L-1

LECTURE HANDOUTS

AI&DS

II/III

Course Name with Code : **19ADC05/ Introduction to Data Science**

Course Teacher :**Dr.P.Srinivasan**

Unit : **I -Introduction** **Date of Lecture:**

Topic of Lecture: Introduction to Data Science

Introduction :

- The Introduction to Data Science class will survey the foundational topics in data science, namely: Data Manipulation. Data Analysis with Statistics and Machine Learning. Data Communication with Information Visualization.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Finding actionable information in large, raw or structured data sets to identify patterns and uncover other insights
- The field primarily seeks to discover answers for areas that are unknown and unexpected

Detailed content of the Lecture:

Data science involves extracting knowledge from data you gather using different methodologies. As a data scientist, you take a complex business problem, compile research from it, creating it into data, then use that data to solve the problem.

What does a data scientist will do?

- Data Acquisition: Here, data scientists take data from all its raw sources, such as databases and flat-files. Then, they integrate and transform it into a homogenous format, collecting it into what is known as a “data warehouse,” a system by which the data can be used to extract information from easily. Also known as ETL, this step can be done with some tools, such as Talend Studio, DataStage and Informatica.

- Data Preparation: This is the most important stage, wherein 60 percent of a data scientist’s time is spent because often data is “dirty” or unfit for use and must be scalable, productive and meaningful. In fact, five sub-steps exist here:

1. Data Cleaning: Important because bad data can lead to bad models, this step handles missing values and null or void values that might cause the models to fail. Ultimately, it improves business decisions and productivity.

2. Data Transformation: Takes raw data and turns it into desired outputs by normalizing it. This step can use, for example, min-max normalization or z-score normalization.

3. Handling Outliers: This happens when some data falls outside the scope of the realm of the rest of the data. Using exploratory analysis, a data scientist quickly uses plots and graphs to determine what to do

with the outliers and see why they're there. Often, outliers are used for fraud detection.

4. Data Integration: Here, the data scientist ensures the data is accurate and reliable.

5. Data Reduction: This compiles multiple sources of data into one, increases storage capabilities, reduces costs and eliminates duplicate, redundant data.

- Data Mining: Here, data scientists uncover the data patterns and relationships to take better business decisions. It's a discovery process to get hidden and useful knowledge, commonly known as exploratory data analysis. Data mining is useful for predicting future trends, recognizing customer patterns, helping to make decisions, quickly detecting fraud and choosing the correct algorithms. Tableau works nicely for data mining.

- Model Building: This goes further than simple data mining and requires building a machine learning model. The model is built by selecting a machine learning algorithm that suits the data, problem statement and available resources.

1. Supervised: Supervised learning algorithms are used when the data is labeled. There are two types:

- Regression: When you need to predict continuous values and variables are linearly dependent, algorithms used are linear and multiple regression, decision trees and random forest

- Classification: When you need to predict categorical values, some of the classification algorithms used are KNN, logistic regression, SVM and Naïve-Bayes

- Unsupervised: Unsupervised learning algorithms are used when the data is unlabeled, there is no labeled data to learn from. There are two types:

- Clustering: This is the method of dividing the objects which are similar between them and dissimilar to others. K-Means and PCA clustering algorithms are commonly used.

- Association-rule analysis: This is used to discover interesting relations between variables, Apriori and Hidden Markov Model algorithm can be used

- Model Maintenance: After gathering data and performing the mining and model building, data scientists must maintain the model accuracy. Thus, they take the following steps:

Video Content / Details of website for further learning (if any):
https://www.simplilearn.com/tutorials/data-science-tutorial/introduction-to-data-science

Important Books/Journals for further learning including the page nos.:

1. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.
2. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

L-2

LECTURE HANDOUTS

AI&DS

II/III

Course Name with Code : **19ADC05/ Introduction to Data Science**

Course Teacher :Dr.P.Srinivasan

Unit : I -Introduction **Date of Lecture:**

Topic of Lecture: Evolution of Data Science

Introduction :

- The term "data science" has been traced back to 1974, when Peter Naur proposed it as an alternative name for computer science. In 1996, the International Federation of Classification Societies became the first conference to specifically feature data science as a topic.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Back in 2001, the term 'data science' was first used in a publication by William Cleveland. Fast forward to 2012 and the Harvard Business Review hailed data science as 'the sexiest job of the 21st century'.
- Fast forward to today and every business wants to employ data scientists

Detailed content of the Lecture:

It happened as a new wave of businesses recognized that data was the key to mastery of modern markets and started making it their strategic focus. In the years since, the field of data science has seen explosive growth as well as some fast-paced developments as higher demand has spurred innovation.

- As far as the field of data science has come since 2010, there's every reason to believe that the next decade will bring even more change. With simultaneous advances in related technology fields and new approaches by the best and brightest minds in the industry, data science in 2030 will bear little resemblance to the state of the art today. Here's a look at how data science is set to evolve over the next decade.

End to End Automation

The first and arguably most important way that data science is going to change in the next decade is that more and more of the work in the field will become fully automated. This is possible in large part due to the rapid advances happening in the development of machine learning and artificial intelligence, which are already making an impact in the world of data science. Right now, it's already becoming more common for data scientists to rely on an automated machine learning pipeline to speed up the process of algorithm selection and hyperparameter tuning.

IoT Brings a Flood of New Data

Data science is, as its name would suggest, all about data. So far, the internet has proven to be the biggest generator of data the world has ever known. Over the next decade, it will be displaced by the internet of things (IoT). Millions of connected devices will come into service over the next ten years, giving data scientists unprecedented access to data of all kinds. They'll use it to gain new insights into old problems, and even provide the rationale for new products and services guided by consumer usage data and other previously-unavailable data streams. In short, IoT will bring data science into the lives of more people than ever before, in more ways than one can count.

Data Visualization Goes VR

Right now, the critical importance of data visualization technology and solutions tends to get lost in the shuffle when discussing data science. Yet, it is the one element that can serve as a bridge between the professionals interpreting data and the people meant to consume the results of their work. Without it, data science would remain in the realm of academics, and businesses couldn't operationalize it to any great effect. That's part of the reason that, according to the Think Big data visualization rankings, there's a growing number of companies all aiming to cement their place as the go-to visualization platform of the moment.

Over the next decade, though, data visualization is going to merge with virtual reality (VR) and augmented reality (AR) technology to create immersive data experiences. In that way, data science consumers will be able to take a hands-on approach to the data they're exploring, allowing for greater digestibility and accessibility for non-technical users. There are already some impressive examples of this coming to market, providing just a small glimpse of what data visualizations might look like in 2030.

Data Privacy Laws Will Become Universal

Right now, the field of data science has advanced so quickly that the legal frameworks necessary to regulate and support it haven't been able to keep pace. In recent years, that's been slowly changing as regulations like the EU's General Data Protection Regulation (GDPR) have come into force, giving individuals additional controls over what businesses can and can't do with their data. In the next few years, it's reasonable to expect to see more regulations of this kind put into place elsewhere in the world.

That's going to have a noticeable effect on the field of data science because it will change the rules of the environment surrounding it. It will place more restrictions on what data may be collected and stored, how it should be safeguarded, and even what kind of analysis it may be used for. In some cases, it may even slow down developments in the field, as efforts to anonymize data or seek additional permissions must be undertaken.

A Brave New World of Data Science

The four broad topics covered here are not the only ones that will reshape the world of data science in the coming decade. In fact, the pace of development in the field all but guarantees that there will be even more changes that nobody working in or covering data science today can yet predict. In the end, though, data science in 2030 will feature processes that are fast, efficient, effective, and informed by more varieties and types of data than are available today. Its conclusions will be communicated via a dazzling array of VR and AR displays. And all of it will be governed by a new set of regulations that protect the rights of all stakeholders. After seeing how far the field has come since its very recent mainstream beginnings, one can only imagine how much further it can and will progress – and the early signs are a wonder to behold.

Video Content / Details of website for further learning (if any):

https://www.simplilearn.com/tutorials/data-science-tutorial/introduction-to-data-science

Important Books/Journals for further learning including the page nos.:

3. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.
4. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

L-3

LECTURE HANDOUTS

AI&DS **II/III**

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : I -Introduction **Date of Lecture:**

Topic of Lecture: Data Science Roles

Introduction :

- AI is a collection of technologies that excel at extracting insights and patterns from large sets of data, then making predictions based on that information. That includes your analytics data from places like Google Analytics, automation platforms, content management systems

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data science gets solutions and results to specific business problems using AI as a tool.
- If data science is to insights, machine learning is to predictions and artificial intelligence is to actions

Detailed content of the Lecture:

Data Scientist Role and Responsibilities

Data scientists work closely with business stakeholders to understand their goals and determine how data can be used to achieve those goals. They design data modeling processes, create algorithms and predictive models to extract the data the business needs, and help analyze the data and share insights with peers. While each project is different, the process for gathering and analyzing data generally follows the below path:

1. Ask the right questions to begin the discovery process
2. Acquire data
3. Process and clean the data
4. Integrate and store data
5. Initial data investigation and exploratory data analysis
6. Choose one or more potential models and algorithms
7. Apply data science techniques, such as machine learning, statistical modeling, and artificial intelligence
8. Measure and improve results
9. Present final result to stakeholders
10. Make adjustments based on feedback
11. Repeat the process to solve a new problem

Common Data Scientist Job Titles

The most common careers in data science include the following roles.

Data scientists: Design data modeling processes to create algorithms and predictive models and perform custom analysis

Data analysts: Manipulate large data sets and use them to identify trends and reach meaningful conclusions to inform strategic business decisions

Data engineers: Clean, aggregate, and organize data from disparate sources and transfer it to data warehouses.

Business intelligence specialists: Identify trends in data sets

Data architects: Design, create, and manage an organization's data architecture

Although the roles of data scientists and data analysts are often conflated, their responsibilities are actually quite different. Put simply, data scientists develop processes for modeling data while data analysts examine data sets to identify trends and draw conclusions. Because of this distinction and the more technical nature of data science, the role of a data scientist is often considered to be more senior than that of a data analyst; however, both positions may be attainable with similar educational backgrounds.

Essential Data Science Skills

Most data scientists use the following core skills in their daily work:

Statistical analysis: Identify patterns in data. This includes having a keen sense of pattern detection and anomaly detection.

Machine learning: Implement algorithms and statistical models to enable a computer to automatically learn from data.

Computer science: Apply the principles of artificial intelligence, database systems, human/computer interaction, numerical analysis, and software engineering.

Programming: Write computer programs and analyze large datasets to uncover answers to complex problems. Data scientists need to be comfortable writing code working in a variety of languages such as Java, R, Python, and SQL.

Data storytelling: Communicate actionable insights using data, often for a non-technical audience.

Data scientists play a key role in helping organizations make sound decisions. As such, they need "soft skills" in the following areas.

Business intuition: Connect with stakeholders to gain a full understanding of the problems they're looking to solve.

Analytical thinking. Find analytical solutions to abstract business issues.

Critical thinking: Apply objective analysis of facts before coming to a conclusion.

Inquisitiveness: Look beyond what's on the surface to discover patterns and solutions within the data.

Interpersonal skills: Communicate across a diverse audience across all levels of an organization.

Video Content / Details of website for further learning (if any):

https://www.northeastern.edu/graduate/blog/what-does-a-data-scientist-do/

Important Books/Journals for further learning including the page nos.:

5. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.
6. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

L-4

LECTURE HANDOUTS

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : I -Introduction Date of Lecture:

Topic of Lecture: Stages in a Data Science Project

Introduction :

- There are altogether 5 steps of a data science project starting from Obtaining Data, Scrubbing Data, Exploring Data, Modelling Data and ending with Interpretation of Data

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes
- Data Science Process goes through Discovery, Data Preparation, Model Planning, Model Building, Operationalize, Communicate Results

Detailed content of the Lecture:

1. **Understand the Business**

Understanding the business or activity that your data project is part of is key to ensuring its success and the first phase of any sound data analytics project. To motivate the different actors necessary to getting your project from design to production, your project must be the answer to a clear organizational need. Before you even think about the data, go out and talk to the people in your organization whose processes or whose business you aim to improve with data. Then, sit down to define a timeline and concrete key performance indicators. I know, planning and processes seem boring, but, in the end, they are an essential first step to kickstart your data initiative!

2. **Get your data**

Connect to a database: Ask your data and IT teams for the data that’s available or open up your private database and start digging through it to understand what information your company has been collecting.
Use APIs: Think of the APIs to all the tools your company’s been using and the data these guys have been collecting. You have to work on getting these all set up so you can use those email open and click stats, the information your sales team put in Pipedrive or Salesforce, the support ticket somebody submitted, etc. If you’re not an expert coder, plugins in Dataiku give you lots of possibilities to bring in external data!
Look for open data: The Internet is full of datasets to enrich what you have with additional information. For example, census data will help you add the average revenue for the district where your user lives or OpenStreetMap can show you how many coffee shops are on a given street

3. **Explore and Clean your data**

Once you've gotten your data, it's time to get to work on it in the third data analytics project phase. Start digging to see what you've got and how you can link everything together to achieve your original goal. Start taking notes on your first analyses and ask questions to business people, the IT team, or other groups to understand what all your variables mean.

The next step (and by far the most dreaded one) is cleaning your data. You've probably noticed that even though you have a country feature, for instance, you've got different spellings, or even missing data. It's time to look at every one of your columns to make sure your data is homogeneous and clean.

4. **Enrich your data set**

Extracting date components (month, hour, day of the week, week of the year, etc.)
Calculating differences between date columns
Flagging national holidays

Another way of enriching data is by joining datasets — essentially, retrieving columns from one dataset or tab into a reference dataset. This is a key element of any analysis, but it can quickly become a nightmare when you have an abundance of sources. Luckily, some tools such as Dataiku allow you to blend data through a simplified process, by easily retrieving data or joining datasets based on specific, fine-tuned criteria

5. **Build helpful visualization**

You now have a nice dataset (or maybe several), so this is a good time to start exploring it by building graphs. When you're dealing with large volumes of data, visualization is the best way to explore and communicate your findings and is the next phase of your data analytics project.

The tricky part here is to be able to dig into your graphs at any time and answer any question someone would have about a given insight. That's when the data preparation comes in handy: you're the guy or gal who did all the dirty work, so you know the data like the palm of your hand

6. **Get Predictive**

The next data science step, phase six of the data project, is when the real fun starts. Machine learning algorithms can help you go a step further into getting insights and predicting future trends. By working with clustering algorithms (aka unsupervised), you can build models to uncover trends in the data that were not distinguishable in graphs and stats. These create groups of similar events (or clusters) and more or less explicitly express what feature is decisive in these results.

More advanced data scientists can go even further and predict future trends with supervised algorithms. By analyzing past data, they find features that have impacted past trends, and use them to build predictions. More than just gaining knowledge, this final step can lead to building entirely new products and processes.

7. **Iterate, Iterate, Iterate**

The main goal in any business project is to prove its effectiveness as fast as possible to justify, well, your job. The same goes for data projects. By gaining time on data cleaning and enriching, you can go to the end of the project fast and get your initial results. This is the final phase of completing your data analytics project and one that is critical to the entire data life cycle.

Video Content / Details of website for further learning (if any):
https://blog.dataiku.com/2019/07/04/fundamental-steps-data-project-success

Important Books/Journals for further learning including the page nos.:

7. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.
8. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-5

AI&DS

II/III

Course Name with Code : **19ADC05/ Introduction to Data Science**

Course Teacher **:Dr.P.Srinivasan**

Unit **: I -Introduction** **Date of Lecture:**

Topic of Lecture: Applications of Data Science in various fields

Introduction :

Data is such an asset that can be used by people to accomplish many feats. With the advancement of technology, the availability of data is also increasing and data science has been successful in analyzing, managing, and tackling the data every day.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes
- Data Science Process goes through Discovery, Data Preparation, Model Planning, Model Building, Operationalize, Communicate Results

Detailed content of the Lecture:

Education

- Educational data is becoming increasingly valuable in higher education, as a growing number of online courses are being used. It even extends to the private sector, where personnel are educated and task issues are solved via online forums, threads, and distributed problem-solving methods through assignments.According to this source, here are many advantages that we will get on using data science in education, like:

- Educational data science would prepare teachers to investigate various types of educational data, as well as to give meaning to educational systems, their issues, and prospective remedies, and to build a deeper knowledge and experimentally verified forms of answers.

- Educators would be able to undertake data visualisation, data reduction and description, and prediction tasks with the help of educational data science.

- For practitioners, data visualisation may make information more intuitive and consumable.

- Many complicated records and fields of data about pupils can be deciphered via data reduction.

- It has helped innumerable people shape their careers and succeed in their lives. Now, education is one of the sectors where Data Science is making drastic changes in improving the entire system and the performance of students, teachers, and other key stakeholders

Airline Route Planning

One of the areas where Data Science is revolutionizing the day-to-day business activities is the Airline industry. For a very long time, the Airline industry has been bearing a substantial decline in revenue generation, and then Data Science emerged as a savior. Previously, due to competition, the airline companies used to provide discounts to customers to attract them. Also, due to the high rates of fuels and the lack of proper analysis for the delay of flights, destination, halts, optimized routes, etc., airlines were unnecessarily spending a lot on extra fuel.Do you know what will this number be when a newer version of the aircraft starts operations? It will be thrice the amount of data mentioned above.

- The data thus collected can be scrutinized and analyzed to improve flight safety.
- Constant innovations with data will help the Aviation sector answer questions like these and collect relevant data.
- Data analytics will help identify major risks and the solutions to ensure passenger safety. This will become extremely crucial when air traffic is expected to double in the next 20 years.
- Identifying potential customers to offer calculated discounts, instead of providing discounts to everyone
- Deciding on the optimized routes by analyzing the traffic on different routes. It helps in saving expensive fuel that gets unnecessarily exhausted otherwise
- Predicting delays in flight
- Setting the cost of flights as per seasons, festivals, and the number of travelers. This is done by analyzing the number of potential travelers and frequent travelers

Healthcare Industry

The advancement in medical science has revolutionized the healthcare industry. There are fast and effective treatments for almost all diseases. However, there are areas of healthcare where Data Science is helping in efficient diagnosis, data management, medical research, etc. through data analysis and visualization.

- Using big data with a combination of large and complex data sets includes electronic medical records, social media, genomic information, and digital body data from wireless health devices.
- With new open-access efforts that seek to utilize the availability of clinical trials, research, and citizen science sources for data sharing.

Video Content / Details of website for further learning (if any):
https://intellipaat.com/blog/applications-of-data-science-real-world-applications/

Important Books/Journals for further learning including the page nos.:

09.Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

10.David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS L-6

AI&DS **II/III**

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : I -Introduction **Date of Lecture:**

Topic of Lecture: Applications of Data Science in various fields

Introduction :

Data is such an asset that can be used by people to accomplish many feats. With the advancement of technology, the availability of data is also increasing and data science has been successful in analyzing, managing, and tackling the data every day.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes
- Data Science Process goes through Discovery, Data Preparation, Model Planning, Model Building, Operationalize, Communicate Results

Detailed content of the Lecture:

Delivery Logistics

When it comes to delivery logistics Data Science is not far behind the race. Many companies use data for optimizing their business operations. It helps in the analysis of profit generation, the causes of loss, the best route for delivery, the time required, and the scope for improvements. Other than that, the application of Data Science in delivery logistics helps the companies analyze the market trend and increase their competence. Further, with the help of route optimization, the number of deliveries increases and the freight cost reduces. By this, companies can boost their profits. This is how the application of Data Science plays a major role in logistics.

Logistics analytics is a term used to describe analytical procedures conducted by organizations to analyze and coordinate the logistical function and supply chain to ensure smooth running of operations in a timely, and cost-effective manner.

Banking and Finance

- Data science combines several disciplines in the process of using statistical and scientific methods on data to get insights. This information is useful for strategic decision making in an organisation. This process is iterative in nature and usually follows these steps – defining the problem, planning the process, collecting data, processing raw data to get it ready for analyses, performing the analyses and then communicating the insights to the stakeholders.
- With so much data available today and easy access to efficient computational power, data science in finance has been successfully implemented to open more doors to data driven decision making.
- Data scientists utilize the behavioral, demographic, and historical purchase data to build a model

that predicts the probability of a customer's response to a promotion or an offer. Therefore, banks can make an efficient, personalized outreach and improve their relationships with customers.

By processing the data, the system can analyze individual customer's financial history, loans, income, and debt clearance. Also, it helps in finding both genuine and suspicious behaviors in transactions. Further, the application of Data Science in Finance helps in the following ways:

Stamping out Tax Fraud:

The economy of a nation depends on its taxpayers. Governments need taxation to maintain the economic infrastructure of countries. If there occurs a fraud activity related to taxation, then it directly affects every citizen of a nation. Therefore, governments have started implementing Data Science to analyze citizen data to prevent tax fraud. With the application of Data Science, the income tax departments keep track of the income and calculate the tax. If the calculated tax is not collected, they track the suspicious taxpayers to take action against them.

Credit Scoring:

Credit scoring is another application of Data Science that helps check the financial civil score of an individual. The credit score is rated out of 10. It helps financial institutions make decisions on sanctioning loans. They decide the loan amount and its sanctioning on the basis of the credit score calculated out of 10.

Filtered Internet Search

Google picks up the most relevant and highly searched long-tail keyword related to your typed keyword. It helps in optimized searching to get better results. This is a type of filtered Internet search and also one of the wonderful applications of Data Science. It is only possible with the help of Data Science. Google collects and stores the data of search history to analyze and visualize it. Then, it uses algorithms and techniques that apply filters on the data to check the frequency of the searched keyword and related topics to show you the best results.

- Data filtering is the process of choosing a smaller part of your data set and using that subset for viewing or analysis.
- Filtering is generally (but not always) temporary – the complete data set is kept, but only part of it is used for the calculation.
- The process of selectively eliminating the data that are not relevant to our decision. Although the implementation of search and filter technology are quite different, they essentially solve the same problem.
- At the abstract level, they narrow the data down to a much smaller set that is relevant to our decision.

Video Content / Details of website for further learning (if any):

https://intellipaat.com/blog/applications-of-data-science-real-world-applications/

Important Books/Journals for further learning including the page nos.:

11. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.
12. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

L-7

AI&DS

II/III

Course Name with Code : **19ADC05/ Introduction to Data Science**

Course Teacher :Dr.P.Srinivasan

Unit : I -Introduction **Date of Lecture:**

Topic of Lecture: Applications of Data Science in various fields

Introduction :

Data is such an asset that can be used by people to accomplish many feats. With the advancement of technology, the availability of data is also increasing and data science has been successful in analyzing, managing, and tackling the data every day.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Science is the area of study which involves extracting insights from vast amounts of data by the use of various scientific methods, algorithms, and processes
- Data Science Process goes through Discovery, Data Preparation, Model Planning, Model Building, Operationalize, Communicate Results

Detailed content of the Lecture:

Product Recommendation Systems

Product recommendation is an effective way of converting leads into sales. All the industries based on sales use recommendation systems for improving their profitability. But, how do these recommendation engines work? It is again a Data Science application. With the help of multiple tools and techniques of Data Science, a system records customer data such as browsing history, products selected for purchase, items added to the cart, etc. Then, it tries to understand the patterns in this data to filter the customers who are likely to make a purchasing decision, i.e., the system filters the leads that might convert into sales. After finding out such potential customers, the recommendation system starts suggesting products to them. It advertises the products on different websites browsed by the same customers.

- Because search is very efficient, we can start with a blank page like Google’s home page and then populate it with more and more relevant data through query refinement.
- Filtering is less efficient, because it often require showing samples from the entire data set for the user to filter upon in order to remove the irrelevant data. That is, the user has to look through the sample data to determine what’s irrelevant. Therefore, true filtering functions are rarely applied to very large data sets at the web scale.
- A Recommender System refers to a system that is capable of predicting the future preference of a set of items for a user, and recommend the top items.
- One key reason why we need a recommender system in modern society is that people have too much options to use from due to the prevalence of Internet.

- In the past, people used to shop in a physical store, in which the items available are limited. For instance, the number of movies that can be placed in a Blockbuster store depends on the size of that store.
- By contrast, nowadays, the Internet allows people to access abundant resources online. Netflix, for example, has an enormous collection of movies.
- Although the amount of available information increased, a new problem arose as people had a hard time selecting the items they actually want to see. This is where the recommender system comes in.

Digital Marketing and Advertising

Marketing Data Science

As a marketing data scientist, you'll be using the usual tools: SQL, R, Python, and a preferred data visualization approach (usually this means Tableau). Depending on the size of the enterprise you're working for, you may also have survey and questionnaire construction responsibilities — which is both an art and a science unto itself. You'll also likely be tasked with analyzing consumer responses, sales call logs, customer service logs, in addition to external data sources such as social media mentions and interactions with your employer's brand. But, that's only the beginning.

Companies want to know what their competitors are and aren't doing so they can target an expansion of their market share for a particular consumer segment. Thus, pulling together and analyzing data about competitors, including pricing, news, and consumer sentiment, will likely be an additional requirement. Data scientists will take these activities a step further and build predictive and prescriptive models to automate the process of deriving actionable insights and producing a recommended course of action.

The Marketing Funnel

While there are mixed views about the utility of the Marketing Funnel, it is a handy classification system for explaining data science use cases as applied to the world of marketing and advertising. Summarily, the Marketing Funnel is separated into three meta-categories: Lead Generation, Lead Nurturing, and Sales (which is at the bottom of the funnel). Each component is discussed in further detail below along with an example of how data science can be used to guide decision making in terms of improving the movement of customers towards a sales conversion.

Data Scientists design algorithms to analyze and visualize customers' data related to their search history, interests, and previously shopped items. Also, the system identifies relevant websites to post ads for marketing. With the help of digital ads, the click-through rate (CTR) of a website increases by a high frequency. By digital marketing, the value of the business increases as advertising enhances the visibility of the company in the market.

Video Content / Details of website for further learning (if any):
https://intellipaat.com/blog/applications-of-data-science-real-world-applications/

Important Books/Journals for further learning including the page nos.:

13. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.

14. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

LECTURE HANDOUTS

L-8

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : I -Introduction **Date of Lecture:**

Topic of Lecture: Data Security Issues

Introduction :

Data security is the process of protecting corporate data and preventing data loss through unauthorized access. This includes protecting your data from attacks that can encrypt or destroy data, such as ransomware, as well as attacks that can modify or corrupt your data.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Applications of Data Science in various fields
- Stages in a Data Science Project

Detailed content of the Lecture:

Data Security

Big data security is an umbrella term that includes all security measures and tools applied to analytics and data processes. Attacks on big data systems – information theft, DDoS attacks, ransomware, or other malicious activities – can originate either from offline or online spheres and can crash a system.

The consequences of information theft can be even worse when organizations store sensitive or confidential information like credit card numbers or customer information. They may face fines because they failed to meet basic data security measures to be in compliance with data loss protection and privacy mandates like the General Data Protection Regulation (GDPR).

Six Big Data Security Challenges

Big Data challenges are not limited to on-premise platforms. They also affect the cloud. The list below reviews the six most common challenges of big data on-premises and in the cloud.

1.Distributed Data

Most big data frameworks distribute data processing tasks throughout many systems for faster analysis. Hadoop, for example, is a popular open-source framework for distributed data processing and storage. Hadoop was originally designed without any security in mind.

2.Cybercriminals

Cybercriminals can force the MapReduce mapper to show incorrect lists of values or key pairs, making the MapReduce process worthless. Distributed processing may reduce the workload on a system, but eventually more systems mean more security issues.

3.Non-Relational Databases

Traditional relational databases use tabular schema of rows and columns. As a result, they cannot handle big data because it is highly scalable and diverse in structure. Non-relational databases, also known as NoSQL databases, are designed to overcome the limitations of relational databases.

Non-relational databases do not use the tabular schema of rows and columns. Instead, NoSQL databases optimize storage models according to data type. As a result, NoSQL databases are more flexible and scalable than their relational alternatives.

NoSQL databases favor performance and flexibility over security. Organizations that adopt NoSQL databases have to set up the database in a trusted environment with additional security measures.

4. Endpoint Vulnerabilities

Cybercriminals can manipulate data on endpoint devices and transmit the false data to data lakes. Security solutions that analyze logs from endpoints need to validate the authenticity of those endpoints.
For example, hackers can access manufacturing systems that use sensors to detect malfunctions in the processes. After gaining access, hackers make the sensors show fake results. Challenges like that are usually solved with fraud detection technologies.

5. Data Mining Solutions

Data mining is the heart of many big data environments. Data mining tools find patterns in unstructured data. The problem is that data often contains personal and financial information. For that reason, companies need to add extra security layers to protect against external and internal threats.

6. Access Controls

Companies sometimes prefer to restrict access to sensitive data like medical records that include personal information. But people that do not have access permission, such as medical researchers, still need to use this data. The solution in many organizations is to grant granular access. This means that individuals can access and see only the information they need to see.Big data technologies are not designed for granular access. A solution is to copy required data to a separate big data warehouse. For example, only the medical information is copied for medical research without patient names and addresses.

Video Content / Details of website for further learning (if any):
https://www.dataversity.net/big-data-security-challenges-and-solutions/

Important Books/Journals for further learning including the page nos.:

15. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

16. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-9

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : I -Introduction **Date of Lecture:**

Topic of Lecture: Data Security Issues

Introduction :

Data security is the process of protecting corporate data and preventing data loss through unauthorized access. This includes protecting your data from attacks that can encrypt or destroy data, such as ransomware, as well as attacks that can modify or corrupt your data.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Security Issues
- Applications of Data Science in various fields

Detailed content of the Lecture:

Addressing Big Data Security Threats

Security tools for big data are not new. They simply have more scalability and the ability to secure many data types. The list below explains common security techniques for big data.

Encryption

Big data encryption tools need to secure data-at-rest and in-transit across large data volumes. Companies also need to encrypt both user and machine-generated data. As a result, encryption tools have to operate on multiple big data storage formats like NoSQL databases and distributed file systems like Hadoop.

User Access Control

User access control is a basic network security tool. The lack of proper access control measures can be disastrous for big data systems. A robust user control policy has to be based on automated role-based settings and policies. Policy-driven access control protects big data platforms against insider threats by automatically managing complex user control levels, like multiple administrator settings.

Intrusion Detection and Prevention

The distributed architecture of big data is a plus for intrusion attempts. An Intrusion Prevention System (IPS) enables security teams to protect big data platforms from vulnerability exploits by examining network traffic. The IPS often sits directly behind the firewall and isolates the intrusion before it does actual damage.

Centralized Key Management

Key management is the process of protecting cryptographic keys from loss or misuse. Centralized key management offers more efficiency as opposed to distributed or application-specific management. Centralized management systems use a single point to secure keys and access audit logs and policies. A reliable key management system is essential for companies handling sensitive information.

A growing number of companies use big data analytics tools to improve business strategies. That gives cybercriminals more opportunities to attack big data architecture. Thus the list of big data security issues continues to grow.

There are many privacy concerns and government regulations for big data platforms. However, organizations and private users do not always know what is happening with their data and where the data is stored.

Data security solution protects your data wherever it lives—on-premises, in the cloud, and in hybrid environments. It also provides security and IT teams with full visibility into how the data is being accessed, used, and moved around the organization.

Our comprehensive approach relies on multiple layers of protection, including:

- Database firewall—blocks SQL injection and other threats, while evaluating for known vulnerabilities.
- User rights management—monitors data access and activities of privileged users to identify excessive, inappropriate, and unused privileges.
- Data masking and encryption—obfuscates sensitive data so it would be useless to the bad actor, even if somehow extracted.
- Data loss prevention (DLP)—inspects data in motion, at rest on servers, in cloud storage, or on endpoint devices.
- User behavior analytics—establishes baselines of data access behavior, uses machine learning to detect and alert on abnormal and potentially risky activity.
- Database activity monitoring—monitors relational databases, data warehouses, big data, and mainframes to generate real-time alerts on policy violations.
- Alert prioritization—Imperva uses AI and machine learning technology to look across the stream of security events and prioritize the ones that matter most.

Video Content / Details of website for further learning (if any):
https://www.dataversity.net/big-data-security-challenges-and-solutions/

Important Books/Journals for further learning including the page nos.:

17. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.
18. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS **L-10**

AI&DS **II/III**

Course Name with Code : **19ADC05/ Introduction to Data Science**

Course Teacher :**Dr.P.Srinivasan**

Unit : **II - Data Collection and Data Pre-Processing** **Date of Lecture:**

Topic of Lecture: Data Collection Strategies

Introduction :

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Security Issues
- Applications of Data Science in various fields

Detailed content of the Lecture:

Data collection is the process of collecting, measuring and analyzing different types of information using a set of standard validated techniques. The main objective of data collection is to gather information-rich and reliable data, and analyze them to make critical business decisions. Once the data is collected, it goes through a rigorous process of data cleaning and data processing to make this data truly useful for businesses.

There are two main methods of data collection in research based on the information that is required, namely:

- Primary Data Collection
- Secondary Data Collection

Primary Data Collection Methods

Primary data refers to data collected from first-hand experience directly from the main source. It refers to data that has never been used in the past. The data gathered by primary data collection methods are generally regarded as the best kind of data in research.

The methods of collecting primary data can be further divided into quantitative data collection methods (deals with factors that can be counted) and qualitative data collection methods (deals with factors that are not necessarily numerical in nature). Here are some of the most common primary data collection.

1. Interviews

Interviews are a direct method of data collection. It is simply a process in which the interviewer asks questions and the interviewee responds to them. It provides a high degree of flexibility because questions can be adjusted and changed anytime according to the situation.

2. Observations

In this method, researchers observe a situation around them and record the findings. It can be used to evaluate the behaviour of different people in controlled (everyone knows they are being observed) and uncontrolled (no one knows they are being observed) situations. This method is highly effective because it is straightforward and not directly dependent on other participants.

For example, a person looks at random people that walk their pets on a busy street, and then uses this data to decide whether or not to open a pet food store in that area.

3. Surveys and Questionnaires

Surveys and questionnaires provide a broad perspective from large groups of people. They can be conducted face-to-face, mailed, or even posted on the Internet to get respondents from anywhere in the world. The answers can be yes or no, true or false, multiple choice, and even open-ended questions. However, a drawback of surveys and questionnaires is delayed response and the possibility of ambiguous answers.

4. Focus Groups

A focus group is similar to an interview, but it is conducted with a group of people who all have something in common. The data collected is similar to in-person interviews, but they offer a better understanding of why a certain group of people thinks in a particular way. However, some drawbacks of this method are lack of privacy and domination of the interview by one or two participants. Focus groups can also be time-consuming and challenging, but they help reveal some of the best information for complex situations.

5. Oral Histories

Oral histories also involve asking questions like interviews and focus groups. However, it is defined more precisely and the data collected is linked to a single phenomenon. It involves collecting the opinions and personal experiences of people in a particular event that they were involved in. For example, it can help in studying the effect of a new product in a particular community.

Video Content / Details of website for further learning (if any):
https://www.simplilearn.com/data-collection-methods-article

Important Books/Journals for further learning including the page nos.:

19. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.
20. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-11

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : II - Data Collection and Data Pre-Processing Date of Lecture:

Topic of Lecture: Data Collection Strategies

Introduction :

Data collection is the process of gathering and measuring information on variables of interest, in an established systematic fashion that enables one to answer stated research questions, test hypotheses, and evaluate outcomes.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Security Issues
- Applications of Data Science in various fields

Detailed content of the Lecture:

Secondary Data Collection Methods

Secondary data refers to data that has already been collected by someone else. It is much more inexpensive and easier to collect than primary data. While primary data collection provides more authentic and original data, there are numerous instances where secondary data collection provides great value to organizations.

Here are some of the most common secondary data collection methods:

1. Internet

The use of the Internet has become one of the most popular secondary data collection methods in recent times. There is a large pool of free and paid research resources that can be easily accessed on the Internet. While this method is a fast and easy way of data collection, you should only source from authentic sites while collecting information.

2. Government Archives

There is lots of data available from government archives that you can make use of. The most important advantage is that the data in government archives are authentic and verifiable. The challenge, however, is that data is not always readily available due to a number of factors. For example, criminal records can come under classified information and are difficult for anyone to have access to them.

3. Libraries

Most researchers donate several copies of their academic research to libraries. You can collect important and authentic information based on different research contexts. Libraries also serve as a storehouse for business directories, annual reports and other similar documents that help businesses in their research.

Use Case: Conducting Customer Surveys to Multiply Sales

A research study was conducted by Rice University Professor Dr. Paul Dholakia and Dr. Vicki Morwitz to see whether a company could influence customers' loyalty or buying habits. The research study was conducted over the course of a year. One group of customers were surveyed and the other set was not surveyed about customer satisfaction. In the next year, the group that took the survey were thrice as likely to renew their loyalty towards the organization than the other group.

Video Content / Details of website for further learning (if any):
https://www.simplilearn.com/data-collection-methods-article

Important Books/Journals for further learning including the page nos.:

21. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.
22. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

LECTURE HANDOUTS

L-12

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : II - Data Collection and Data Pre-Processing Date of Lecture:

Topic of Lecture: Data Pre-Processing

Introduction :

Data processing occurs when data is collected and translated into usable information. Usually performed by a data scientist or team of data scientists, it is important for data processing to be done correctly as not to negatively affect the end product, or data output.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Security Issues
- Data Collection Strategies

Detailed content of the Lecture:

Data processing occurs when data is collected and translated into usable information. Usually performed by a data scientist or team of data scientists, it is important for data processing to be done correctly as not to negatively affect the end product, or data output.

Steps Involved in Data Preprocessing:

1. Data Cleaning
2. Data Transformation
3. Data Reduction

1. Data Cleaning:

The data can have many irrelevant and missing parts. To handle this part, data cleaning is done. It involves handling of missing data, noisy data etc.

- **(a). Missing Data:**

This situation arises when some data is missing in the data. It can be handled in various ways. Some of them are:

1. **Ignore the tuples:**
   This approach is suitable only when the dataset we have is quite large and multiple values are missing within a tuple.

2. **Fill the Missing values:**

There are various ways to do this task. You can choose to fill the missing values manually, by attribute mean or the most probable value.

- **(b). Noisy Data:**

Noisy data is a meaningless data that can't be interpreted by machines.It can be generated due to faulty data collection, data entry errors etc. It can be handled in following ways :

1. **Binning Method:**

This method works on sorted data in order to smooth it. The whole data is divided into segments of equal size and then various methods are performed to complete the task. Each segmented is handled separately. One can replace all data in a segment by its mean or boundary values can be used to complete the task.

2. **Regression:**

Here data can be made smooth by fitting it to a regression function.The regression used may be linear (having one independent variable) or multiple (having multiple independent variables).

3. **Clustering:**

This approach groups the similar data in a cluster. The outliers may be undetected or it will fall outside the clusters.

Video Content / Details of website for further learning (if any):
https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

Important Books/Journals for further learning including the page nos.:

23. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.

24. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS **L-13**

AI&DS **II/III**

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : II - Data Collection and Data Pre-Processing Date of Lecture:

Topic of Lecture: Data Pre-Processing

Introduction :

Data processing occurs when data is collected and translated into usable information. Usually performed by a data scientist or team of data scientists, it is important for data processing to be done correctly as not to negatively affect the end product, or data output.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Collection Strategies
- Data Pre-Processing

Detailed content of the Lecture:

2. Data Transformation:

This step is taken in order to transform the data in appropriate forms suitable for mining process. This involves following ways:

1. **Normalization:**
   It is done in order to scale the data values in a specified range (-1.0 to 1.0 or 0.0 to 1.0)

2. **Attribute Selection:**
   In this strategy, new attributes are constructed from the given set of attributes to help the mining process.

3. **Discretization:**
   This is done to replace the raw values of numeric attribute by interval levels or conceptual levels.

4. **Concept Hierarchy Generation:**
   Here attributes are converted from lower level to higher level in hierarchy. For Example-The attribute “city” can be converted to “country”.

3. Data Reduction:

Since data mining is a technique that is used to handle huge amount of data. While working with huge volume of data, analysis became harder in such cases. In order to get rid of this, we uses data reduction technique. It aims to increase the storage efficiency and reduce data storage and analysis costs.

The various steps to data reduction are:

1. **Data Cube Aggregation:**
Aggregation operation is applied to data for the construction of the data cube.

2. **Attribute Subset Selection:**
The highly relevant attributes should be used, rest all can be discarded. For performing attribute selection, one can use level of significance and p- value of the attribute.the attribute having p-value greater than significance level can be discarded.

3. **Numerosity Reduction:**
This enable to store the model of data instead of whole data, for example: Regression Models.

4. **Dimensionality Reduction:**
This reduce the size of data by encoding mechanisms.It can be lossy or lossless. If after reconstruction from compressed data, original data can be retrieved, such reduction are called lossless reduction else it is called lossy reduction.

Video Content / Details of website for further learning (if any):
https://www.geeksforgeeks.org/data-preprocessing-in-data-mining/

Important Books/Journals for further learning including the page nos.:

25. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.

26. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-14

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : II - Data Collection and Data Pre-Processing Date of Lecture:

Topic of Lecture: Data Cleaning

Introduction :

Data cleansing includes more actions than removing data, such as fixing spelling and syntax errors, standardizing data sets, and correcting mistakes such as missing codes, empty fields, and identifying duplicate records.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Pre-Processing Overview
- Data Pre-Processing

Detailed content of the Lecture:

Data cleaning, or data cleansing, is the important **process of correcting or removing incorrect, incomplete, or duplicate data within a dataset**. Data cleaning **should be the first step in your workflow**. When working with large datasets and combining various data sources, there's a strong possibility you may duplicate or mislabel data. If you have inaccurate or incorrect data, it will lose its quality, and your algorithms and outcomes become unreliable.

Data cleaning **differs from data transformation because you're actually removing data that doesn't belong in your dataset**. With data transformation, you're changing your data to a different format or structure. Data transformation processes are sometimes referred to as data wrangling or data munging. The data cleaning process is what we'll focus on today.

Handling missing data

It's common for large datasets to have some missing values. Maybe the person recording the data forgot to input them, or maybe they began collecting those missing data variables late into the data collection process. No matter what, **missing data should be managed before working with datasets**.

Filtering unwanted outliers

Outliers hold essential information about your data, but at the same time take your focus away from the main group. It’s a good idea to **examine your data with and without outliers**. If you discover you want to use them, be sure to choose a robust method that can handle your outliers. If you decide against using them, you can just drop them.

Standardizing your data

The **data in your feature variables should be standardized**. It makes examining and modeling your data a lot easier. For example, let’s look at two values we’ll call “dog” and “cat” that are in the “animal” variable. If you collected data, you may receive different data values that you didn’t anticipate, such as:

- DOG, CAT (entered in all caps)
- Dog, Cat (entered with first letters capitalized)
- dof, cart (entered as typos)

Dropping dirty data and duplication

Dirty data includes any data points that are wrong or just shouldn’t be there. Duplicates occur when data points are repeated in your dataset. **If you have a lot of duplicates, it can throw off the training of your machine learning model**.

To handle dirty data, you can either drop them or use a replacement (like converting incorrect data points into the correct ones). To handle duplication issues, you can just drop them from your data.

Removing blank data

You obviously can’t use blank data for data analysis. Blank data is a major issue for analysts because it weakens the quality of the data. You should ideally **remove blank data in the data collection phase**, but you can also write a program to do this for you.

Eliminating white space

White space is a small but common issue within many data structures. A TRIM function will help you eliminate white space.

Video Content / Details of website for further learning (if any):
https://www.educative.io/blog/what-is-data-cleaning

Important Books/Journals for further learning including the page nos.:

27. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

28. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L-15

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : II - Data Collection and Data Pre-Processing Date of Lecture:

Topic of Lecture: Data Integration

Introduction :

Data integration is a common industry term referring to the requirement to combine data from multiple separate business systems into a single unified view, often called a single view of the truth. This unified view is typically stored in a central data repository known as a data warehouse.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Pre-Processing Overview
- Data Cleaning

Detailed content of the Lecture:

Data integration is the practice of consolidating data from disparate sources into a single dataset with the ultimate goal of providing users with consistent access and delivery of data across the spectrum of subjects and structure types, and to meet the information needs of all applications and business processes. The data integration process is one of the main components in the overall data management process, employed with increasing frequency as big data integration and the need to share existing data continues to grow.

Data integration architects develop data integration software programs and data integration platforms that facilitate an automated data integration process for connecting and routing data from source systems to target systems. This can be achieved through a variety of data integration techniques, including:

- Extract, Transform and Load: copies of datasets from disparate sources are gathered together, harmonized, and loaded into a data warehouse or database
- Extract, Load and Transform: data is loaded as is into a big data system and transformed at a later time for particular analytics uses
- Change Data Capture: identifies data changes in databases in real-time and applies them to a data warehouse or other repositories
- Data Replication: data in one database is replicated to other databases to keep the information the

information synchronized to operational uses and for backup

- Data Virtualization: data from different systems are virtually combined to create a unified view rather than loading data into a new repository
- Streaming Data Integration: a real time data integration method in which different streams of data are continuously integrated and fed into analytics systems and data stores

Data integration techniques are available across a broad range of organizational levels, from fully automated to manual methods. Typical tools and techniques for data integration include:

- Manual Integration or Common User Interface: There is no unified view of the data. Users operate with all relevant information accessing all the source systems.
- Application Based Integration: requires each application to implement all the integration efforts; manageable with a small number of applications
- Middleware Data Integration: transfers integration logic from an application to a new middleware layer
- Uniform Data Access: leaves data in the source systems and defines a set of views to provide a unified view to users across the enterprise
- Common Data Storage or Physical Data Integration: creates a new system in which a copy of the data from the source system is stored and managed independently of the original system

Developers may use Structured Query Language (SQL) to code a data integration system by hand. There are also data integration toolkits available from various IT vendors that streamline, automate, and document the development process.

Video Content / Details of website for further learning (if any):
https://www.omnisci.com/technical-glossary/data-integration

Important Books/Journals for further learning including the page nos.:

29. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.
30. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-16

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : II - Data Collection and Data Pre-Processing Date of Lecture:

Topic of Lecture: Data Transformation

Introduction :

Data integration is a common industry term referring to the requirement to combine data from multiple separate business systems into a single unified view, often called a single view of the truth. This unified view is typically stored in a central data repository known as a data warehouse.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Cleaning
- Data Integration

Detailed content of the Lecture:

Data transformation is the process of changing the format, structure, or values of data. For data analytics projects, data may be transformed at two stages of the data pipeline. Organizations that use on-premises data warehouses generally use an ETL (extract, transform, load) process, in which data transformation is the middle step. Today, most organizations use cloud-based data warehouses, which can scale compute and storage resources with latency measured in seconds or minutes. The scalability of the cloud platform lets organizations skip preload transformations and load raw data into the data warehouse, then transform it at query time — a model called ELT (extract, load, transform).

Processes such as data integration, data migration, data warehousing, and data wrangling all may involve data transformation.

Data transformation may be constructive (adding, copying, and replicating data), destructive (deleting fields and records), aesthetic (standardizing salutations or street names), or structural (renaming, moving, and combining columns in a database).

An enterprise can choose among a variety of ETL tools that automate the process of data transformation. Data analysts, data engineers, and data scientists also transform data using scripting languages such as Python or domain-specific languages like SQL.

Benefits and challenges of data transformation

Transforming data yields several benefits:

- Data is transformed to make it better-organized. Transformed data may be easier for both humans and computers to use.

- Properly formatted and validated data improves data quality and protects applications from potential landmines such as null values, unexpected duplicates, incorrect indexing, and incompatible formats.

- Data transformation facilitates compatibility between applications, systems, and types of data. Data used for multiple purposes may need to be transformed in different ways..

However, there are challenges to transforming data effectively:

- Data transformation can be expensive. The cost is dependent on the specific infrastructure, software, and tools used to process data. Expenses may include those related to licensing, computing resources, and hiring necessary personnel.

- Data transformation processes can be resource-intensive. Performing transformations in an on-premises data warehouse after loading, or transforming data before feeding it into applications, can create a computational burden that slows down other operations.

- If you use a cloud-based data warehouse, you can do the transformations after loading because the platform can scale up to meet demand.

Data transformation can increase the efficiency of analytic and business processes and enable better data-driven decision-making. The first phase of data transformations should include things like data type conversion and flattening of hierarchical data. These operations shape data to increase compatibility with analytics systems. Data analysts and data scientists can implement further transformations additively as necessary as individual layers of processing. Each layer of processing should be designed to perform a specific set of tasks that meet a known business or technical requirement.

Video Content / Details of website for further learning (if any):
https://www.stitchdata.com/resources/data-transformation/

Important Books/Journals for further learning including the page nos.:

31. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.

32. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L-17

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : II - Data Collection and Data Pre-Processing Date of Lecture:

Topic of Lecture: Data Reduction

Introduction :

Data reduction refers to the process of selecting, focusing, simplifying, abstracting, and transforming the data that appear in written up field notes or transcriptions.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Integration
- Data Transformation

Detailed content of the Lecture:

Data reduction is a process that reduced the volume of original data and represents it in a much smaller volume. Data reduction techniques ensure the integrity of data while reducing the data. The time required for data reduction should not overshadow the time saved by the data mining on the reduced data set. In this section, we will discuss data reduction in brief and we will discuss different methods of data reduction.

The time required for data reduction should not overshadow the time saved by the data mining on the reduced data set. In this section, we will discuss data reduction in brief and we will discuss different methods of data reduction.

When you collect data from different data warehouses for analysis, it results in a huge amount of data. It is difficult for a data analyst to deal with this large volume of data.This is why reducing data becomes important. Data reduction technique reduces the volume of data yet maintains the integrity of the data.

Data reduction does not affect the result obtained from data mining that means the result obtained from data mining before data reduction and after data reduction is the same (or almost the same).
The only difference occurs in the efficiency of data mining. Data reduction increases the efficiency of data mining. In the following section, we will discuss the techniques of data reduction.

Data Reduction Techniques

Techniques of data deduction include dimensionality reduction, numerosity reduction and data compression.

1. Dimensionality Reduction

Dimensionality reduction **eliminates the attributes** from the data set under consideration thereby reducing the volume of original data. In the section below, we will discuss three methods of dimensionality reduction.

a. Wavelet Transform

In the wavelet transform, a data vector X is transformed into a numerically different data vector X' such that both X and X' vectors are of the same length. Then how it is useful in reducing data? The data obtained from the wavelet transform can be truncated. The compressed data is obtained by retaining the smallest fragment of the strongest of wavelet coefficients.Wavelet transform can be applied to data cube, sparse data or skewed data.

b. Principal Component Analysis

Let us consider we have a data set to be analyzed that has tuples with n attributes, then the principal component analysis identifies k independent tuples with n attributes that can represent the data set.

In this way, the original data can be cast on a much smaller space. In this way, the dimensionality reduction can be achieved. Principal component analysis can be applied to sparse, and skewed data.

c. Attribute Subset Selection

The large data set has many attributes some of which are irrelevant to data mining or some are redundant. The attribute subset selection makes it sure that even after eliminating the unwanted attributes we get a good subset of original attributes such that the resulting probability of data distribution is as close as possible to the original data distribution using all the attributes.

2. Numerosity Reduction

The numerosity reduction reduces the volume of the original data and represents it in a much smaller form. This technique includes two types parametric and non-parametric numerosity reduction.

Video Content / Details of website for further learning (if any):
https://binaryterms.com/data-reduction.html

Important Books/Journals for for further learning including the page nos.:

33. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.

34. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L-18

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : II - Data Collection and Data Pre-Processing Date of Lecture:

Topic of Lecture: Data Discretization

Introduction :

Data Discretization is the process through which we can transform continuous variables, models or functions into a discrete form. We do this by creating a set of contiguous intervals (or bins) that go across the range of our desired variable/model/function.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Transformation
- Data Reduction

Detailed content of the Lecture:

Data discretization is the process of converting continuous data into discrete buckets by grouping it. Discretization is also known for easy maintainability of the data. Training a model with discrete data becomes faster and more effective than when attempting the same with continuous data. Although continuous-valued data contains more information, huge amounts of data can slow the model down. Here, discretization can help us strike a balance between both. Some famous methods of data discretization are binning and using a histogram. Although data discretization is useful, we need to effectively pick the range of each bucket, which is a challenge.

Data discretization refers to a method of converting a huge number of data values into smaller ones so that the evaluation and management of data become easy. In other words, data discretization is a method of converting attributes values of continuous data into a finite set of intervals with minimum data loss. There are two forms of data discretization first is supervised discretization, and the second is unsupervised discretization. Supervised discretization refers to a method in which the class data is used. Unsupervised discretization refers to a method depending upon the way which operation proceeds. It means it works on the top-down splitting strategy and bottom-up merging strategy.

Some Famous techniques of data discretization

Histogram analysis

Histogram refers to a plot used to represent the underlying frequency distribution of a continuous data set. Histogram assists the data inspection for data distribution. For example, Outliers, skewness representation, normal distribution representation, etc.

Binning

Binning refers to a data smoothing technique that helps to group a huge number of continuous values into smaller values. For data discretization and the development of idea hierarchy, this technique can also be used.

Cluster Analysis

Cluster analysis is a form of data discretization. A clustering algorithm is executed by dividing the values of x numbers into clusters to isolate a computational feature of x.

Data discretization using decision tree analysis

Data discretization refers to a decision tree analysis in which a top-down slicing technique is used. It is done through a supervised procedure. In a numeric attribute discretization, first, you need to select the attribute that has the least entropy, and then you need to run it with the help of a recursive process. The recursive process divides it into various discretized disjoint intervals, from top to bottom, using the same splitting criterion.

Data discretization using correlation analysis

Discretizing data by linear regression technique, you can get the best neighboring interval, and then the large intervals are combined to develop a larger overlap to form the final 20 overlapping intervals. It is a supervised procedure.

Video Content / Details of website for further learning (if any):
https://www.javatpoint.com/discretization-in-data-mining

Important Books/Journals for further learning including the page nos.:

35. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.

36. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-19

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : III - Exploratory Data Analytics Date of Lecture:

Topic of Lecture: Descriptive Statistics

Introduction :

Descriptive statistics are **used to describe the basic features of the data in a study**. They provide simple summaries about the sample and the measures. Together with simple graphics analysis, they form the basis of virtually every quantitative analysis of data.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Reduction
- Data Discretization

Detailed content of the Lecture:

Descriptive Statistics is summarizing the data at hand through certain numbers like mean, median etc. so as to make the understanding of the data easier. It does not involve any generalization or inference beyond what is available. This means that the descriptive statistics are just the representation of the data (sample) available and not based on any theory of probability.

Commonly Used Measures

- Measures of Central Tendency
- Measures of Dispersion (or Variability)

Measures of Central Tendency

A Measure of Central Tendency is a one number summary of the data that typically describes the center of the data. These one number summary is of three types.

Mean : Mean is defined as the ratio of the sum of all the observations in the data to the total number of observations. This is also known as Average. Thus mean is a number around which the entire data set is spread.

Median : Median is the point which divides the entire data into two equal halves. One-half of the data is less than the median, and the other half is greater than the same. Median is calculated by first arranging the data in either ascending or descending order.

3. Mode : Mode is the number which has the maximum frequency in the entire data set, or in other words,mode is the number that appears the maximum number of times. A data can have one or more than one mode.

If there is only one number that appears maximum number of times, the data has one mode, and is called **Uni-modal**.

If there are two numbers that appear maximum number of times, the data has two modes, and is called **Bi-modal**.

If there are more than two numbers that appear maximum number of times, the data has more than two modes, and is called **Multi-modal**.

Measures of Dispersion (or Variability)

Measures of Dispersion describes the spread of the data around the central value (or the Measures of Central Tendency)

1.Absolute Deviation from Mean — The Absolute Deviation from Mean, also called Mean Absolute Deviation (MAD), describe the variation in the data set, in sense that it tells the average absolute distance of each data point in the set.

2. Variance — Variance measures how far are data points spread out from the mean. A high variance indicates that data points are spread widely and a small variance indicates that the data points are closer to the mean of the data set.

3. Standard Deviation — The square root of Variance is called the Standard Deviation.

4. Range — Range is the difference between the Maximum value and the Minimum value in the data set.

5. Quartiles — Quartiles are the points in the data set that divides the data set into four equal parts. Q1, Q2 and Q3 are the first, second and third quartile of the data set.

6. Skewness — The measure of asymmetry in a probability distribution is defined by Skewness. It can either be positive, negative or undefined.

7. Kurtosis — Kurtosis describes the whether the data is light tailed (lack of outliers) or heavy tailed (outliers present) when compared to a Normal distribution. There are three kinds of Kurtosis:

Video Content / Details of website for further learning (if any):

https://towardsdatascience.com/descriptive-statistics-f2beeaf7a8df

Important Books/Journals for further learning including the page nos.:

37. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

38. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-20

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : III - Exploratory Data Analytics Date of Lecture:

Topic of Lecture: Mean

Introduction :

Mean is defined as the ratio of the sum of all the observations in the data to the total number of observations. This is also known as Average. Thus mean is a number around which the entire data set is spread.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Data Discretization
- Descriptive Statistics

Detailed content of the Lecture:

Descriptive statistics involves describing, summarizing and organizing the data so it can be easily understood. Graphical displays are often used along with the quantitative measures to enable clarity of communication.

- Methods of describing the characteristics of a data set.
- Useful because they allow you to make sense of the data.
- Helps exploring and making conclusions about the data in order to make rational decisions.
- Includes calculating things such as the average of the data, its spread and the shape it produces.

The following measures are used to describe a data set:

- Measures of position (also referred to as central tendency or location measures).
- Measures of spread (also referred to as variability or dispersion measures).
- Measures of shape.

Measures of Position:

- Position Statistics measure the data central tendency.
- Central tendency refers to where the data is centered.
- You may have calculated an average of some kind.

Despite the common use of average, there are different statistics by which we can describe the average of a data set:

- Mean
- Median
- Mode

Mean :

Mean is defined as the ratio of the sum of all the observations in the data to the total number of observations. This is also known as Average. Thus mean is a number around which the entire data set is spread.

Mean is sum of observed values in a data divided by the numberof observations.

- The total of all the values divided by the size of the data set.
- It is the most commonly used statistic of position.
- It is easy to understand and calculate.
- It works well when the distribution is symmetric and there are no outliers.
- The mean of a sample is denoted by ‘x-bar’.
- The mean of a population is denoted by ‘μ’.

Video Content / Details of website for further learning (if any):
https://incois.gov.in/documents/ITCOocean/C4_descriptive%20statistics.pdf

Important Books/Journals for further learning including the page nos.:

39. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

40. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-21

AI&DS

II/III

Course Name with Code : **19ADC05/ Introduction to Data Science**

Course Teacher : **Dr.P.Srinivasan**

Unit : **III - Exploratory Data Analytics** **Date of Lecture:**

Topic of Lecture: Standard Deviation

Introduction :

The square root of Variance is called the Standard Deviation. The average distance of the data points from their own mean. A low standard deviation indicates that the data points are clustered around the mean.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Descriptive Statistics
- Mean

Detailed content of the Lecture:

Descriptive statistics involves describing, summarizing and organizing the data so it can be easily understood. Graphical displays are often used along with the quantitative measures to enable clarity of communication.

- Methods of describing the characteristics of a data set.
- Useful because they allow you to make sense of the data.
- Helps exploring and making conclusions about the data in order to make rational decisions.
- Includes calculating things such as the average of the data, its spread and the shape it produces.

The following measures are used to describe a data set:

- Measures of position (also referred to as central tendency or location measures).
- Measures of spread (also referred to as variability or dispersion measures).
- Measures of shape.

Standard Deviation

- The average distance of the data points from their own mean.
- A low standard deviation indicates that the data points are clustered around the mean.
- A large standard deviation indicates that they are widely scattered around the mean.

- The standard deviation of a sample is denoted by ‘s’.
- The standard deviation of a population is denoted by “µ”.

Standard deviation is the measurement of the average distance between each quantity and mean. That is, how data is spread out from the mean. A low standard deviation indicates that the data points tend to be close to the mean of the data set, while a high standard deviation indicates that the data points are spread out over a wider range of values.There are situations when we have to choose between sample or population Standard Deviation.When we are asked to find SD of some part of a population, a segment of population; then we use sample Standard Deviation.

$$\text{S.D.} = \sqrt{\frac{1}{n-1}\sum_{i=0}^{n}(x-\bar{x})^{2}}$$

where x̅ is mean of a sample.But when we have to deal with a whole population, then we use population Standard Deviation.

$$\text{S.D.} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x-\mu)^{2}}$$

where µ is mean of a population.

Video Content / Details of website for further learning (if any):
https://incois.gov.in/documents/ITCOocean/C4_descriptive%20statistics.pdf

Important Books/Journals for further learning including the page nos.:

41. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.
42. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-22

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : III - Exploratory Data Analytics Date of Lecture:

Topic of Lecture: Skewness and Kurtosis

Introduction :

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. Kurtosis is a measure of whether the data are heavy-tailed (profusion of outliers) or light-tailed (lack of outliers) relative to a normal distribution.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Mean
- Standard Deviation

Detailed content of the Lecture:

Skewness

Skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable about its mean. The skewness value can be positive or negative, or undefined.

In a perfect normal distribution, the tails on either side of the curve are exact mirror images of each other.When a distribution is skewed to the left, the tail on the curve's left-hand side is longer than the tail on the right-hand side, and the mean is less than the mode. This situation is also called negative skewness.

When a distribution is skewed to the right, the tail on the curve's right-hand side is longer than the tail on the left-hand side, and the mean is greater than the mode. This situation is also called positive skewness. The direction of skewness is given by the sign. A zero means no skewness at all.

A negative value means the distribution is negatively skewed. A positive value means the distribution is positively skewed.

The coefficient compares the sample distribution with a normal distribution. The larger the value, the larger the distribution differs from a normal distribution.

Kurtosis

The exact interpretation of the measure of Kurtosis used to be disputed but is now settled. It's about the existence of outliers. Kurtosis is a measure of whether the data are heavy-tailed (profusion of outliers) or light-tailed (lack of outliers) relative to a normal distribution

There are three types of Kurtosis

Mesokurtic

Mesokurtic is the distribution that has similar kurtosis as normal distribution kurtosis, which is zero.

Leptokurtic

Distribution is the distribution that has kurtosis greater than a Mesokurtic distribution. Tails of such distributions are thick and heavy. If the curve of distribution is more peaked than the Mesokurtic curve, it is referred to as a Leptokurtic curve.

Platykurtic

Distribution is the distribution that has kurtosis lesser than a Mesokurtic distribution. Tails of such distributions thinner. If a curve of a distribution is less peaked than a Mesokurtic curve, it is referred to as a Platykurtic curve.

The main difference between skewness and kurtosis is that the skewness refers to the degree of symmetry, whereas the kurtosis refers to the degree of presence of outliers in the distribution.

Video Content / Details of website for further learning (if any):
https://towardsdatascience.com/understanding-descriptive-statistics-c9c2b0641291

Important Books/Journals for further learning including the page nos.:

43. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.

44. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

Estd. 2000

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

LECTURE HANDOUTS **L-23**

AI&DS **II/III**

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : III - Exploratory Data Analytics Date of Lecture:

Topic of Lecture: Box Plots

Introduction :

A boxplot is a graph that gives you a good indication of how the values in the data are spread out.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Standard Deviation
- Skewness and Kurtosis

Detailed content of the Lecture:

A Box and Whisker Plot (or Box Plot) is a convenient way of visually displaying the data distribution through their quartiles. It is a graphical rendition of statistical data based on the minimum, first quartile, median, third quartile, and maximum. The term “box plot” comes from the fact that the graph looks like a rectangle with lines extending from the top and bottom. Because of the extending lines, this type of graph is sometimes called a box-and-whisker plot.

Let us understand these 5 components of the box plot

Median Value

Value or quantity that falls halfway between a set of values arranged in an ascending or descending order. When the set contains an odd number of values, the median value is exactly in middle. If the number of values is even, the median is computed by averaging the two numbers closest to the middle.

Lower Quartile(Q1)

The lower quartile is also known as the first quartile, splits the lower 25% of the data. Quartiles are three points that divide the data set into four equal groups. Each group represents the one-fourth of the data set. The lower quartile is the middle value of the lower half.

Upper Quartile(Q3)

Upper quartile is also known as the third quartile. It splits lowest 75% (or highest 25%) of data. It can be also seen as the middle value of the upper half.

Interquartile Range(Q3-Q1)

The Interquartile range is from Q1 to Q3. It is the difference between the lower quartile and upper quartile. The IQR is often seen as a better measure of a spread than the range (highest value-lowest value) as it is not affected by outliers.

Highest Value

This point in the box plot represents the highest value in the data distribution over which the box plot is built which is not an outlier. This point does not correspond to the highest value in your dataset. Suppose you have some data like 65,76,87,100,105,100000. Here the largest value is 100000 but it is most likely to be an outlier and hence the box plot will not mark this as the maximum value. The most feasible option will be 105 as the maximum value of the box plot.

Lowest Value

This point in the box plot represents the lowest value in the data distribution over which the box plot is built and is not an outlier (smallest value in the Interquartile range of the distribution). This point does not correspond to the smallest value in your dataset. Suppose you have some data like 0.005,65,76,87,100,105. Here the smallest value is 0.005 but it is most likely to be an outlier and hence the box plot will not mark this as the minimum value. The most feasible option will be 65 as the minimum value of the box plot.

Video Content / Details of website for further learning (if any):
https://dimensionless.in/what-is-a-box-plot/

Important Books/Journals for further learning including the page nos.:

45. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

46. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-24

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : III - Exploratory Data Analytics Date of Lecture:

Topic of Lecture: Pivot Table

Introduction :

A pivot table is a similar operation that is commonly seen in spreadsheets and other programs that operate on tabular data. The pivot table takes simple column-wise data as input, and groups the entries into a two-dimensional table that provides a multidimensional summarization of the data.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Skewness and Kurtosis
- Box Plots

Detailed content of the Lecture:

Pivot Table is everywhere. In fact, we have built Pivot Table a few times for various analytics products including BI when we used to work at Oracle. We found that people were using Pivot Table more often than any other chart types, because it's just easy to summarize data in a 'table' format, not 'chart' format, and to see the exact numbers. With Exploratory Desktop, you can always run a few dplyr+tidyr commands to present your data in 'Pivot Table' view, but it is a bit too much especially when you can use other tools like Excel to do it quickly.

Pivot Table makes it super easy to not only summarize (aggregate) data but also spot outliers or patterns quickly by using color. And, just like any other visualization (chart) types you can share it with reproducible data preparation steps simply by clicking a button and start having a conversation around the data.

Here are some highlights of Pivot Table that I'm going to walk you through quickly.

- Pivot Table Basic
- Pivot Table with Color
- Pivot Table with Date / Time Data

- Sharing Pivot Table

1.Pivot Table Basic

Pivot Table can be found under Viz Type under Viz tab. We have renamed it from Chart view to Viz view for an obvious reason. ;)

2. Pivot Table with Color

You can also use color to format the Pivot Table cells, which would make it easier to spot the patterns or trends in the data.

3. Pivot Table with Date Aggregation

When you assign columns with Date / Time data type, you can do something more. As you would expect with other Viz types (Bar, Line, etc.) you can set the Date / Time aggregation level here as well.

4. Sharing Pivot Table

And just like any other Visualization types, you can share it with the reproducible data preparation steps.

Video Content / Details of website for further learning (if any):
https://blog.exploratory.io/introducing-pivot-table-1c9c949fd2d6

Important Books/Journals for further learning including the page nos.:

47. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

48. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-25

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : III - Exploratory Data Analytics Date of Lecture:

Topic of Lecture: Heat Map

Introduction :

A heatmap is a graphical representation of data in which data values are represented as colors. That is, it uses color in order to communicate a value to the reader.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Box Plots
- Pivot Table

Detailed content of the Lecture:

A heat map is data analysis software that uses color the way a bar graph uses height and width: as a data visualization tool.If you’re looking at a web page and you want to know which areas get the most attention, a heat map shows you in a visual way that’s easy to assimilate and make decisions from.

A heat map uses a warm-to-cool color spectrum to show you which parts of a page receive the most attention.

This heat map, for example, shows how far down the page visitors have scrolled:

With a heat map, the data for your web page is right there: the CTA above the fold glows bright orange, or it doesn’t.If it doesn’t, consider moving it higher up the page so it aligns with where visitors are paying the most attention.

Heat mapping software works by collecting data from a web page and displaying that data over the web page itself. First, we take a snapshot of the web page at the URL you select. When the HTML of the page loads, versions of what’s loaded are sent to our servers too, thanks to a short section of JavaScript inserted into your site’s code Then we create a map of all the elements on your page – everything that anybody might interact with. These and their tags and their parent elements are what we use to build a map of user activity on your page.

Next, we collect all the activity data. Every time a user does something on your web page, we flag it.(We also use a piece of code to make sure we know it’s unique users, so you don’t get duplicate or multiple users messing with your figures.)

There’s a lot going on under the hood to make this happen, but the whole point of a heat map is that you don’t need to necessarily know about this to get value from it.Suppose someone comes to your page and clicks on your CTA. We’ll record that click and add it to all the other information we have about that page.

A heat map analysis gives you a visual overview of where your visitors click on your page — the more clicks, the brighter the area, creating what we call “hotspots.”This means that when you look at your heat map, you can quickly see which areas of the page get a lot of action and which don’t.

Video Content / Details of website for further learning (if any):
https://www.crazyegg.com/blog/understanding-using-heatmaps-studies/

Important Books/Journals for further learning including the page nos.:

49. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.
50. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-26

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher : Dr.P.Srinivasan

Unit : III - Exploratory Data Analytics Date of Lecture:

Topic of Lecture: Correlation Statistics

Introduction :

Correlation is used to find the relationship between two variables which is important in real life because we can predict the value of one variable with the help of other variables, who is being correlated with it.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Pivot Table
- Heat Map

Detailed content of the Lecture:

Correlation is used to find the relationship between two variables which is important in real life because we can predict the value of one variable with the help of other variables, who is being correlated with it. It is a type of Bivariate statistics since two variables are involved here. It is a **statistical technique** that helps us to analyze the relationship between two or more variables.

Correlation: It is a numerical measure of the direction and magnitude of the mutual relationship between the variables(X and Y).

It may happen because of several reasons like:

1. Mutual dependence Between the variables: Both the variables may be mutually influencing each other so that neither can be designated as the cause and the other the effect.

When two variables(X and Y) affect each other mutually, we cannot say X is the cause or Y is the cause.

For Example, The price of a commodity is affected by demand and supply.

2. Due to pure chance: In a small sample, X and Y are highly correlated but in the universe X and Y are not correlated.

For Example, Correlation between income and weight of a person. This may be due to:

– Sampling fluctuations

– Bias of investigator in selecting the sample.

Such a relation is called a **non-sense or spurious relation.**

3. Correlation due to any third common factor: Both the correlated variables may be influenced by one or other variables.

– X and Y don’t have a direct correlation.

For Example, It is between the production of tea and rice per hectare. Here they are not directly correlated instead the cause is the good rainfall well in time.

Utility of Correlation

1. It is very useful for Economists to study the relationships between variables.

2. It helps in measuring the degree of relationship between the variables.

3. We can also test the significance of the relationship.

4. Sampling error can also be calculated by knowing the correlation.

5. It is the basis for the study of regression.

6. Estimate the value of one variable based on the other variable.

7. It is used to determine the relationship between datasets in business.

Based on the degree of correlation:

1. Positive correlation: It is said to be positive when the values of the two variables move in the same direction so that an increase in one variable is followed by an increase in the other variable or a decrease in one variable is followed by a decrease in the other variable.

Two variables X and Y are going in the same direction.

If X rises, Y also rises, and vice-versa.

Examples of positive correlation are (a) Age and Income, (b) Amount of rainfall, and the yield of the crop.

2. Negative correlation: It is said to be negative when the values of the two variables move in the opposite direction so that an increase in one variable is followed by a decrease in the other variable.

Two variables X and Y are going in the opposite direction.

If X rises, Y falls, and vice versa.

Examples of negative correlation are (a) Height above sea level and temperature, (b) Sales of woolen clothes and temperature.

Video Content / Details of website for further learning (if any):

https://www.analyticsvidhya.com/blog/2021/04/intuition-behind-correlation-definition-and-its-types/

Important Books/Journals for further learning including the page nos.:

51. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

52. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-27

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : III - Exploratory Data Analytics Date of Lecture:

Topic of Lecture: ANOVA

Introduction :

ANOVA helps you find out whether the differences between groups of data are statistically significant. It works by analyzing the levels of variance within the groups through samples taken from each of them.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Heat Map
- Correlation Statistics

Detailed content of the Lecture:

A common approach to figure out a reliable treatment method would be to analyse the days it took the patients to be cured. We can use a statistical technique which can compare these three treatment samples and depict how different these samples are from one another. Such a technique, which compares the samples on the basis of their means, is called ANOVA.

Analysis of variance (ANOVA) is a statistical technique that is used to check if the means of two or more groups are significantly different from each other. ANOVA checks the impact of one or more factors by comparing the means of different samples. We can use ANOVA to prove/disprove if all the medication treatments were equally effective or not.

Another measure to compare the samples is called a t-test. When we have only two samples, t-test and ANOVA give the same results. However, using a t-test would not be reliable in cases where there are more than 2 samples. If we conduct multiple t-tests for comparing more than two samples, it will have a compounded effect on the error rate of the result.

Before we get started with the applications of ANOVA, I would like to introduce some common terminologies used in the technique.

Grand Mean

Mean is a simple or arithmetic average of a range of values. There are two kinds of means that we use in ANOVA calculations, which are separate sample means $(\mu_1, \mu_2 \, \& \, \mu_3)$ and the grand mean (μ). . The grand mean is the mean of sample means or the mean of all observations combined, irrespective of the sample.

Video Content / Details of website for further learning (if any):
https://www.analyticsvidhya.com/blog/2018/01/anova-analysis-of-variance/

Important Books/Journals for further learning including the page nos.:

53. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.
54. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L-28

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : IV - Model Development Date of Lecture:

Topic of Lecture: Simple and Multiple Regression

Introduction :

Simple linear regression has only one x and one y variable. Multiple linear regression has one y and two or more x variables. For instance, when we predict rent based on square feet alone that is simple linear regression.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Correlation Statistics
- ANOVA

Detailed content of the Lecture:

Simple Linear Regression

Simple linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.

The core idea is to obtain a line that best fits the data. The best fit line is the one for which total prediction error (all data points) are as small as possible. Error is the distance between the point to the regression line.

Co-efficient from Normal equations

Apart from above equation co-efficient of the model can also be calculated from normal equation.

$$\text{Theta} = (X^T X)^{-1} X^T Y$$

Theta contains co-efficient of all predictors including constant term 'b0'. Normal equation performs computation by taking inverse of input matrix. Complexity of the computation will increase as the number of features increase. It gets very slow when number of features grow large.

Optimizing using gradient descent

Complexity of the normal equation makes it difficult to use, this is where gradient descent method comes into picture. Partial derivative of the cost function with respect to the parameter can give optimal co-efficient value.

Residual Analysis

Randomness and unpredictability are the two main components of a regression model.

Prediction = Deterministic + Statistic

Deterministic part is covered by the predictor variable in the model. Stochastic part reveals the fact that the expected and observed value is unpredictable. There will always be some information that are missed to cover. This information can be obtained from the residual information.

Video Content / Details of website for further learning (if any):
https://towardsdatascience.com/linear-regression-detailed-view-ea73175f6e86

Important Books/Journals for further learning including the page nos.:

55. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.

56. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-29

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : IV - Model Development Date of Lecture:

Topic of Lecture: Model Evaluation using Visualization

Introduction :

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future.

Prerequisite knowledge for Complete understanding and learning of Topic:

- ANOVA
- Simple and Multiple Regression

Detailed content of the Lecture:

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future. Evaluating model performance with the data used for training is not acceptable in data science because it can easily generate overoptimistic and overfitted models. There are two methods of evaluating models in data science, Hold-Out and Cross-Validation. To avoid overfitting, both methods use a test set (not seen by the model) to evaluate model performance.

In this method, the mostly large dataset is randomly divided to three subsets:

1. **Training set** is a subset of the dataset used to build predictive models.
2. **Validation set** is a subset of the dataset used to assess the performance of model built in the training phase. It provides a test platform for fine tuning model's parameters and selecting the best-performing model. Not all modeling algorithms need a validation set.
3. **Test set** or unseen examples is a subset of the dataset to assess the likely future performance of a model. If a model fit to the training set much better than it fits the test set, overfitting is probably the cause.

Cross-Validation

When only a limited amount of data is available, to achieve an unbiased estimate of the model performance we use k-fold cross-validation. In k-fold cross-validation, we divide the data into k subsets of equal size. We build models k times, each time leaving out one of the subsets from training and use it as the test set. If k equals the sample size, this is called "leave-one-out".

Model evaluation can be divided to two sections:

- Classification Evaluation
- Regression Evaluation

Video Content / Details of website for further learning (if any):
https://www.saedsayad.com/model_evaluation.htm

Important Books/Journals for further learning including the page nos.:

57. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.

58. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-30

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : IV - Model Development Date of Lecture:

Topic of Lecture: Model Evaluation using Visualization

Introduction :

Model Evaluation is an integral part of the model development process. It helps to find the best model that represents our data and how well the chosen model will work in the future.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Simple and Multiple Regression
- Model Evaluation using Visualization

Detailed content of the Lecture:

Confusion Matrix

A confusion matrix shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data. The matrix is NxN, where N is the number of target values (classes). Performance of such models is commonly evaluated using the data in the matrix. The following table displays a 2x2 confusion matrix for two classes (Positive and Negative).

- **Accuracy** : the proportion of the total number of predictions that were correct.
- **Positive Predictive Value** or **Precision** : the proportion of positive cases that were correctly identified.
- **Negative Predictive Value** : the proportion of negative cases that were correctly identified.
- **Sensitivity** or **Recall** : the proportion of actual positive cases which are correctly identified.
- **Specificity** : the proportion of actual negative cases which are correctly identified.

Gain and Lift Charts

Gain or lift is a measure of the effectiveness of a classification model calculated as the ratio between the results obtained with and without the model. Gain and lift charts are visual aids for evaluating performance of classification models. However, in contrast to the confusion matrix that evaluates models on the whole population gain or lift chart evaluates model performance in a portion of the population.

Lift Chart

The lift chart shows how much more likely we are to receive positive responses than if we contact a random sample of customers. For example, by contacting only 10% of customers based on the predictive model we will reach 3 times as many respondents, as if we use no model.

K-S Chart

K-S or Kolmogorov-Smirnov chart measures performance of classification models. More accurately, K-S is a measure of the degree of separation between the positive and negative distributions. The K-S is 100 if the scores partition the population into two separate groups in which one group contains all the positives and the other all the negatives. On the other hand, If the model cannot differentiate between positives and negatives, then it is as if the model selects cases randomly from the population. The K-S would be 0. In most classification models the K-S will fall between 0 and 100, and that the higher the value the better the model is at separating the positive from negative cases

ROC Chart

The ROC chart is similar to the gain or lift charts in that they provide a means of comparison between classification models. The ROC chart shows false positive rate (1-specificity) on X-axis, the probability of target=1 when its true value is 0, against true positive rate (sensitivity) on Y-axis, the probability of target=1 when its true value is 1. Ideally, the curve will climb quickly toward the top-left meaning the model correctly predicted the cases. The diagonal red line is for a random model (ROC101).

Area Under the Curve (AUC)

Area under ROC curve is often used as a measure of quality of the classification models. A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1. In practice, most of the classification models have an AUC between 0.5 and 1.

Video Content / Details of website for further learning (if any):
https://www.saedsayad.com/model_evaluation.htm

Important Books/Journals for further learning including the page nos.:

59. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.

60. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-31

AI&DS

II/III

Course Name with Code : **19ADC05/ Introduction to Data Science**

Course Teacher :**Dr.P.Srinivasan**

Unit : **IV - Model Development** **Date of Lecture:**

Topic of Lecture: Residual Plot

Introduction :

A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a nonlinear model is more appropriate.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Model Evaluation using Visualization
- Simple and Multiple Regression

Detailed content of the Lecture:

Residuals in a statistical or machine learning model are the differences between observed and predicted values of data. They are a diagnostic measure used when assessing the quality of a model. They are also known as errors.

Residuals are important when determining the quality of a model. You can examine residuals in terms of their magnitude and/or whether they form a pattern.

Where the residuals are all 0, the model predicts perfectly. The further residuals are from 0, the less accurate the model. In the case of linear regression, the greater the sum of squared residuals, the smaller the R-squared statistic, all else being equal.

Where the average residual is not 0, it implies that the model is systematically biased (i.e., consistently over- or under-predicting).

Where residuals contain patterns, it implies that the model is qualitatively wrong, as it is failing to explain some property of the data. The existence of patterns invalidates most statistical tests.

Diagnosing problems by looking for patterns in residuals

If some of the residuals are relatively large compared to others, either the data or model may be flawed. The next step is to investigate and work out what has specifically led to the unusually large residual.

In the example above, the residuals for January and February are much further from 0 than the residuals for the other months. Thus, it may be worthwhile to investigate what was unusual about January. Unusually large residuals are called outliers or extreme values.

Another common type of pattern in residuals is when we can predict the value of residuals based on the preceding values of residuals. This is known variously as autocorrelation, serial correlation, and serial dependence. The residuals in this case to seem to have a snake-like pattern - evidence of autocorrelation.

Another type of pattern is where the degree of variation in the residuals seems to change over time. This pattern is known as heteroscedasticity. In the plot above, we can see some evidence of heteroscedasticity, with residuals in months 1 through 4 being further from 0 than the residuals from months 5 through 12.

Another type of pattern relates to the distribution of the residuals. In some situations, it can be informative to see if the residuals are distributed in accordance with the normal distribution.

Different types of residuals

Sometimes residuals are scaled (i.e., divided by a number) to make them easier to interpret. In particular, standardized and studentized residuals typically rescale the residuals so that values of more than 1.96 from 0 equate to a p-value of 0.05. Different software packages use terminology inconsistently. Also, most packages do not adjust for multiple comparison correction errors, so be careful when reading documentation using scaled residuals.

Video Content / Details of website for further learning (if any):
https://www.displayr.com/learn-what-are-residuals/

Important Books/Journals for further learning including the page nos.:

61. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.
62. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L-32

AI&DS

II/III

Course Name with Code : **19ADC05/ Introduction to Data Science**

Course Teacher :**Dr.P.Srinivasan**

Unit : **IV - Model Development** **Date of Lecture:**

Topic of Lecture: Distribution Plot

Introduction :

Distribution plots visually assess the distribution of sample data by comparing the empirical distribution of the data with the theoretical values expected from a specified distribution.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Model Evaluation using Visualization
- Residual Plot

Detailed content of the Lecture:

Distribution plots visually assess the distribution of sample data by comparing the empirical distribution of the data with the theoretical values expected from a specified distribution. Use distribution plots in addition to more formal hypothesis tests to determine whether the sample data comes from a specified distribution. To learn about hypothesis tests, see Hypothesis Testing.

Statistics and Machine Learning Toolbox™ offers several distribution plot options:

Normal Probability Plots — Use `normplot` to assess whether sample data comes from a normal distribution. Use `probplot` to create Probability Plots for distributions other than normal, or to explore the distribution of censored data.

Quantile-Quantile Plots — Use `qqplot` to assess whether two sets of sample data come from the same distribution family. This plot is robust with respect to differences in location and scale.

Cumulative Distribution Plots — Use `cdfplot` or `ecdf` to display the empirical cumulative distribution function (cdf) of the sample data for visual comparison to the theoretical cdf of a specified distribution.

Normal Probability Plots

We can use normal probability plots to assess whether data comes from a normal distribution. Many statistical procedures make the assumption that an underlying distribution is normal. Normal probability plots can provide some assurance to justify this assumption or provide a warning of problems with the assumption. An analysis of normality typically combines normal probability plots with hypothesis tests for normality.

Quantile-Quantile Plots

Use quantile-quantile (q-q) plots to determine whether two samples come from the same distribution family. Q-Q plots are scatter plots of quantiles computed from each sample, with a line drawn between the first and third quartiles. If the data falls near the line, it is reasonable to assume that the two samples come from the same distribution. The method is robust with respect to changes in the location and scale of either distribution.

Create a quantile-quantile plot by using the `qqplot` function.

Cumulative Distribution Plots

An empirical cumulative distribution function (cdf) plot shows the proportion of data less than or equal to each *x* value, as a function of *x*. The scale on the *y*-axis is linear; in particular, it is not scaled to any particular distribution. Empirical cdf plots are used to compare data cdfs to cdfs for particular distributions. To create an empirical cdf plot, use the `cdfplot` function or the `ecdf` function.

Video Content / Details of website for further learning (if any):

https://www.mathworks.com/help/stats/distribution-plots.html

Important Books/Journals for further learning including the page nos.:

63. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

64. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L-33

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher : Dr.P.Srinivasan

Unit : IV - Model Development Date of Lecture:

Topic of Lecture: Polynomial Regression

Introduction :

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Residual Plot
- Distribution Plot

Detailed content of the Lecture:

Polynomial Regression is a regression algorithm that models the relationship between a dependent(y) and independent variable(x) as nth degree polynomial. The Polynomial Regression equation is given below:

$$y= b_0+b_1x_1+ b_2x_1^2+ b_2x_1^3+\ldots\ldots\ b_nx_1^n$$

- It is also called the special case of Multiple Linear Regression in ML. Because we add some polynomial terms to the Multiple Linear regression equation to convert it into Polynomial Regression.
- It is a linear model with some modification in order to increase the accuracy.
- The dataset used in Polynomial regression for training is of non-linear nature.
- It makes use of a linear regression model to fit the complicated and non-linear functions and datasets.
- Hence, "In Polynomial regression, the original features are converted into Polynomial features of required degree (2,3,..,n) and then modeled using a linear model."

Need for Polynomial Regression:

The need of Polynomial Regression in ML can be understood in the below points:

- If we apply a linear model on a **linear dataset**, then it provides us a good result as we have seen in

Simple Linear Regression, but if we apply the same model without any modification on a **non-linear dataset**, then it will produce a drastic output. Due to which loss function will increase, the error rate will be high, and accuracy will be decreased.

- So for such cases, **where data points are arranged in a non-linear fashion, we need the Polynomial Regression model**. We can understand it in a better way using the below comparison diagram of the linear dataset and non-linear dataset.
- In the above image, we have taken a dataset which is arranged non-linearly. So if we try to cover it with a linear model, then we can clearly see that it hardly covers any data point. On the other hand, a curve is suitable to cover most of the data points, which is of the Polynomial model.
- Hence, *if the datasets are arranged in a non-linear fashion, then we should use the Polynomial Regression model instead of Simple Linear Regression.*

Steps for Polynomial Regression:

The main steps involved in Polynomial Regression are given below:

- Data Pre-processing
- Build a Linear Regression model and fit it to the dataset
- Build a Polynomial Regression model and fit it to the dataset
- Visualize the result for Linear Regression and Polynomial Regression model.
- Predicting the output.

Video Content / Details of website for further learning (if any):
https://www.javatpoint.com/machine-learning-polynomial-regression

Important Books/Journals for further learning including the page nos.:

65. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

66. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-34

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher : Dr.P.Srinivasan

Unit : IV - Model Development Date of Lecture:

Topic of Lecture: Polynomial Pipelines

Introduction :

Pipelines sequentially apply a list of transformers and a final predictor (classifier or regressor). Intermediate steps of the pipeline must be 'transformers', that is, they must implement **fit()** and **transform()** methods.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Distribution Plot
- Polynomial Regression

Detailed content of the Lecture:

Building a machine learning pipeline

Scikit-learn refers to machine learning algorithms as **estimators**. There are three different types of estimators: **classifiers**, **regressors**, and **transformers**. The classifiers and regressors are called **predictors**.

As our analysis and workflow become more complicated, you may need to apply multiple transformations to your data before it is ready for a supervised machine learning model. Pipelines sequentially apply a list of transformers and a final predictor (classifier or regressor). Intermediate steps of the pipeline must be 'transformers', that is, they must implement **fit()** and **transform()** methods. The final predictor only needs to implement the **fit()** method.

In our workflow:

1. **StandardScaler()** is a transformer.

2. **PCA()** is a transformer.

3. **PolynomialFeatures()** is a transformer.

4. **LinearRegression()** is a predictor.

So, we can build a pipeline for our model using Scikit-learn **Pipeline()** class. It sequentially applies the above list of transformers and the final predictor. Here is the code.

By using pipelines, we can easily build complex models with less code!

Let’s call the **fit()** method of our pipeline.

When calling ***poly_reg_model.fit(X, y)***, the following process occurs:

X_scaled = sc.fit_transform(X)

X_pca = pca.fit_transform(X_scaled)

X_poly = poly_features.fit_transform(X_pca)

lin_reg.fit(X_poly, y)

Let’s call the predict() method of our pipeline.

When calling ***poly_reg_model.predict(X_new)***, the following process occurs:

X_new_scaled = sc.transform(X)

X_new_pca = pca.transform(X_scaled)

X_new_poly = poly_features.transform(X_pca)

lin_reg.predict(X_new_poly)

So, ***poly_reg_model.predict(X_new)*** returns predictions for our new data.

Video Content / Details of website for further learning (if any):
https://towardsdatascience.com/polynomial-regression-with-a-machine-learning-pipeline-7e27d2dedc87

Important Books/Journals for further learning including the page nos.:

67. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

68. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-35

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : IV - Model Development Date of Lecture:

Topic of Lecture: Measures for In-sample Evaluation

Introduction :

Methods for evaluating a model's performance are divided into 2 categories: namely, holdout and Cross-validation. Both methods use a test set (i.e data not seen by the model) to evaluate model performance.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Polynomial Regression
- Polynomial Pipelines

Detailed content of the Lecture:

Methods for evaluating a model's performance are divided into 2 categories: namely, holdout and Cross-validation. Both methods use a test set (i.e data not seen by the model) to evaluate model performance. It's not recommended to use the data we used to build the model to evaluate it. This is because our model will simply remember the whole training set, and will therefore always predict the correct label for any point in the training set. This is known as overfitting.

Holdout

The purpose of holdout evaluation is to test a model on different data than it was trained on. This provides an unbiased estimate of learning performance.

In this method, the dataset is *randomly* divided into three subsets:

Training set is a subset of the dataset used to build predictive models.

Validation set is a subset of the dataset used to assess the performance of the model built in the training phase. It provides a test platform for fine-tuning a model's parameters and selecting the best performing model. Not all modeling algorithms need a validation set.

Test set, or unseen data, is a subset of the dataset used to assess the likely future performance of a model. If a model fits to the training set much better than it fits the test set, overfitting is probably the cause.

The holdout approach is useful because of its speed, simplicity, and flexibility. However, this technique is often associated with high variability since differences in the training and test dataset can result in meaningful differences in the estimate of accuracy.

Cross-Validation

Cross-validation is a technique that involves partitioning the original observation dataset into a training set, used to train the model, and an independent set used to evaluate the analysis.

The most common cross-validation technique is k-fold cross-validation, where the original dataset is partitioned into k equal size subsamples, called folds. The k is a user-specified number, usually with 5 or 10 as its preferred value. This is repeated k times, such that each time, one of the k subsets is used as the test set/validation set and the other k-1 subsets are put together to form a training set. The error estimation is averaged over all k trials to get the total effectiveness of our model.

For instance, when performing five-fold cross-validation, the data is first partitioned into 5 parts of (approximately) equal size. A sequence of models is trained. The first model is trained using the first fold as the test set, and the remaining folds are used as the training set. This is repeated for each of these 5 splits of the data and the estimation of accuracy is averaged over all 5 trials to get the total effectiveness of our model. As can be seen, every data point gets to be in a test set exactly once and gets to be in a training set k-1 times. This significantly reduces bias, as we're using most of the data for fitting, and it also significantly reduces variance, as most of the data is also being used in the test set. Interchanging the training and test sets also adds to the effectiveness of this method.

Model Evaluation Metrics

Model evaluation metrics are required to quantify model performance. The choice of evaluation metrics depends on a given machine learning task (such as classification, regression, ranking, clustering, topic modeling, among others). Some metrics, such as precision-recall, are useful for multiple tasks. Supervised learning tasks such as classification and regression constitutes a majority of machine learning applications. In this article, we focus on metrics for these two supervised learning models.

Classification Metrics

In this section we will review some of the metrics used in classification problems, namely:

- Classification Accuracy
- Confusion matrix
- Logarithmic Loss
- Area under curve (AUC)
- F-Measure

Video Content / Details of website for further learning (if any):
https://heartbeat.comet.ml/introduction-to-machine-learning-model-evaluation-fa859e1b2d7f

Important Books/Journals for further learning including the page nos.:

69. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.

70. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L-36

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : IV - Model Development Date of Lecture:

Topic of Lecture: Prediction and Decision Making

Introduction :

Predictive-Decision Model a novel integration of prediction analytics with decision modeling, where predictions are optimized and decisions are predicted.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Polynomial Pipelines
- Measures for In-sample Evaluation

Detailed content of the Lecture:

Prediction refers to the output of an algorithm after it has been trained on a historical dataset and applied to new data when forecasting the likelihood of a particular outcome, such as whether or not a customer will churn in 30 days. The algorithm will generate probable values for an unknown variable for each record in the new data, allowing the model builder to identify what that value will most likely be.

The word “prediction” can be misleading. In some cases, it really does mean that you are predicting a future outcome, such as when you’re using machine learning to determine the next best action in a marketing campaign. Other times, though, the “prediction” has to do with, for example, whether or not a transaction that already occurred was fraudulent. In that case, the transaction already happened, but you’re making an educated guess about whether or not it was legitimate, allowing you to take the appropriate action.

Machine learning model predictions allow businesses to make highly accurate guesses as to the likely outcomes of a question based on historical data, which can be about all kinds of things – customer churn likelihood, possible fraudulent activity, and more. These provide the business with insights that result in tangible business value. For example, if a model predicts a customer is likely to churn, the business can target them with specific communications and outreach that will prevent the loss of that customer.

Predictive Analytics Examples

Predictive analytics helps teams in industries as diverse as finance, healthcare, pharmaceuticals, automotive, aerospace, and manufacturing.

Automotive – Breaking new ground with autonomous vehicles

Companies developing driver assistance technology and new autonomous vehicles use predictive analytics to analyze sensor data from connected vehicles and to build driver assistance algorithms.

Aerospace – Monitoring aircraft engine health
To improve aircraft up-time and reduce maintenance costs, an engine manufacturer created a real-time analytics application to predict subsystem performance for oil, fuel, liftoff, mechanical health, and controls.

Energy Production – Forecasting electricity price and demand
Sophisticated forecasting apps use models that monitor plant availability, historical trends, seasonality, and weather.

Financial Services – Developing credit risk models
Financial institutions use machine learning techniques and quantitative tools to predict credit risk.

Industrial Automation and Machinery – Predicting machine failures
A plastic and thin film producer saves 50,000 Euros monthly using a health monitoring and predictive maintenance application that reduces downtime and minimizes waste.

Medical Devices – Using pattern-detection algorithms to spot asthma and COPD
An asthma management device records and analyzes patients' breathing sounds and provides instant feedback via a smart phone app to help patients manage asthma and COPD.

Predictive analytics is the process of using data analytics to make predictions based on data. This process uses data along with analysis, statistics, and machine learning techniques to create a predictive model for forecasting future events.

Video Content / Details of website for further learning (if any):
https://www.mathworks.com/discovery/predictive-analytics.html

Important Books/Journals for further learning including the page nos.:

71. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

72. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L-37

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : V - Model Evaluation Date of Lecture:

Topic of Lecture: Generalization Error

Introduction :

Generalisation error in statistics is generally the out-of-sample error which is the measure of how accurately a model can predict values for previously unseen data.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Measures for In-sample Evaluation
- Prediction and Decision Making

Detailed content of the Lecture:

The gap between predictions and observed data is induced by **model inaccuracy**, **sampling error**, and **noise**. Some of the errors are reducible but some are not. Choosing the right algorithm and tuning parameters could improve model accuracy, but we will never be able to make our predictions 100% accurate.

An important way to understand generalization error is **bias-variance decomposition**.

Intuitively speaking, **bias** is the **error rate** in the world of big data. A model has a high bias when, for example, it fails to capture meaningful patterns in the data. Bias is measured by the differences between the *expected predicted values* and the *observed values*, in the dataset D when the prediction variables are at the level of x (X=x).

In contrast with bias, **variance** is an algorithm's **flexibility** to learn patterns in the observed data. **Variance** is the amount that an algorithm will change if a **different dataset** is used. A model is of high variance when, for instance, it tries too hard that it not only captures the pattern of meaningful features but also that the meaningless error (**overfitting**).

Interpretation

Bias measures the deviation between the expected output of our model and the real values, so it indicates the **fit of our model**.

Variance measures the amount that the outputs of our model will change if a different dataset is used. It is the impacts of using different datasets.

Noise is the irreducible error, the **lowest bound of generalization error** for the current task that any model will not be able to get rid of, indicating the difficulty of this task.

These 3 components above determine the model's ability to react to new unseen data rather than just the data that it was trained on.

Bias-Variance Tradeoff

Bias-Variance Tradeoff as a Function of Model Capacity
Generalization error could be measured by MSE. As the model capacity increases, the bias decreases as the model fits the training datasets better. However, the variance increases, as your model become sophisticated to fit more patterns of the current dataset, changing datasets (even if they come from the same distribution) would be impactful. As a data scientist, our challenge lies in finding the optimal capacity — where both bias and variance are low.

Video Content / Details of website for further learning (if any):
https://medium.com/@yixinsun_56102/understanding-generalization-error-in-machine-learning-e6c03b203036

Important Books/Journals for further learning including the page nos.:

73. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.

74. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS **L-38**

AI&DS **II/III**

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : V - Model Evaluation Date of Lecture:

Topic of Lecture: Out-of-Sample Evaluation Metrics

Introduction :

Performance of a data model developed by data scientists is a direct way to measure their efficiency.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Prediction and Decision Making
- Generalization Error

Detailed content of the Lecture:

Evaluating your machine learning algorithm is an essential part of any project. Your model may give you satisfying results when evaluated using a metric ***say accuracy_score*** but may give poor results when evaluated against other metrics such as ***logarithmic_loss*** or any other such metric. Most of the times we use classification accuracy to measure the performance of our model, however it is not enough to truly judge our model. In this post, we will cover different types of evaluation metrics available.

Classification Accuracy

Classification Accuracy is what we usually mean, when we use the term accuracy. It is the ratio of number of correct predictions to the total number of input samples.

$$Accuracy = \frac{Number\ of\ Correct\ predictions}{Total\ number\ of\ predictions\ made}$$

It works well only if there are equal number of samples belonging to each class.

Logarithmic Loss

Logarithmic Loss or Log Loss, works by penalising the false classifications. It works well for multi-class classification. When working with Log Loss, the classifier must assign probability to each class for all the samples.

Confusion Matrix

Confusion Matrix as the name suggests gives us a matrix as output and describes the complete performance of the model.

There are 4 important terms :

True Positives : The cases in which we predicted YES and the actual output was also YES.

True Negatives : The cases in which we predicted NO and the actual output was NO.

False Positives : The cases in which we predicted YES and the actual output was NO.

False Negatives : The cases in which we predicted NO and the actual output was YES.

Accuracy for the matrix can be calculated by taking average of the values lying across the **“main diagonal”**

Area Under Curve

Area Under Curve(AUC) is one of the most widely used metrics for evaluation. It is used for binary classification problem. *AUC* of a classifier is equal to the probability that the classifier will rank a randomly chosen positive example higher than a randomly chosen negative example.

Mean Absolute Error

Mean Absolute Error is the average of the difference between the Original Values and the Predicted Values. It gives us the measure of how far the predictions were from the actual output. However, they don’t gives us any idea of the direction of the error i.e. whether we are under predicting the data or over predicting the data.

Mean Squared Error(MSE)

Mean Squared Error(MSE) is quite similar to Mean Absolute Error, the only difference being that MSE takes the average of the **square** of the difference between the original values and the predicted values. The advantage of MSE being that it is easier to compute the gradient, whereas Mean Absolute Error requires complicated linear programming tools to compute the gradient. As, we take square of the error, the effect of larger errors become more pronounced then smaller error, hence the model can now focus more on the larger errors.

Video Content / Details of website for further learning (if any):

https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

Important Books/Journals for further learning including the page nos.:

75. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

76. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

IQAC

LECTURE HANDOUTS

L-39

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : V - Model Evaluation Date of Lecture:

Topic of Lecture: Cross Validation

Introduction :

Cross-Validation is a statistical method of evaluating and comparing learning algorithms by dividing data into two segments: one used to learn or train a model and the other used to validate the model

Prerequisite knowledge for Complete understanding and learning of Topic:

- Generalization Error
- Out-of-Sample Evaluation Metrics

Detailed content of the Lecture:

Cross validation is a technique for assessing how the statistical analysis generalises to an independent data set.It is a technique for evaluating machine learning models by training several models on subsets of the available input data and evaluating them on the complementary subset of the data. Using cross-validation, there are high chances that we can detect over-fitting with ease. We will discuss the most popular method of them i.e the K-Fold Cross Validation. The others are also very effective but less common to use.

So let’s take a minute to ask ourselves why we need cross-validation —

We have been splitting the data set into a training set and testing set (or holdout set). But, the accuracy and metrics are highly biased upon how the split was performed, it depends upon whether the data set was shuffled, which part was taken for training and testing, how much, so on. Moreover, it is not representative of the model’s ability to generalize. This leads us to cross validation.

K-Fold Cross Validation

First I would like to introduce you to a golden rule — *“Never mix training and test data”*. Your first step should always be to **isolate the test data-set** and use it only for final evaluation. Cross-validation will thus be performed on the training set.

Initially, the entire training data set is broken up in *k* equal parts. The first part is kept as the hold out (testing) set and the remaining *k-1* parts are used to train the model. Then the trained model is then tested on the holdout set. The above process is repeated k times, in each case we keep on changing the holdout set. Thus, every data point get an equal opportunity to be included in the test set.

Usually, k is equal to 3 or 5. It can be extended even to higher values like 10 or 15 but it becomes extremely computationally expensive and time-consuming. Let us have a look at how we can implement this with a few lines of Python code and the Sci-kit Learn API.

Although it might be computationally expensive, cross-validation is essential for evaluating the performance of the learning model.

Feel free to have a look at the other cross-validation score evaluation methods which I have included in the references section, at the end of this article.

Video Content / Details of website for further learning (if any):
https://towardsdatascience.com/cross-validation-430d9a5fee22

Important Books/Journals for further learning including the page nos.:

77. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

78. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L-40

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : V - Model Evaluation Date of Lecture:

Topic of Lecture: Over fitting

Introduction :

Overfitting is an error that occurs in data modeling as a result of a particular function aligning too closely to a minimal set of data points. Financial professionals are at risk of overfitting a model based on limited data and ending up with results that are flawed.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Out-of-Sample Evaluation Metrics
- Cross Validation

Detailed content of the Lecture:

Overfitting is a concept in data science, which occurs when a statistical model fits exactly against its training data. When this happens, the algorithm unfortunately cannot perform accurately against unseen data, defeating its purpose. Generalization of a model to new data is ultimately what allows us to use machine learning algorithms every day to make predictions and classify data.

When machine learning algorithms are constructed, they leverage a sample dataset to train the model. However, when the model trains for too long on sample data or when the model is too complex, it can start to learn the “noise,” or irrelevant information, within the dataset. When the model memorizes the noise and fits too closely to the training set, the model becomes “overfitted,” and it is unable to generalize well to new data. If a model cannot generalize well to new data, then it will not be able to perform the classification or prediction tasks that it was intended for.

Low error rates and a high variance are good indicators of overfitting. In order to prevent this type of behavior, part of the training dataset is typically set aside as the “test set” to check for overfitting. If the training data has a low error rate and the test data has a high error rate, it signals overfitting.

If overtraining or model complexity results in overfitting, then a logical prevention response would be either to pause training process earlier, also known as, “early stopping” or to reduce complexity in the model by eliminating less relevant inputs. However, if you pause too early or exclude too many important features, you may encounter the opposite problem, and instead, you may underfit your model. Underfitting occurs when the model has not trained for enough time or the input variables are not significant enough to determine a meaningful relationship between the input and output variables.

To understand the accuracy of machine learning models, it’s important to test for model fitness. K-fold cross-validation is one of the most popular techniques to assess accuracy of the model.

In k-folds cross-validation, data is split into k equally sized subsets, which are also called “folds.” One of the k-folds will act as the test set, also known as the holdout set or validation set, and the remaining folds will train the model. This process repeats until each of the fold has acted as a holdout fold. After each evaluation, a score is retained and when all iterations have completed, the scores are averaged to assess the performance of the overall model.

Avoid Overfitting

While using a linear model helps us avoid overfitting, many real-world problems are nonlinear ones. In addition to understanding how to detect overfitting, it is important to understand how to avoid overfitting altogether. Below are a number of techniques that you can use to prevent overfitting:

- **Early stopping:** As we mentioned earlier, this method seeks to pause training before the model starts learning the noise within the model. This approach risks halting the training process too soon, leading to the opposite problem of underfitting. Finding the “sweet spot” between underfitting and overfitting is the ultimate goal here.
- **Train with more data:** Expanding the training set to include more data can increase the accuracy of the model by providing more opportunities to parse out the dominant relationship among the input and output variables. That said, this is a more effective method when clean, relevant data is injected into the model. Otherwise, you could just continue to add more complexity to the model, causing it to overfit.
- **Data augmentation:** While it is better to inject clean, relevant data into your training data, sometimes noisy data is added to make a model more stable. However, this method should be done sparingly.
- **Feature selection:** When you build a model, you’ll have a number of parameters or features that are used to predict a given outcome, but many times, these features can be redundant to others. Feature selection is the process of identifying the most important ones within the training data and then eliminating the irrelevant or redundant ones. This is commonly mistaken for dimensionality reduction, but it is different. However, both methods help to simplify your model to establish the dominant trend in the data

Video Content / Details of website for further learning (if any):
https://www.ibm.com/cloud/learn/overfitting

Important Books/Journals for further learning including the page nos.:

79. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

80. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

DESIGNING YOUR FUTURE
Estd. 2000

IQAC

LECTURE HANDOUTS

L-41

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : V - Model Evaluation Date of Lecture:

Topic of Lecture: Under Fitting

Introduction :

Underfitting destroys the accuracy of our machine learning model. Its occurrence simply means that our model or the algorithm does not fit the data well enough.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Cross Validation
- Over fitting

Detailed content of the Lecture:

Underfitting is a scenario in data science where a data model is unable to capture the relationship between the input and output variables accurately, generating a high error rate on both the training set and unseen data. It occurs when a model is too simple, which can be a result of a model needing more training time, more input features, or less regularization. Like overfitting, when a model is underfitted, it cannot establish the dominant trend within the data, resulting in training errors and poor performance of the model. If a model cannot generalize well to new data, then it cannot be leveraged for classification or prediction tasks. Generalization of a model to new data is ultimately what allows us to use machine learning algorithms every day to make predictions and classify data.

High bias and low variance are good indicators of underfitting. Since this behavior can be seen while using the training dataset, underfitted models are usually easier to identify than overfitted ones.

If undertraining or lack of complexity results in underfitting, then a logical prevention strategy would be to increase the duration of training or add more relevant inputs. However, if you train the model too much or add too many features to it, you may overfit your model, resulting in low bias but high variance (i.e. the bias-variance tradeoff). In this scenario, the statistical model fits too closely against its training data, rendering it unable to generalize well to new data points. It's important to note that some types of models can be more prone to overfitting than others, such as decision trees or KNN.

Identifying overfitting can be more difficult than underfitting because unlike underfitting, the training data performs at high accuracy in an overfitted model. To assess the accuracy of an algorithm, a technique called k-fold cross-validation is typically used.In k-folds cross-validation, data is split into k equally sized subsets, which are also called “folds.” One of the k-folds will act as the test set, also known as the holdout set or validation set, and the remaining folds will train the model. This process repeats until each of the fold has acted as a holdout fold. After each evaluation, a score is retained and when all iterations have completed, the scores are averaged to assess the performance of the overall model.

Avoid Underfitting

Since we can detect underfitting based off of the training set, we can better assist at establishing the dominant relationship between the input and output variables at the onset. By maintaining adequate model complexity, we can avoid underfitting and make more accurate predictions. Below are a few techniques that can be used to reduce underfitting:

Decrease regularization

Regularization is typically used to reduce the variance with a model by applying a penalty to the input parameters with the larger coefficients. There are a number of different methods, such as L1 regularization, Lasso regularization, dropout, etc., which help to reduce the noise and outliers within a model. However, if the data features become too uniform, the model is unable to identify the dominant trend, leading to underfitting. By decreasing the amount of regularization, more complexity and variation is introduced into the model, allowing for successful training of the model.

Increase the duration of training

As mentioned earlier, stopping training too soon can also result in underfit model. Therefore, by extending the duration of training, it can be avoided. However, it is important to cognizant of overtraining, and subsequently, overfitting. Finding the balance between the two scenarios will be key.

Feature selection

With any model, specific features are used to determine a given outcome. If there are not enough predictive features present, then more features or features with greater importance, should be introduced. For example, in a neural network, you might add more hidden neurons or in a random forest, you may add more trees. This process will inject more complexity into the model, yielding better training results.

Video Content / Details of website for further learning (if any):
https://www.ibm.com/cloud/learn/underfitting

Important Books/Journals for further learning including the page nos.:

81. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

82. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-42

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : V - Model Evaluation Date of Lecture:

Topic of Lecture: Model Selection

Introduction :

Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset. Model selection is a process that can be applied both across different types of models (e.g. logistic regression, SVM, KNN, etc.)

Prerequisite knowledge for Complete understanding and learning of Topic:

- Over fitting
- Under Fitting

Detailed content of the Lecture:

Model selection is the process of selecting one final machine learning model from among a collection of candidate machine learning models for a training dataset.

Model selection is a process that can be applied both across different types of models (e.g. logistic regression, SVM, KNN, etc.) and across models of the same type configured with different model hyperparameters (e.g. different kernels in an SVM).

Model selection is the process of choosing one of the models as the final model that addresses the problem.

Considerations for Model Selection

Fitting models is relatively straightforward, although selecting among them is the true challenge of applied machine learning.

Firstly, we need to get over the idea of a "*best*" model.

All models have some predictive error, given the statistical noise in the data, the incompleteness of the data sample, and the limitations of each different model type. Therefore, the notion of a perfect or best model is not useful. Instead, we must seek a model that is "*good enough.*"

The best approach to model selection requires "*sufficient*" data, which may be nearly infinite depending on the complexity of the problem.

In this ideal situation, we would split the data into training, validation, and test sets, then fit candidate models on the training set, evaluate and select them on the validation set, and report the performance of the final model on the test set.

Instead, there are two main classes of techniques to approximate the ideal case of model selection; they are:

Probabilistic Measures: Choose a model via in-sample error and complexity.

Resampling Methods: Choose a model via estimated out-of-sample error.

Four commonly used probabilistic model selection measures include:

Akaike Information Criterion (AIC).

Bayesian Information Criterion (BIC).

Minimum Description Length (MDL).

Structural Risk Minimization (SRM).

Probabilistic measures are appropriate when using simpler linear models like linear regression or logistic regression where the calculating of model complexity penalty (e.g. in sample bias) is known and tractable.

Resampling Methods

Resampling methods seek to estimate the performance of a model (or more precisely, the model development process) on out-of-sample data.

This is achieved by splitting the training dataset into sub train and test sets, fitting a model on the sub train set, and evaluating it on the test set. This process may then be repeated multiple times and the mean performance across each trial is reported.

It is a type of Monte Carlo estimate of model performance on out-of-sample data, although each trial is not strictly independent as depending on the resampling method chosen, the same data may appear multiple times in different training datasets, or test datasets.

Three common resampling model selection methods include:

Random train/test splits.

Cross-Validation (k-fold, LOOCV, etc.).

Bootstrap.

Most of the time probabilistic measures (described in the previous section) are not available, therefore resampling methods are used.

By far the most popular is the cross-validation family of methods that includes many subtypes.

Video Content / Details of website for further learning (if any):

https://machinelearningmastery.com/a-gentle-introduction-to-model-selection-for-machine-learning/

Important Books/Journals for further learning including the page nos.:

83. Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

84. David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

MUTHAYAMMAL ENGINEERING COLLEGE

(An Autonomous Institution)

(Approved by AICTE, New Delhi, Accredited by NAAC & Affiliated to Anna University)
Rasipuram - 637 408, Namakkal Dist., Tamil Nadu

Estd. 2000

LECTURE HANDOUTS

L-43

AI&DS

II/III

Course Name with Code : **19ADC05/ Introduction to Data Science**

Course Teacher :**Dr.P.Srinivasan**

Unit : **V - Model Evaluation Date of Lecture:**

Topic of Lecture: Prediction by using Ridge Regression

Introduction :

Ridge regression is the method used for the analysis of multicollinearity in multiple regression data. It is most suitable when a data set contains a higher number of predictor variables than the number of observations.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Under Fitting
- Model Selection

Detailed content of the Lecture:

Ridge regression is a model tuning method that is used to analyse any data that suffers from multicollinearity. This method performs L2 regularization. When the issue of multicollinearity occurs, least-squares are unbiased, and variances are large, this results in predicted values to be far away from the actual values.

The cost function for ridge regression:

Min(||Y – X(theta)||^2 + λ||theta||^2)

Lambda is the penalty term. λ given here is denoted by an alpha parameter in the ridge function. So, by changing the values of alpha, we are controlling the penalty term. Higher the values of alpha, bigger is the penalty and therefore the magnitude of coefficients is reduced.

It shrinks the parameters. Therefore, it is used to prevent multicollinearity

It reduces the model complexity by coefficient shrinkage

Ridge Regression Models

For any type of regression machine learning models, the usual regression equation forms the base which is written as:

Y = XB + e

Where Y is the dependent variable, X represents the independent variables, B is the regression coefficients to be estimated, and e represents the errors are residuals.

Once we add the lambda function to this equation, the variance that is not evaluated by the general model is considered. After the data is ready and identified to be part of L2 regularization, there are steps that one can undertake.

Standardization

In ridge regression, the first step is to standardize the variables (both dependent and independent) by subtracting their means and dividing by their standard deviations. This causes a challenge in notation since we must somehow indicate whether the variables in a particular formula are standardized or not. As far as standardization is concerned, all ridge regression calculations are based on standardized variables. When the final regression coefficients are displayed, they are adjusted back into their original scale. However, the ridge trace is on a standardized scale.

Bias and variance trade-off

Bias and variance trade-off is generally complicated when it comes to building ridge regression models on an actual dataset. However, following the general trend which one needs to remember is:

The bias increases as λ increases.

The variance decreases as λ increases.

Assumptions of Ridge Regressions

The assumptions of ridge regression are the same as that of linear regression: linearity, constant variance, and independence. However, as ridge regression does not provide confidence limits, the distribution of errors to be normal need not be assumed.

Now, let's take an example of a linear regression problem and see how ridge regression if implemented, helps us to reduce the error.

We shall consider a data set on Food restaurants trying to find the best combination of food items to improve their sales in a particular region.

Regularization

Value of alpha, which is a hyperparameter of Ridge, which means that they are not automatically learned by the model instead they have to be set manually. We run a grid search for optimum alpha values

To find optimum alpha for Ridge Regularization we are applying GridSearchCV

Video Content / Details of website for further learning (if any):

https://www.mygreatlearning.com/blog/what-is-ridge-regression/

Important Books/Journals for further learning including the page nos.:

85. Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.

86. David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-44

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : V - Model Evaluation Date of Lecture:

Topic of Lecture: Testing Multiple Parameters

Introduction :

A technique for testing a hypothesis in which multiple variables are modified. The goal of multivariate testing is to determine which combination of variations performs the best out of all of the possible combinations.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Model Selection
- Prediction by using Ridge Regression

Detailed content of the Lecture:

Hypothesis Testing

Hypothesis Testing in statistics is a method to test the results of experiments or surveys to see if you have meaningful results. It is useful when you want to infer about a population based on a sample or correlation between two or more samples.

Null Hypothesis

This hypothesis states that there is no significant difference between sample and population or among different populations. It is denoted by H_0.

Ex – We assume that the mean of 2 samples is equal.

Alternate Hypothesis

The statement contrary to the null hypothesis comes under the alternate hypothesis. It is denoted by H_1.

Ex – We assume that the mean of the 2 samples is unequal.

Critical Value

It is a point on the scale of the test statistic beyond which the null hypothesis is rejected. Higher the critical value, lower the probability of 2 samples belonging to the same distribution. The critical value for any test can

p-value

p-value stands for ‘probability value’; it tells how likely it is that a result occurred by chance alone. Basically, the p-value is used in hypothesis testing to help you support or reject the null hypothesis. The smaller the p-value, the stronger the evidence to reject the null hypothesis.

Degree of freedom

The degree of freedom is the number of independent variables. This concept is used in calculating t statistic and chi-square statistic.

A statistical test is a way to determine whether the random variable is following the null hypothesis or alternate hypothesis. It basically tells whether the sample and population or two/ more samples have significant differences. You can use various descriptive stats such as mean, median, mode, range, or standard deviation for this purpose. However, we generally use the mean. The statistical test gives you a number which is then compared with the p-value. If its value is more than the p-value you accept the null hypothesis, else you reject it.

The procedure for implementing each statistical test will be as follows:

- We calculate the statistic value using the mathematical formula
- We then calculate the critical value using statistic tables
- With the help of critical value, we calculate the p-value
- If p-value> 0.05 we accept the null hypothesis else we reject it

Now you have an understanding of feature selection and statistical tests, we can move towards the implementation of various statistical tests along with their meaning. Before that, I will show you the dataset and this dataset will be used to perform all tests.

Video Content / Details of website for further learning (if any):
https://www.analyticsvidhya.com/blog/2021/06/feature-selection-using-statistical-tests/

Important Books/Journals for further learning including the page nos.:

Cathy O’Neil and Rachel Schutt , “Doing Data Science”, O'Reilly, 2015.

David Dietrich, Barry Heller, Beibei Yang, “Data Science and Big data Analytics”,EMC 2013

Course Teacher

Verified by HOD

LECTURE HANDOUTS

L-45

AI&DS

II/III

Course Name with Code : 19ADC05/ Introduction to Data Science

Course Teacher :Dr.P.Srinivasan

Unit : V - Model Evaluation Date of Lecture:

Topic of Lecture: Testing Multiple Parameters by using Grid Search

Introduction :

Grid search refers to a technique used to identify the optimal hyperparameters for a model. Unlike parameters, finding hyperparameters in training data is unattainable.

Prerequisite knowledge for Complete understanding and learning of Topic:

- Prediction by using Ridge Regression
- Testing Multiple Parameters

Detailed content of the Lecture:

Hyperparameters are model parameters whose values are set before training. For example, the number of neurons of a feed-forward neural network is a hyperparameter, because we set it before training. Another example of hyperparameter is the number of trees in a random forest or the penalty intensity of a Lasso regression. They are all numbers that are set before the training phase and their values affect the behavior of the model.

Why should we tune the hyperparameters of a model? Because we don't really know their optimal values in advance. A model with different hyperparameters is, actually, a different model so it may have a lower performance.

In the case of neural networks, a low number of neurons could lead to underfitting and a high number could lead to overfitting. In both cases, the model is not good, so we need to find the intermediate number of neurons that leads to the best performance.

If the model has several hyperparameters, we need to find the best combination of values of the hyperparameters searching in a multi-dimensional space. That's why hyperparameter tuning, which is the process of finding the right values of the hyperparameters, is a very complex and time-expensive task.

Let's see two of the most important algorithms for hyperparameter tuning, that are grid search and random search.

Grid search

Grid search is the simplest algorithm for hyperparameter tuning. Basically, we divide the domain of the hyperparameters into a discrete grid. Then, we try every combination of values of this grid, calculating some performance metrics using cross-validation. The point of the grid that maximizes the average value in cross-validation, is the optimal combination of values for the hyperparameters.

Grid search is an exhaustive algorithm that spans all the combinations, so it can actually find the best point in the domain. The great drawback is that it's very slow. Checking every combination of the space requires a lot of time that, sometimes, is not available. Don't forget that every point in the grid needs k-fold cross-validation, which requires *k* training steps. So, tuning the hyperparameters of a model in this way can be quite complex and expensive. However, if we look for the best combination of values of the hyperparameters, grid search is a very good idea.

Video Content / Details of website for further learning (if any):
https://www.yourdatateacher.com/2021/05/19/hyperparameter-tuning-grid-search-and-random-search/

Important Books/Journals for further learning including the page nos.:

Cathy O'Neil and Rachel Schutt , "Doing Data Science", O'Reilly, 2015.

David Dietrich, Barry Heller, Beibei Yang, "Data Science and Big data Analytics",EMC 2013

Course Teacher

Verified by HOD