Analyzing and Mining Data

Overview of Statistical Analysis

Statistics

- Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of numerical or quantitative data.
    - It's all around us in our day to day lives. Whether we're talking about average income, average age, or highest-paid professions—it's all statistics.
    - Today, statistics is being applied across industries for decision-making based on data.
- For example,
    - researchers using statistics to analyze data from the production of vaccines to ensure safety and efficacy
    - companies using statistics to reduce customer churn by gaining greater insight into customer requirements.

Statistical Analysis

- Statistical Analysis is the application of statistical methods to a sample of data in order to develop an understanding of what that data represents.
    - It includes collecting and scrutinizing every data sample in a set of items from which samples can be drawn.
- A sample, in Statistics, is a representative selection drawn from a total population,
- where population is a discrete group of people or things that can be identified by at least one common characteristic for purposes of data collection and analysis.
    - For example, in a certain use case, population may be all people in a state that have a driving license, and a sample of this population that is a part, or subset, of the population could be men drivers over the age of 50.
- Statistical methods are mainly useful to ensure that data is interpreted correctly, and apparent relationships are meaningful and not just happening by chance.
- Whenever we collect data from a sample, there are two different types of statistics we can run.

- Descriptive statistics to summarize information about the sample
- Inferential statistics to make inferences or generalizations about the broader population.

Descriptive Statistics

- Descriptive Statistics enables you to present data in a meaningful way allowing simpler interpretation of the data.
- Data is described using summary charts, tables, and graphs without any attempts to draw conclusions about the population from which the sample is taken.
- The objective is to make it easier to understand and visualize raw data without making conclusions regarding any hypotheses that were made.
    - For example, we want to describe the English test scores in a specific class of 25 student s.
    - We record the test scores of all students, calculate the summary statistics, and produce a graph.
- Some of the common measures of Descriptive Statistical Analysis include Central Tendency, Dispersion, and Skewness: Central Tendency, or locating the center of a data sample.
    - Some of the common measures of central tendency include mean, median, and mode.
    - These measures tell you where most values in your dataset fall.
    - So, in the earlier example, the mean score, or the mathematical average, of the class of 25 students would be the sum total of the scores of all 25 students, divided by 25, that is, the number of students.
    - If you order the above dataset from the smallest score value to the highest score value of the 25 students and pick the middle value— that is the value with 12 values to the left and 12 values to the right of a score value, that score value would be the median for this dataset. If 12 students have scored less than 75%, and 12 students have scored greater than 75%, then the median is 75. Median is unique for each dataset and is not affected by outliers.

    - Mode is the value that occurs most frequently in a set of observations. For example, if the most common score in this group of 25 students is 72%, then that is the mode for this dataset. So, you can see how looking at your dataset through these values can help you get a clearer understanding of your dataset.
- Dispersion is the measure of variability in a dataset. Common measures of statistical dispersion are Variance, Standard Deviation, and Range.
- Variance defines how far away the data points fall from the center, that is, the distribution of values.
    - When a distribution has lower variability, the values in a dataset are more consistent. However, when the variability is higher, the data points are more dissimilar, and extreme values become more likely. Understanding variability can help you grasp the likelihood of an event happening.
- Standard deviation tells you how tightly your data is clustered around the mean.
- And Range gives you the distance between the smallest and largest values in your datasets.
- Skewness is the measure of whether the distribution of values is symmetrical around a central value or skewed left or right.
    - Skewed data can affect which types of analyses are valid to perform. These are some of the basic and most commonly used descriptive statistics tools, but there are other tools as well, for example, using correlation and scatterplots to assess the relationships of paired data.

Inferential Statistics

- The second type of statistical analysis is Inferential Statistics. Inferential statistics takes data from a sample to make inferences about the larger population from which the sample was drawn.
    - Using methods of inferential statistics, you can draw generalizations that apply the results of the sample to the population as a whole.
- Some common methodologies of Inferential Statistics include Hypothesis Testing, Confidence Intervals, and Regression Analysis:

- Hypothesis Testing—For example, for studying the effectiveness of a vaccine by comparing outcomes in a control group, hypothesis tests can tell you whether the efficacy of a vaccine observed in a control group is likely to exist in the population as well.
- Confidence Intervals incorporate the uncertainty and sample error to create a range of values the actual population value is like to fall within.
- Regression Analysis incorporates hypothesis tests that help determine whether the relationships observed in the sample data actually exist in the population rather than just the sample.

Conclusion

- There are various software packages to perform statistical data analysis, such as Statistical Analysis System (or SAS), Statistical Package for the Social Sciences (or SPSS), and Stat Soft.
- Statistics form the core of data mining by:
    - Providing measures and methodologies necessary for data mining; and Identifying patterns that help identify differences between random noise and significant findings.
    - Both data mining, which we will learn more about in this course, and Statistics, as techniques of data analysis, help in better decision-making.

What is Data Mining?

Overview

- Data mining or the process of extracting knowledge from data, is the heart of the data analysis process.
- It is an interdisciplinary field that involves the use of pattern recognition technologies, statistical analysis and mathematical techniques.
- Its goal is to identify correlations in data, find patterns and variations. Understand trends and predict probabilities.

Patterns and Trends

- You'll hear about patterns and trends frequently in the context of data analysis, so let's first understand these concepts. Pattern recognition is the discovery of regularity's or commonality's in data.
- Consider the log data for logins to an application in an organization. It contains information such as the username, login timestamp, time spent in each login session, and activities performed.
    - When we analyze this data to gain insights into the habits or behaviors of users, for example, the time of the day when maximum users tend to login or user roles that typically spend the maximum hours logged into the application or modules in the workflow application that are being used where examining the data manually or through tools to uncover patterns hidden in the data.
- A trend, on the other hand, is the general tendency of a set of data to change overtime.
    - For example, global warming in the short term, like a year on year basis temperatures may remain the same or go up or down by a few degrees, but the overall global temperatures continue to increase overtime, making global warming a trend.

Applications of Data Mining

- Data mining has applications across industries and disciplines.
    - For example, profiling customer behaviors needs and disposable income in order to offer targeted campaigns
    - financial institutions, tracking customer transactions for unusual behaviors, and flagging fraudulent transactions using data mining models.
    - The use of statistical models to predict a patients likelihood for specific health conditions and prioritizing treatment.
    - Accessing performance data of students to predict achievement levels and make a focused effort to provide support where required.
    - Helping investigation agencies deploy police force where the likelihood of crime is higher
    - and aligning supply and logistics with demand forecasts.

Data Mining Techniques

- There are several techniques you can use to detect patterns and build accurate models for discovery, be it descriptive, diagnostic, predictive, or prescriptive modeling.
- Let's understand some of the most commonly used techniques.
- Classification is a technique that classifies attributes into target categories
    - for example, classifying customers into low, medium, or high spenders based on how much they earn.
- Clustering is similar to classification, but involves grouping data into clusters so they can be treated as groups.
    - For example, clustering customers based on geographic regions
- Anomaly or outlier detection is a technique that helps find patterns and data that are not normal or unexpected.
    - For example, spikes in the usage of a credit card that can flag possible misuse.
- Association rule mining is a technique that helps establish our relationship between two data events.
    - For example, the purchase of a laptop being frequently accompanied by the purchase of a cooling pad.
- Sequential patterns is the technique that traces a series of events that take place in a sequence.
    - For example, tracing a customer shopping trail from the time they log into an online store to the time they log out.
- Affinity grouping is a technique used to discover Co occurrence in relationships.
    - This technique is widely used in on line stores for cross selling and up selling their products by recommending products to people based on the purchase history of other people who purchased the same item.
- Decision trees help build classification models in the form of a tree structure with multiple branches, where each branch represents a probable occurrence.
    - This technique helps to build a clear understanding of the relationship between input and output.
- Regression is a technique that helps identify the nature of the relationship between two variables, which could be causal or correlational.

- For example, based on factors such as location and covered area, a regression model could be used to predict the value of a house.

Conclusion

- Data mining essentially helps separate the noise from the real information and helps businesses focus their energies on only what is relevant.

Tools for Data Mining?

Overview

- In this video, we will learn about some of the commonly used software and tools for data mining, such as: Spreadsheets, R-Language, Python, IBM SPSS Statistics, IBM Watson Studio; and SAS.

Spreadsheets

- Spreadsheets, such as Microsoft Excel and Google Sheets, are commonly used for performing basic data mining tasks.
- Spreadsheets can be used to host data that has been exported from other systems in an easily accessible and easy-to-read format. You can pivot tables to showcase specific aspects of your data, which is vital when you have huge amounts of data to sort through and analyze.
- They also make it relatively easier to make comparisons between different sets of data.
- Add-ins available for Excel, such as the Data Mining Client for Excel, XLMiner, and KnowledgeMiner for Excel, allow you to perform common mining tasks, such as classification, regression, association rules, clustering, and model building.
- GoogleSheets also has an array of add-ons that can be used for analysis and mining, such as Text Analysis, Text Mining, Google Analytics.

R-language

- R is one of the most widely used languages for performing statistical modeling and computations by statisticians and data miners.

- R is packaged with hundreds of libraries explicitly built for data mining operations such as regression, classification, data clustering, association rule mining, text mining, outlier detection, and social network analysis.
- Some of the popular R packages include tm and twitteR.
    - tm, a framework for text mining applications within R, provides functions for text mining.
    - twitteR provides a framework for mining tweets.
- R Studio is a popularly used open-source Integrated Development Envionrment (or IDE) for working with the R programming language.

Python

- Python libraries like Pandas and NumPy are commonly used for Data Mining. Pandas is an open-source module for working with data structures and analysis. It is possibly one of the most popular libraries for data analysis in Python.
- It allows you to upload data in any format and provides a simple platform to organize, sort, and manipulate that data.
- Using Pandas, you can:
    - perform basic numerical computations such as mean, median, mode, and range;
    - calculate statistics and answer questions regarding correlation between data and distribution of data;
    - explore data visually and quantitatively;
    - visualize data with help from other Python libraries.
- NumPy is a tool for mathematical computing and data preparation in Python.
    - NumPy offers a host of built-in functions and capabilities for data mining.
- Jupyter Notebooks have become the tool of choice for Data Scientists and Data Analysts when working with Python to perform data mining and statistical analysis.

IBM SPSS Statistics

- SPSS stands for Statistical Process for Social Sciences.
- While the name suggests its original usage in the field of Social Sciences,

    - it is popularly used for advanced analytics, text analytics, trend analysis, validation of assumptions, and translation of business problems into data science solutions.
- SPSS is closed source and requires a license for use.
- SPSS has an easy to use interface that requires minimal coding for complex tasks.
    - It comprises of efficient data management tools and is popular because of its in-depth analysis capabilities and accurate data results.

IBM Watson Studio

- IBM Watson Studio, included in the IBM Cloud Pak for Data, leverages a collection of open source tools such as Jupyter notebooks, and extends them with closed source IBM tools that make it a powerful environment for data analysis and data science.
    - It is available through a web browser on the public cloud, private cloud, and as a desktop app.
    - Watson Studio enables team members to collaborate on projects, that can range from simple exploratory analysis to building machine learning and AI models.
    - It also includes SPSS Modeller flows that enable you to quickly develop predictive models for your business data.

SAS

- SAS Enterprise Miner is a comprehensive, graphical workbench for data mining.
    - It provides powerful capabilities for interactive data exploration, which enables users to identify relationships within data.
    - SAS can manage information from various sources, mine and transform data, and analyze statistics.
    - It offers a graphical user interface for non-technical users.
- With SAS, you can:
    - identify patterns in the data using a range of available modeling techniques;
    - explore relationships and anomalies in data;
    - analyze big data;
    - validate the reliability of findings from the data analysis process.

- SAS is very easy to use because of its syntax and is also easy to debug. It has the ability to handle large databases and offers high security to its users.

Conclusion

- In this video, we have learned about just a few of the data mining tools available today.
- Your decision regarding the best tool for your needs will be driven by the data size and structures the tool supports, the features it offers, its data visualization capabilities, infrastructure needs, ease of use, and learnability.
- It's fairly common to use a combination of data mining tools to meet all your needs.

Summary

- In this lesson, you have learned the following information:
    - Statistics is a branch of mathematics dealing with the collection, analysis, interpretation, and presentation of numerical or quantitative data.
    - Statistical Analysis involves the use of statistical methods in order to develop an understanding of what the data represents.
- Statistical Analysis can be:
    - Descriptive; that which provides a summary of what the data represents. Common measures include Central Tendency, Dispersion, and Skewness.
    - Inferential; that which involves making inferences, or generalizations, about data. Common measures include Hypothesis Testing, Confidence Intervals, and Regression Analysis.
- Data Mining, simply put, is the process of extracting knowledge from data. It involves the use of pattern recognition technologies, statistical analysis, and mathematical techniques, in order to identify correlations, patterns, variations, and trends in data.
- There are several techniques that can help mine data, such as, classifying attributes of data, clustering data into groups, establishing relationships between events, variables, and input and output.
- A variety of software and tools are available for analyzing and mining data. Some of the popularly used ones include Spreadsheets, R-Language, Python, IBM SPSS Statistics, IBM

Watson Studio, and SAS, each with their own set of characteristics, strengths, limitations, and applications.