Week 6: Day 2 Daily Morning Challenge

Day 2: Thursday 28th January 2020

Question 1: Briefly explain the concept of hypothesis testing and its importance in the data science process

Hypothesis testing is a statistical method that is used in making statistical decisions using experimental data. Hypothesis Testing is basically an assumption that we make about the population parameter. Using Hypothesis Testing, we try to interpret or draw conclusions about the population using sample data. A Hypothesis Test evaluates two mutually exclusive statements about a population to determine which statement is best supported by the sample data.

Hypothesis testing in the data science process is important because we need to draw inferences or some conclusion about the overall population or data by conducting some statistical tests on a sample.

Question 2: Briefly explain the concepts and differences of cluster and outlier analysis

Cluster analysis is a group of multivariate techniques whose primary purpose is to group objects (e.g., respondents, products, or other entities) based on the characteristics they possess. It is a means of grouping records based upon attributes that make them similar. If plotted geometrically, the objects within the clusters will be close together, while the distance between clusters will be farther apart. The primary objective of cluster analysis is to define the structure of the data by placing the most similar observations into groups.

An outlier is an object that deviates significantly from the rest of the objects. They can be caused by measurement or execution error. The analysis of outlier data is referred to as outlier analysis or outlier mining. Outlier analysis refers to the problem of finding patterns in data that do not conform to expected normal behaviour. These anomalous patterns are often referred to as outliers, anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants in different application domains. Outlier analysis has been a widely researched problem and finds immense use in a wide variety of application.

<u>Differences:</u>
A cluster is formed when several data points lie in a small interval.
An outlier has a value that is much greater than or much less than other data in the set.
An outlier may significantly affect the mean of a data set while a cluster may not.

Question 3: Outline the major tasks that are involved in data preprocessing

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Data preprocessing prepares raw data for further processing. Data preprocessing is extremely important because it allows improving the quality of the raw experimental data. The primary aim of data preprocessing is to minimise or, eventually, eliminate those small data contributions associated with the experimental error.

Steps Involved in Data Preprocessing.

- 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.
- Data Transformation: This step is taken in order to transform the data in appropriate forms suitable for mining process such as Normalization, Attribute Selection, Discretization etc
- 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.