Name: A. Shiva Surya Saran
Roll No.: BE-B 20

Data Analytics

Assignment 2

Unit 1: Basic Data Analytic Tools

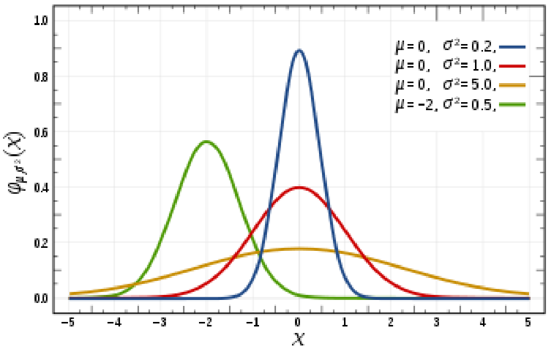
1. Explain Statistical Methods for Evaluation with Hypothesis testing and difference of means.

Statistical Methods for Evaluation

- Visualization is useful for data exploration and presentation, but statistics is crucial because it may exist throughout the entire Data Analytics Lifecycle.
- Statistical techniques are used during the initial data exploration and data preparation model building, evaluation of final models and assessment of how the new models improve the situation when deployed in the field.
- There are some useful statistical tools that are listed below:

Hypothesis testing

- When comparing populations, such as testing evaluating the difference of the means from the 2 samples of data a common technique to assess the difference of the significance of the differences hypothesis testing.



Distribution of Two Samples of Data

- The Basic concept of hypothesis testing is to form an assertion and tested with data. Hey when performing hypothesis testing the common assumption is that there is no difference between 2 samples. This assumption is used as the default position for building the test or conducting a scientific experiment. Statisticians refer to this as the null hypothesis (H_0).

- The alternative hypothesis H_A is that there is a difference between 2 samples.
  Eg. if the task is to identify the effect of drug A compared to drug B on patient, the null hypothesis and alternative hypothesis would be this:
  - H_0: Drug A and Drug B have the same effect on patient
  - $H_{A:}$ Drug A has a greater effect then Dru B on patients.

- Following table shows example of null hypothesis and alternative hypothesis

Applications	Null Hypothesis	Alternative Hypothesis
Accuracy forecast	Model X doesn't predict better than the existing model.	Model X predicts better than the existing model
Recommendation engine	Algorithm Y does not produce better recommendations than the current algorithm being used.	Algorithm Y produces better recommendations than the current algorithm being used this
Regression modelling	This variable does not affect the outcome because its coefficient is zero.	This variable affect outcome because its coefficient is not zero.

Difference of means

Hypothesis testing is a common approach draw inferences on whether or not the 2 populations, denoted pop one and pop 2, are different from each other. This provides to hypothesis test to compare the means of the respective populations based on sample randomly drawn from each population.

Specifically, the 2 hypothesis tests in this Point, consider the following null and alternative hypothesis

$H_0 : \mu_1 = \mu_2$
$H_A : \mu_1 \neq \mu_2$

μ_1 and μ_2 denote the population means of Pop 1 and Pop 2 respectively.

The basic testing approach is to compare the observed sample means, X1 and X2, corresponding to each population. If the values of X1 bar and xtube are approximately equal to each other, the distributions of expand my next to bar overlap substantially as shown in the figure and the null hypothesis is supported

A large observed difference between the sample means indicates that the null hypothesis should be restricted formally, the difference in means can be tested using

- Students T-test
- Weish test

2. Write Short Note On:

a. ***Wilcoxon rank-sum test***

The Wilcoxon rank-sum test is a nonparametric hypothesis test that checks whether two populations are identically distributed. Wilcoxon test does not assume anything about the population distribution, it is generally considered more robust than the t-test. This test is often performed as a two-sided test and, thus, the research hypothesis indicates that the populations are not equal as opposed to specifying directionality.

Steps in the test are:

- Assign ranks by arranging the observations from smallest to largest.
- Summation of ranks in each group
- U test is performed

b. **Type 1 Error**

A type 1 error is also known as a false positive and occurs when a researcher incorrectly rejects a true null hypothesis. This means that your report that your findings are significant when in fact they have occurred by chance.

Type 2 Error

A type II error is also known as a false negative and occurs when a researcher fails to reject a null hypothesis which is false. Here a researcher concludes there is not a significant effect, when there really is.

3. Write Short Note On:

a. **Power and Sample size**

The power of a test is the probability of correctly rejecting the null hypothesis. It is denoted by 1- β, where β is the probability of a type II error.

b. **ANOVA**

Analysis of Variance (ANOVA) is a generalization of the hypothesis testing of the difference of two population means. ANOVA tests if any of the population means differ from the other population means. The null hypothesis of ANOVA is that all the population means are equal.

4. Explain Clustering in detail.

In general, clustering is the use of unsupervised techniques for grouping similar objects. In machine learning unsupervised refers to the problem of finding hidden structure within unbiased data.

Clustering techniques unsupervised in one sense that the data scientist does not determine, in advance, the labels to apply to the clusters. hey structure of data describes the objects of interest in determines how best to group the objects,

Example:
Based on customers personal income, it is straight forward to divide the customers into 3 groups depending on the arbitrary selected values. The customers could be divided into 3 groups as follows:

- Earn less than $10,000
- Earn between $10,000 and $99,000
- Earn on $100,000 more

In this case, the income levels were chosen somewhat subjectively based on easy to communicate points of delineation. However, such groupings do not indicator natural affinity of the customers within each group

Clustering is a method often used for exploratory analysis of the data hey. In clustering, there no predictions made rather clustering methods find the similarities between objects according to the object attributes and groups the similar objects into clusters

5. Explain K-Means with example.

Given a collection of objects each with N measurable attributes, K means is an analytical technique that were chosen value of K, identifies K clusters of objects based on the object's proximity to the centre of K groups.

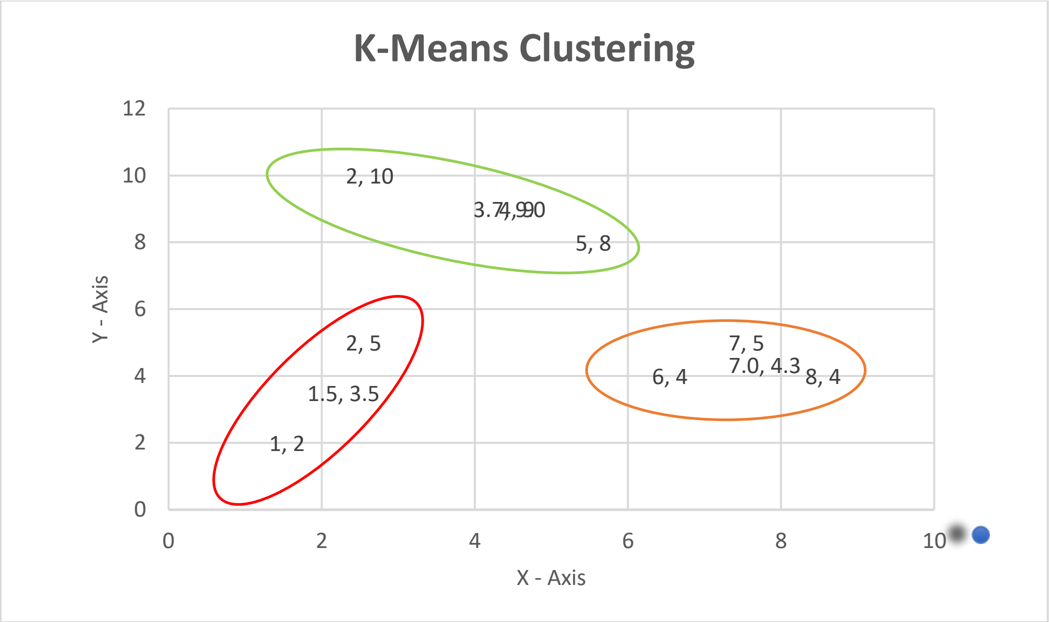
The centres determined as the arithmetic average (mean) of each cluster n-dimensional vector of attributes.

Following example shows 3 cluster of objects with 2 attributes. Each object in the data set is represented by a small dot colour coded to the closest large circle, the true mean of the cluster.

Example:
The following eight point (with(X,Y) representing locations)into three cluster:
A1(2,10),A2(2,5),A3(8,4),A4(5,8),A5(7,5),A6(6,4),A7(1,2),A8(4,9)
(Using Manhattan distance algorithm)

Using Manhattan distance algorithm

Initial cluster centres are: A1(2,10) , A2(2,5) and A3 (8,4)



Calculations:

			2	10	2	5	8	4	
	Points		Distance Mean 1		Distance Mean 2		Distance Mean 3		Cluster
A1	2	10	0		5		12		1
A2	2	5	5		0		7		2
A3	8	4	12		7		0		3
A4	5	8	5		6		7		1
A5	7	5	10		10		2		3
A6	6	4	10		5		2		3
A7	1	2	9		4		9		2
A8	4	9	3		6		9		1

Cluster Grouping:

Cluster 1	A1	A4	A8
Cluster 2	A2	A7	
Cluster 3	A3	A5	A6

New Cluster Centres:

New C 1	3.7	9.0
New C 2	1.5	3.5
New C 3	7.0	4.3