ASSIGNMENT – 3
MACHINE LEARNING

1. Which of the following is an application of clustering?
a. Biological network analysis
b. Market trend prediction
c. Topic modeling
d. All of the above
Ans: d. All of the above

2. On which data type, we cannot perform cluster analysis?
a. Time series data
b. Text data
c. Multimedia data
d. None
Ans : d. None

3. Netflix's movie recommendation system uses-
a. Supervised learning
b. Unsupervised learning
c. Reinforcement learning and Unsupervised learning
d. All of the above
Ans: c. Reinforcement learning and Unsupervised learning

4. The final output of Hierarchical clustering is-
a. The number of cluster centroids
b. The tree representing how close the data points are to each other
c. A map defining the similar data points into individual groups
d. All of the above
Ans : b. The tree representing how close the data points are to each other

5. Which of the step is not required for K-means clustering?
a. A distance metric
b. Initial number of clusters
c. Initial guess as to cluster centroids
d. None
Ans : d. None

6. Which is the following is wrong?
a. k-means clustering is a vector quantization method
b. k-means clustering tries to group n observations into k clusters
c. k-nearest neighbour is same as k-means
d. None
Ans : c. k-nearest neighbour is same as k-means

7. Which of the following metrics, do we have for finding dissimilarity between two clusters in hierarchical clustering?
i. Single-link
ii. Complete-link
iii. Average-link

Options:
a. 1 and 2
b. 1 and 3
c. 2 and 3
d. 1, 2 and 3
Ans : d. 1, 2 and 3

8. Which of the following are true?
i. Clustering analysis is negatively affected by multicollinearity of features
ii. Clustering analysis is negatively affected by heteroscedasticity

Options:
a. 1 only
b. 2 only
c. 1 and 2
d. None of them
Ans : . 1 only

9. In the figure above, if you draw a horizontal line on y-axis for y=2. What will be the number of clusters formed?
a. 2
b. 4
c. 3
d. 5
Ans : A 2

13. What is the importance of clusterin?

Ans : Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. Data mining is the process of analysing data from different perspectives and summarizing it into useful information. Data mining involves the anomaly detection, association rule learning, classification, regression,

summarization and clustering. In this paper, clustering analysis is done. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. Clustering is important in data analysis and data mining applications.

Clustering has a myriad of uses in a variety of industries. Some common applications for clustering include the following:

- market segmentation
- social network analysis
- search result grouping
- medical imaging
- image segmentation
- anomaly detection

After clustering, each cluster is assigned a number called a cluster ID. Now, you can condense the entire feature set for an example into its cluster ID. Representing a complex example by a simple cluster ID makes clustering powerful. Extending the idea, clustering data can simplify large datasets.

14. How can I improve my clustering performance

Ans : Centroid-based Clustering

Centroid-based clustering organizes the data into non-hierarchical clusters, in contrast to hierarchical clustering defined below. k-means is the most widely-used centroid-based clustering algorithm. Centroid-based algorithms are efficient but sensitive to initial conditions and outliers. This course focuses on k-means because it is an efficient, effective, and simple clustering algorithm.

Figure 1: Example of centroid-based clustering.

Density-based Clustering

Density-based clustering connects areas of high example density into clusters. This allows for arbitrary-shaped distributions as long as dense areas can be connected. These algorithms have difficulty with data of varying densities and high dimensions. Further, by design, these algorithms do not assign outliers to clusters.

Figure 2: Example of density-based clustering.

Distribution-based Clustering

This clustering approach assumes data is composed of distributions, such as **Gaussian distributions**. In Figure 3, the distribution-based algorithm clusters data into three Gaussian distributions. As distance from the distribution's center increases, the probability that a point belongs to the distribution decreases. The bands show that decrease in probability. When you do not know the type of distribution in your data, you should use a different algorithm.

Figure 3: Example of distribution-based clustering.

Hierarchical Clustering

Hierarchical clustering creates a tree of clusters. Hierarchical clustering, not surprisingly, is well suited to hierarchical data, such as taxonomies. See *Comparison of 61 Sequenced Escherichia coli Genomes* by Oksana Lukjancenko, Trudy Wassenaar & Dave Ussery for an example. In addition, another advantage is that any number of clusters can be chosen by cutting the tree at the right level.

Figure 4: Example of a hierarchical tree clustering animals.

WORKSHEET 3

SQL

1. Write SQL query to create table **Customers**

Ans : Create table Customers (
customerNumber float,
customerName char(50),
contactLastName char(50),
phone float,
addressLine1 char(100),
addressLine2 char(100),
city char(50),
state char(50),
postalcode int,
country char(100),
slaesRepEmpNo int,
creditLimit float);

2. Write SQL query to create table **Orders.**

Ans: Create table order (
orderNumber int,
orderDate datetime,
requiredDate datetime,
shipingDate datetime,
status text,
comment text,
customerNo int);

3. Write SQL query to show all the columns data from the **Orders** Table.
Ans : Select * from orders;

4. Write SQL query to show all the comments from the **Orders** Table.
Ans : Select * commets from orders

5. Write a SQL query to show orderDate and Total number of orders placed on that date, from **Orders** table.
Ans : select orderdate, count(orderdate) from order

6. Write a SQL query to show employeNumber, lastName, firstName of all the employees from **employees** table.
Ans : select employeNumber, latName, firstName from employees

STATISTICS WORKSHEET-3

1. Which of the following is the correct formula for total variation?
Ans : a) Total Variation = Residual Variation – Regression Variation

2. Collection of exchangeable binary outcomes for the same covariate data are called outcomes.
Ans : c) binomial

3. How many outcomes are possible with Bernoulli trial?
Ans : a) 2

4. If Ho is true and we reject it is called
Ans : a) Type-I error

5. Level of significance is also called:
a) Power of the test
b) Size of the test
c) Level of confidence
d) Confidence coefficient
Ans :

6. The chance of rejecting a true hypothesis decreases when sample size is:
Ans : b) Increase

7. Which of the following testing is concerned with making decisions using data?
a) Probability
b) Hypothesis
c) Causal
d) None of the mentioned

8. What is the purpose of multiple testing in statistical inference?
a) Minimize errors
b) Minimize false positives
c) Minimize false negatives
d) All of the mentioned

9. Normalized data are centred at and have units equal to standard deviations of the original data
Ans : a) 0

10. What Is Bayes' Theorem?

Ans : Bayes theorem, in simple words, determines the conditional probability of an event A given that event B has already occurred. Bayes theorem is also known as the Bayes Rule or Bayes Law. It is a method to determine the probability of an event based on the occurrences of prior events. It is used to calculate conditional probability. Bayes theorem calculates the probability based on the hypothesis. Now, let us state the theorem and its proof. Bayes theorem states that the conditional probability of an event A, given the occurrence of another event B, is equal to the product of the likelihood of B, given A and the probability of A. It is given as:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Here, P(A) = how likely A happens(Prior knowledge)- The probability of a hypothesis is true before any evidence is present.

P(B) = how likely B happens(Marginalization)- The probability of observing the evidence.

P(A/B) = how likely A happens given that B has happened(Posterior)-The probability of a hypothesis is true given the evidence.

P(B/A) = how likely B happens given that A has happened(Likelihood)- The probability of seeing the evidence if the hypothesis is true.

11. What is z-score?

Ans : Z score is also known as a standard score and is used to represent the number of standard deviations by which a raw score is above or below the mean. A z score is usually used as part of a z test to draw interpretations about population data. This score helps to compare data from different normal distributions.

A z score can be positive, negative, or zero depending upon the position of the raw score with respect to the mean. To determine a z score the knowledge of the population mean and the standard deviation is required. In this article, we will learn more about a z score, its formula, and how to calculate it.

To calculate a z score, knowledge of the mean and standard deviation is required.

When the population mean and population standard deviation are known then the z score formula is given as follows:

$$z = \frac{x-\mu}{\sigma}$$

μ = population mean

σ = population standard deviation

x = raw score

12. What is t-test?

Ans : The t test tells you how significant the differences between group means are. It lets you know if those differences in means could have happened by chance. The t test is usually used when data sets follow a normal distribution but you don't know the population variance.

For example, you might flip a coin 1,000 times and find the number of heads follows a normal distribution for all trials. So you can calculate the sample variance from this data, but the population variance is unknown. Or, a drug company may want to test a new cancer drug to find out if it improves life expectancy. In an experiment, there's always a control group (a group who are given a placebo, or "sugar pill"). So while the control group may show an average life expectancy of +5 years, the group taking the new drug might have a life expectancy of +6 years. It would seem that the drug might work. But it could be due to a fluke. To test this, researchers would use a Student's t-test to find out if the results are repeatable for an entire population.

In addition, a t test uses a t-statistic and compares this to t-distribution values to determine if the results are statistically significant.

However, note that you can only uses a t test to compare two means. If you want to compare three or more means, use an ANOVA instead.

13. What is percentile?

Ans : The percentile formula determines the performance of a person over others. The percentile formula is used in finding where a student stands in the test compared to other candidates. A percentile is a number where a certain percentage of scores fall below the given number.

The percentile formula is used when we need to compare the exact values or numbers over the other numbers from the given data i.e. the accuracy of the number. Often percentile and percentage are taken as one but both are different concepts. A percentage is where the fraction is considered as one term while percentile is the value below the percentage found from the given data. In our day-to-day life, percentile formulas are usually helpful in finding the test scores or biometric measurements. Hence, the percentile formula is:

Percentile = (n/N) × 100

Or

The percentile of x is the ratio of the number of values below x to the total number of values multiplied by 100. i.e., the percentile formula is

Percentile = (Number of Values Below "x" / Total Number of Values) × 100

14. What is ANOVA?
Ans: An ANOVA test is a type of statistical test used to determine if there is a statistically significant difference between two or more categorical groups by testing for differences of means using variance.

Another Key part of ANOVA is that it splits the independent variable into 2 or more groups. For example, one or more groups might be expected to influences the dependent variable while the other group is used as a control group, and is not expected to influence the dependent variable.

Assumptions of ANOVA

The assumptions of the ANOVA test are the same as the general assumptions for any parametric test:

1. An ANOVA can only be conducted if there is **no relationship between the subjects** in each sample. This means that subjects in the first group cannot also be in the second group (e.g. independent samples/between-groups).
2. The different groups/levels must have **equal sample sizes**.
3. An ANOVA can only be conducted if the dependent variable is **normally distributed**, so that the middle scores are most frequent and extreme scores are least frequent.
4. Population variances must be equal (i.e. homoscedastic). **Homogeneity of variance** means that the deviation of scores (measured by the range or standard deviation for example) is similar between populations.

15. How can ANOVA help?
Ans : Types of ANOVA Tests
There are different types of ANOVA tests. The two most common are a "One-Way" and a "Two-Way."

The difference between these two types depends on the number of independent variables in your test.

One-way ANOVA
A one-way ANOVA (analysis of variance) has one categorical independent variable (also known as a factor) and a normally distributed continuous (i.e., interval or ratio level) dependent variable.

The independent variable divides cases into two or more mutually exclusive levels, categories, or groups.

The one-way ANOVA test for differences in the means of the dependent variable is broken down by the levels of the independent variable.

An example of a one-way ANOVA includes testing a therapeutic intervention (CBT, medication, placebo) on the incidence of depression in a clinical sample.

Note: Both the One-Way ANOVA and the Independent Samples t-Test can compare the means for two groups. However, only the One-Way ANOVA can compare the means across three or more groups.

Two-way (factorial) ANOVA
A two-way ANOVA (analysis of variance) has two or more categorical independent variables (also known as a factor), and a normally distributed continuous (i.e., interval or ratio level) dependent variable.

The independent variables divide cases into two or more mutually exclusive levels, categories, or groups. A two-way ANOVA is also called a factorial ANOVA.

An example of a factorial ANOVAs include testing the effects of social contact (high, medium, low), job status (employed, self-employed, unemployed, retired), and family history (no family history, some family history) on the incidence of depression in a population.