Machine Learning_3

1. d (All of the above i.e., biological network analysis,market trend prediction, topic modeling are the application of clustering)

2. d (None i.e., we can't perform on time series data, text data & multimedia data)

3. c (Reinforcement learning only is used by Netflix's movie recommendation system. Reinforcement learning algorithms don't need any information in advance; they learn from data during process, while unsupervised learning algorithms learns patterns from untagged data)

4. b (The tree representing how close the data points are to each others)

5. d (None i.e., doesn't require a distance metric, initial number of clusters and initial guess as to cluster centroids)

6. c (k-nearest neighbor is same as k-means)

7. d (1,2,&3 i.e., Single link, Complete link, & Average link)

8. a (1 only i.e., clustering analysis is negatively affected by multicollinearity of features)

9. a (2)

10. b (Given a database of information about your users, automatically group them into different market segments)

11. a (Because for single link or min version of hierarchical clustering, the proximity of two clusters is defined to be the min of the distance between any two points in the different clusters. And from table distance between points 3&6 is 0.11 and that is the height at which they are joined into one cluster.

12. b (Because for maximum version of hierarchical clustering the proximity of two clusters is defined to be the maximum of the distance between any two points in the different clusters)

13. **Importance of clustering** :

A. Clustering is an unsupervised machine learning tool which is used to analysed the factors having a big impact on sales and customer satisfaction.

B. It is an exploratory tool which helps to discover new information and patterns in the data.

C. It is an invaluable tppl used to boost revenue, cut costs or sometimes even both.

14. I can improve my clustering performance by using independent component analysis and unsupervised feature learning.

Statistics Worksheet-3

1. b (Total variation=Residual variation+Regression variation)

2. c (Binomial)

3. a (2)

4. a (Type-I error)

5. b (Size of the test)

6. b (Increase)

7. b (Hypothesis)

8. d (All of the above i.e., minimize errors, minimize false positives, and minimize false negatives)

9. a (0)

10. **Bayes' Theorem**: It is a probability theorem which states that the conditional probability of an event, based on the occurrence of another event, is equal to the likelihood of the second event guven the first event multiplied by the probability

of the first event. It is applied in the transition probability calculations.
Formula: P(A/B)= P(B/A).P(A)/P(B)

where,

A,B= events
P(A/B)= probability of A given B is true
P(B/A)= probability of B given A is true
P(A),P(B)= The independent probabilities of A&B

11. **Z-score**: It is also known as standard score, which tells you where the score lies on a normal distribution curve. Z-score shows how far away from the mean either above or below, a value is situated. Basically it indicated how much a given value differes from standard deviation.

$$Z= x-\mu/\sigma,$$

where, Z= standard score
x= observed value
μ= mean of the sample
σ= SD of the sample

12. **t-test**: It is a statistical test that is used to compare the means of two groups. t-test often used in hypothesis testing to determine whether a process or treatment actually has an effect on the population of interest, or whether two groups are different from one another. It can only be used in pairwise comparison i.e. only between two groups.

13. **Percentile**: It is measure used in statistics. Percentile is a comparison score between a particular score and the scores of the rest of a group. It shows the percentage of scores that a particular scores surpassed.

$$n=(P/100)*N$$

where, n=ordinal rank of given value
P= percentile
N= no of values in data set

14. **ANOVA**: Stands for Analysis of Variance. It is an analytical tool used in statistics to compare variances across the means(or average) of different groups. It splits an observed aggregate variability found inside a data set inti two parts i.e. systematic factors and random factors.
The systematic factors have a statistical influence on a given data set, while the random factors do not.

15. ANOVA can help to determine the influence that independent variables have on the dependent variable in a regression study.
It is also helpful for testing one or more variables.