Machine Learning Assignment-1

1. B – 4
2. D – 1, 2 and 4
3. D – formulating the clustering problem
4. A – Euclidean distance
5. B – Divisive method
6. D – All answers are correct
7. A – Divide data points into groups
8. B – Unsupervised learning
9. A – K means clustering
10. A - K means clustering algorithm
11. D – All of the above
12. A – Labeled data
13. The cluster analysis is calculated with three steps –
    a) Calculating the distances
    b) Linking the clusters
    c) Choosing the solution by selecting the right number of clusters
14. We user the average silhouette coefficient value of all objects in the data set.
15. Cluster analysis is a method to organize data by clustering data points in a particular cluster. It is a way of putting data points with similar characteristics in one group so that they differ from other data points of other clusters.
    Types of cluster analysis are: Hard clustering and Soft clustering.

    Statistics Worksheet-1

1. A – true
2. A – central limit theorem
3. B – modelling bounded count data
4. D – all of the mentioned
5. C – poisson
6. B – false
7. B – hypothesis
8. A – 0
9. C – Outliers cannot conform to the regression relationship
10. A normal distribution is a type of continuous probability distribution in which most data points cluster towards the middle of the range, while the rest taper off symmetrically toward either extreme.
11. Missing data can be dealt with in a variety of ways. I believe the most common reaction is to ignore it. Choosing to make no decision, on the other hand, indicates that your statistical programme will make the decision for you.
    Imputation is the process of substituting an estimate for missing values and analysing the entire data set as if the imputed values were the true observed values.
    The following are some of the most prevalent methods:
    a) Mean imputation
    b) Substitution
    c) Hot deck imputation
    d) Cold deck imputation
    e) Regression imputation

f) Interpolation and extrapolation

12. A/B testing, also known as split testing, refers to a randomized experimentation process wherein two or more versions of a variable (web page, page element, etc.) are shown to different segments of website visitors at the same time to determine which version leaves the maximum impact and drives business metrics.
13. Mean imputation is typically considered terrible practice since it ignores feature correlation.
14. Linear regression analysis is used to predict the value of a variable based on the value of another variable. The variable you want to predict is called the dependent variable. The variable you are using to predict the other variable's value is called the independent variable.
15. There are three real branches of statistics: data collection, descriptive statistics and inferential statistics.