MACHINE LEARNING ASSIGNEMENT – 2

QUESTION		ANSWER
Q1.	-	A
Q2.	-	D
Q3.	-	A
Q4.	-	A
Q5.	-	B
Q6.	-	B
Q7.	-	A
Q8.	-	D
Q9.	-	A
Q10.	-	D
Q11.	-	D

Q12. Yes, K means is quite sensitive to outliers because k-means tries to optimize the sum of the squares. And thus, a large deviation such as outliers gets a lot of weight also the mean is easily influenced by extreme values. The k-means algorithm updates the cluster centres by taking the average of all the data points that are closer to each cluster centre. When all the points are packed nicely together, the average makes sense. However, when you have outliers, this can affect the average calculation of the whole cluster. As a result, this will push your cluster centre closer to the outliers.

For example: The mean of 2,2,2,3,3,3,4,4,4 is 3.

If we add 23 to that, the mean becomes 5, which is larger than any of the other values present here.

That is, (2+2+2+3+3+3+4+4+4+23) / 10 = 50/10 = 5 which is larger than other values.

Since in k-means, you'll be taking the mean a lot, you wind up a lot of outlier-sensitive calculations. That's why we have the k-medians algorithm. It just uses the median rather than the mean and less is sensitive to outliers.

Q13. K-means are relatively simple to implement and scales to large data sets. Generalizes to clusters of different shapes and sizes, such as elliptical clusters. Other clustering algorithms with better features tend to be more expensive. In this case, k-means becomes a great solution for pre-clustering, reducing the space into disjoint smaller sub-spaces where other clustering algorithms can be applied. K-means is like the Exchange Sort algorithm. Easy to understand, helps one get into the topic, but should never be used for anything real, ever. In the case of Exchange Sort is better because it can stop early if the array is partially sorted.

Q14. No, the K-means clustering is based on a non-**deterministic algorithm**. This means that running the algorithm several times on the same data, could give different results. However, to ensure consistent results, FCS Express performs k-means clustering using a deterministic method.

A deterministic algorithm is an algorithm which given a particular input, will always produce the same output, with the underlying machine always passing through the same sequence of states. Since K-mean clustering doesn't have this property. The k-means clustering algorithm attempts to split a given anonymous data set (a set containing no information as to class identity) into a fixed number (k) of clusters. Initially k number of so-called centroids is chosen. Each centroid is thereafter set to the arithmetic mean of the cluster it defines. The non-deterministic nature of K-Means is due to its random selection of data points as initial centroids. The key idea of the algorithm is to select data points which belong to dense regions and which are adequately separated in feature space as the initial centroids.

SQL ASSIGNMENT – 2

QUESTION		ANSWER
Q1.	-	D
Q2.	-	C
Q3.	-	A
Q4.	-	A
Q5.	-	B
Q6.	-	C
Q7.	-	A
Q8.	-	C
Q9.	-	B
Q10.	-	D
Q11.	-	B
Q12.	-	C
Q13.	-	A
Q14.	-	B&C
Q15.	-	A&D

STATISTICS ASSIGNMENT – 2

QUESTION		ANSWER
Q1.	-	B
Q2.	-	C
Q3.	-	D
Q4.	-	C
Q5.	-	D

Q6.	-	B
Q7.	-	A
Q8.	-	B
Q9.	-	D
Q10.	-	A
Q11.	-	C
Q12.	-	D
Q13.	-	D
Q14.	-	A
Q15.	-	D