Quizzes : Module 5

Data-Driven Identification

In the LAMS Sequence, you have learned the theory behind this module. It is also expected that you have attempted the quizzes embedded within the LAMS Sequence, and have used the “unlimited attempts” opportunity to score 100%. Here are the quiz questions, consolidated with their answers and corresponding feedback. This is for your after-LAMS revision.

Question 1

What do you think is a major problem of the k-Means algorithm? You may choose multiple options.

Answer Choice	Verdict	Explanation
The number of clusters k has to be chosen first, or guessed!	Correct	True. This is the major drawback of k-Means algorithm. In fact, this is true for most of the commonly used clustering algorithms. As they belong to the group of Unsupervised Learning, you do not know the number of clusters to start with.
The choice for initial centroids or cluster centers dictate the algorithm.	Correct	True. The k-Means algorithm is heavily dependent on the initial choice of cluster centers or centroids. If you want to make the algorithm converge faster, or more reliably, you should use a better cluster initialization strategy. Google for k-Means++.
The k-Means algorithm prefers clusters of similar size and shape.	Correct	True. An implicit assumption of k-Means is that the points belong to the cluster whose centroid is "closest" to them. This promotes the selection of "spherical" clusters, generally of equal diameter. Thus, we get clusters of similar shape and size, which may not be true in the context of practical data.
The convergence rate of k-Means algorithm is too slow.	Wrong	Actually, the convergence for k-Means is quite fast. It is one of the most preferred methods for clustering in case of big data.

Reference Module 5 Topic 1 : Clustering Patterns Slide 8

General Comment : Go through k-Means algorithm once again in the lecture, and try to identify as many implicit assumptions as you can find. The assumptions in an algorithm are commonly the source of major drawbacks, as they prevent the algorithms from generalizing.

Question 2

Suppose you start the k-Means algorithm with all the cluster centers initialized to the same value. What will happen?

Answer Choice	Verdict	Explanation
The algorithm will end up identifying only one cluster, with all points.	Correct	True. Initializing all cluster centers to the same value is technically identical to choosing a single cluster center.
The convergence of the algorithm to the k clusters will be really slow.	Wrong	Wrong. The algorithm will converge in a single iteration, to a single cluster, containing all the points.
Each point in the dataset will be assigned to all k clusters with equal probability.	Wrong	Wrong. k-Means algorithm deterministically assigns points to their nearest cluster. Thus, there is no probabilistic assignment of points possible in this procedure. All the points will be deterministically assigned to a single cluster in this case.

Reference Module 5 Topic 1 : Clustering Patterns Slide 6 and Slide 7 and Slide 8

General Comment : Go through the k-Means algorithm in the lecture, once again, and set all cluster centers to the same value to start with. Run through the first one or two steps of the algorithm manually to check what happens. You will find something amusing.

Question 3

What will be the Within Sum of Squares if k is equal to the number of points in the data?

Answer Choice	Verdict	Explanation
Zero	Correct	True. In this case, each point is its own cluster and its own cluster center. Hence, the Within SS should be 0.
Total Sum of Squares	Wrong	No. It can't be the total sum of squares, as that is the case only for k = 1. In all other cases, it decreases.
Optimal	Wrong	No. This choice of k is not optimal, as it goes against the core motive of clustering. If you allocate every point to its own cluster, then there is no point in finding common structures in the data. This is similar to "overfitting" in prediction.

Reference Module 5 Topic 1 : Clustering Patterns Slide 9 and Slide 10

General Comment : This is a nice thought experiment -- try visualizing the case where number of clusters is same as the number of points. You may do it with other scenarios too, say, when there is only one cluster of all the points.

Question 4

In case of Local Outlier Factor algorithm, what will happen if we set K = 1?

Answer Choice	Verdict	Explanation
K = 1 emphasizes local anomalies within the data, even if they are not at the boundary.	Correct	True. Think of a point right in the middle of the dataset, which does not have too many neighbors. It's like a bald patch within the data, with a single point in it. With K = 1, such a point will be identified as an anomaly. With large K, you will miss it.
K = 1 diminishes local anomalies within the data, and only identifies those at the boundary.	Wrong	Wrong. Think of a point right in the middle of the dataset, which does not have too many neighbors. It's like a bald patch within the data, with a single point in it. With K = 1, such a point will be identified as an anomaly.
Value of K has no effect on anomaly detection in the Local Outlier Factor algorithm.	Wrong	In fact, value of K has a lot of effect on anomaly detection. It controls which anomalies to highlight.

Reference Module 5 Topic 2 : Anomaly Detection Slide 4 and Slide 5 and Slide 6

General Comment : This is not obvious from the lecture. Try to think about it deeply, and read up on the Internet about the algorithm, if you wish. Note that K is the number of neighbors you are considering for outlier detection.

Question 5

What is the relation between outliers of two Uni-Variate datasets and the anomalies in their joint Bi-Variate dataset? That is, what is the relation between outliers in the boxplots of X and Y with the anomalies in the jointplot of X-Y?

Answer Choice	Verdict	Explanation
Points NOT tagged as outliers in individual variables X and Y may be tagged anomalies in the X-Y case.	Correct	True. There may be a point that's equal to the median in X, but is a huge outlier in Y. This point will most likely be tagged as an anomaly in the X-Y joint distribution. Try picturing such an example yourself, and you'll know.
Outliers in individual variables X and Y will not be tagged as anomalies in the bi-variate jointplot of X-Y.	Wrong	Generally, the outliers in X and Y (individually) have a high probability of becoming anomalies in the bi-variate case too. Can you think of a case where it wouldn't? That is, can you construct a hypothetical case where outliers in X do not become anomalies in the X-Y bivariate distribution?

Reference Module 5 Topic 2 : Anomaly Detection Slide 9

General Comment : This is not obvious from the lecture. Try to think about it deeply, and read up on the Internet about the algorithm, if you wish. Think about what outliers mean along each axis, and in general over multiple axes.