Review Question 1

(1/1 point)

Only one of following statements is true. Which one is it?

- (selected) Its possible for samples from two different clusters to be more similar to one another than their intra-cluster neighbors, if the the two clusters are large and located near one another ✔
- Real world data typically comes labeled
- Unsupervised clustering aims to group your samples based on their labels
- Centroids are records that live in your dataset and share the same feature space so that a meaningful distance can be calculated between them and your samples

EXPLANATION

Lab

This was a tricky problem. Real world data doesn't usually come labeled. Also unsupervised algorithms don't have or use labels either. Lastly, your centroids don't live within your dataset, they only share its feature space.

The correct answer is that although disliked, samples from separate clusters can actually be more similar than their own intra-cluster samples under specific circumstances. Consider the following two clusters:

(A_____._____B)(C_____._____D)

The centroids are the marked with '.' and you can examine samples A and B, both part of the first cluster, and samples C and D part of the second cluster. Due to the close proximity of the clusters, B and C are actually closer or more similar to one another than either A and B, or C and D are!

You have used 1 of 2 submissions

Review Question 2

(1/1 point)

Once again, only a single one of the following statements is correct. Do you know which one it is?

- It's possible for a sample to be assigned to two clusters; but only if its equidistant from either cluster.
- The K-Means algorithm scans your dataset to detect clusters using an iterative assignment / update cycle. The algorithm returns the number of clusters found, as well as their centroid position.

- As a clustering algorithm, K-Means is really only useful for grouping your samples
- K-Means assumes your features are either length normalized, or that their length encodes a specific meaning. ✔

EXPLANATION

A sample can only have a single cluster assignment.

Also, you are the one responsible for specifying the number of clusters. K-Means won't tell you the number of clusters in your data.

Besides assigning a cluster to your samples, there are many uses for the centroid locations. Review th reading please.

Since K-Means cluster assignment depends on an Euclidean length metric, your features have to either be length normalized, or have appropriate units for the algorithm to perform properly.

You have used 1 of 2 submissions