Unsupervised Learning

1. For which of the following tasks might K-means clustering be a suitable algorithm? Select all that apply.

> Given a database of information about your users, automatically group them into different market segments.

Explanation: You can use K-means to cluster the database entries, and each cluster will correspond to a different market segment.

> Given sales data from a large number of products in a supermarket, figure out which products tend to form coherent groups (say are frequently purchased together) and thus should be put on the same shelf.

Explanation: If you cluster the sales data with K-means, each cluster should correspond to coherent groups of items.

- Given historical weather records, predict the amount of rainfall tomorrow (this would be a real-valued output)
- Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

2. Suppose we have three cluster centroids $\mu_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} -3 \\ 0 \end{bmatrix}$ and $\mu_3 = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$. Furthermore, we have a training example $x^{(i)} = \begin{bmatrix} 3 \\ 1 \end{bmatrix}$. After a cluster assignment step, what will $c^{(i)}$ be?

- $c^{(i)} = 2$

> $c^{(i)} = 3$

- $c^{(i)} = 1$
- $c^{(i)}$ is not assigned

Explanation: $x^{(i)}$ is closest to μ_3, so $c^{(i)} = 3$

3. K-means is an iterative algorithm, and two of the following steps are repeatedly carried out in its inner-loop. Which two?

- The cluster centroid assignment step, where each cluster centroid μ_i is assigned (by setting $c^{(i)}$ to the closest training example $x^{(i)}$.

> The cluster assignment step, where the parameters $c^{(i)}$ are updated.

Explanation: This is the correst first step of the K-means loop.

- Move each cluster centroid μ_k, by setting it to be equal to the closest training example $x^{(i)}$

> Move the cluster centroids, where the centroids μ_k are updated.

Explanation: The cluster update is the second step of the K-means loop.

4. Suppose you have an unlabeled dataset $\{x^{(1)}, \ldots, x^{(m)}\}$. You run K-means with 50 different random initializations, and obtain 50 different clusterings of the data. What is the recommended way for choosing which one of these 50 clusterings to use?

> Compute the distortion function $J(c^{(1)}, \ldots, c^{(m)}, \mu_1, \ldots, \mu_k)$, and pick the one that minimizes this.

- Use the elbow method.
- Manually examine the clusterings, and pick the best one.
- Plot the data and the cluster centroids, and pick the clustering that gives the most "coherent" cluster centroids.

Explanation: A lower value for the distortion function implies a better clustering, so you should choose the clustering with the smallest value for the distortion function.

5. Which of the following statements are true? Select all that apply.

- The standard way of initializing K-means is setting $\mu_1 = \cdots = \mu_k$ to be equal to a vector of zeros.

> If we are worried about K-means getting stuck in bad local optima, one way to ameliorate (reduce) this problem is if we try using multiple random initializations.

Explanation: Since each run of K-means is independent, multiple runs can find different optima, and some should avoid bad local optima.

- Since K-Means is an unsupervised learning algorithm, it cannot overfit the data, and thus it is always better to have as large a number of clusters as is computationally feasible.

> For some datasets, the "right" or "correct" value of K (the number of clusters) can be ambiguous, and hard even for a human expert looking carefully at the data to decide.

Explanation: In many datasets, different choices of K will give different clusterings which appear quite reasonable. With no labels on the data, we cannot say one is better than the other.