HMC CS 158
Quiz 8.5: Clustering

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

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

   (b) Given historical weather records, predict if tomorrow's weather will be sunny or rainy.

   (c) From the user usage patterns on a website, figure out what different groups of users exist.

   (d) Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.

   (e) Given many emails, you want to determine if they are spam or non-spam emails.

   (f) Given a set of news articles from many different news websites, find out what are the main topics covered.

2. Suppose we have three cluster centroids $\boldsymbol{\mu}_1 = \binom{1}{2}$, $\boldsymbol{\mu}_2 = \binom{-3}{0}$, and $\boldsymbol{\mu}_3 = \binom{4}{2}$. Furthermore, we have a training example $\boldsymbol{x}^{(i)} = \binom{-2}{1}$. After a cluster assignment step, what will $c^{(i)}$ be?

3. Suppose you have an unlabeled dataset $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}\}$. 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?

   (a) Always pick the final (50^{th}) clustering found, since by that time, it is more likely to have converged to a good solution.

   (b) For each of the clusterings, compute $\frac{1}{n}\sum_{i=1}^{n} ||\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_{c^{(i)}}||^2$, and pick the one that minimizes this.

   (c) Use the elbow method.

   (d) Plot the data and the cluster centroids, and pick the clustering that gives the most "coherent" cluster centroids.

   (e) The answer is ambiguous, and there is no good way of choosing.

This quiz is adapted from course material by Andrew Ng (Stanford).

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

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

   (b) The standard way of initializing k-means is setting $\boldsymbol{\mu}_1 = \ldots = \boldsymbol{\mu}_k$ to be equal to a vector of zeros.

   (c) A good way to initialize k-means is to select k (distinct) examples from the training set and set the cluster centroids equal to these selected examples.

   (d) k-means will always give the same results regardless of the initialization of the centroids.

   (e) Once an example has been assigned to a particular centroid, it will never be reassigned to another different centroid.

   (f) On every iteration of k-means, the cost function $J(c^{(1)}, \ldots, c^{(n)}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)$ (the distortion function) should either stay the same or decrease; in particular, it should not increase.