Week 8-1: Unsupervised Learning

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 marketing segments.

   (b) 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.

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

   To find which cluster $x^{(i)}$ will be assigned to, we must find the distance between $x^{(i)}$ and each of the centroids. The shortest distance will be the resulting cluster group.
   $||\mu_1 - x^{(i)}|| = \sqrt{(1-3)^2 + (2-1)^2} = \sqrt{(-2)^2 + (1)^2} = \sqrt{4+1} = \sqrt{5}$
   $||\mu_2 - x^{(i)}|| = \sqrt{37}$
   $||\mu_3 - x^{(i)}|| = \sqrt{2}$. We observe that $\sqrt{2}$ is the lowest and thus $c^{(i)} = \boxed{\mu_3} = 3$

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

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

   (b) Move the cluster centroids, where the centroids μ_k are updated.

4. Suppose you have an unlabeled dataset $\{x^{(1)}, ..., x^{(m)}\}$. 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?

   For each of the clusterings, compute $\frac{1}{m} \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||$ and pick the one that minimizes this.

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

   (a) A good way to intialize K-means is to select K (distinct) examples from the training set and set the cluster centroids equal to these selected examples.

   (b) On every iteratin of K-means, the cost function $J(c^{(1)}, ..., c^{(m)}, \mu_1., , , .\mu_k)$ (the distortion function) should either stay the same or decrease; in particular, it should not increase.