WEEK 8 QUIZ 1

Unsupervised Learning

LATEST SUBMISSION GRADE

80%

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

- [x] Given a database of information about your users, automatically group them into different market segments.

  ✓ **Correct**

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

- [x] 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.

  ✓ **Correct**

  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?

- ○ $c^{(i)}$ is not assigned
- ○ $c^{(i)} = 1$
- ○ $c^{(i)} = 2$
- ◉ $c^{(i)} = 3$

✓ **Correct**

$x^{(i)}$ 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?

- [x] The cluster assignment step, where the parameters $c^{(i)}$ are updated.

  ✓ **Correct**
  This is the correst first step of the K-means loop.

- [x] Move the cluster centroids, where the centroids μ_k are updated.

  ✓ **Correct**
  The cluster update is the second step of the K-means loop.

- [ ] Using the elbow method to choose K.

- [ ] Feature scaling, to ensure each feature is on a comparable scale to the others.

---

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?

- Always pick the final (50th) clustering found, since by that time it is more likely to have converged to a good solution.
- The answer is ambiguous, and there is no good way of choosing.
- For each of the clusterings, compute $\frac{1}{m} \sum_{i=1}^{m} ||x^{(i)} - \mu_{c^{(i)}}||^2$, and pick the one that minimizes this.
- The only way to do so is if we also have labels $y^{(i)}$ for our data.

✓ **Correct**

This function is the distortion function. Since a lower value for the distortion function implies a better clustering, 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.

- [ ] 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.

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

  ! **This should not be selected**

  This is a poor initialization, since every centroid needs to start in a different location. Otherwise, each will be updated in the same way at each iteration and they will never spread out into different clusters.

- [x] 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.

  ✓ **Correct**

  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.