

Feedback — XIII. Clustering

[Help Center](#)

You submitted this quiz on **Wed 8 Apr 2015 9:00 AM CEST**. You got a score of **5.00** out of **5.00**.

Question 1

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

Your Answer	Score	Explanation
<input type="checkbox"/> Given sales data from a large number of products in a supermarket, estimate future sales for each of these products.	✓ 0.25	Such a prediction is a regression problem, and K-means does not use labels on the data, so it cannot perform regression.
<input checked="" type="checkbox"/> 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.	✓ 0.25	If you cluster the sales data with K-means, each cluster should correspond to coherent groups of items.
<input type="checkbox"/> Given many emails, you want to determine if they are Spam or Non-Spam emails.	✓ 0.25	Classifying input as spam / non-spam requires labels for the data, which K-means does not use.
<input checked="" type="checkbox"/> Given a database of information about your users, automatically group them into different market segments.	✓ 0.25	You can use K-means to cluster the database entries, and each cluster will correspond to a different market segment.
Total	1.00 / 1.00	

Question 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 $x^{(i)} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}$. After a cluster assignment step, what will $c^{(i)}$ be?

Your Answer	Score	Explanation
<input type="radio"/> $c^{(i)} = 1$		
<input checked="" type="radio"/> $c^{(i)} = 2$	✓ 1.00	$x^{(i)}$ is closest to μ_2 , so $c^{(i)} = 2$
<input type="radio"/> $c^{(i)}$ is not assigned		
<input type="radio"/> $c^{(i)} = 3$		
Total	1.00 / 1.00	

Question 3

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

Your Answer	Score	Explanation
<input checked="" type="checkbox"/> Move the cluster centroids, where the centroids μ_k are updated.	✓ 0.25	The cluster update is the second step of the K-means loop.
<input type="checkbox"/> Test on the cross-validation set.	✓ 0.25	Any sort of testing is outside the scope of the K-means algorithm itself.
<input type="checkbox"/> The cluster centroid assignment step, where each cluster centroid μ_i is assigned (by setting $c^{(i)}$) to the closest training example $x^{(i)}$.	✓ 0.25	This is not a correct description of the cluster assignment step.
<input checked="" type="checkbox"/> The cluster assignment step, where the parameters $c^{(i)}$ are updated.	✓ 0.25	This is the correct first step of the K-means loop.
Total	1.00 / 1.00	

Question 4

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

Your Answer	Score	Explanation
<input type="radio"/> Manually examine the clusterings, and pick the best one.		
<input type="radio"/> Plot the data and the cluster centroids, and pick the clustering that gives the most "coherent" cluster centroids.		
<input checked="" type="radio"/> 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.	✓ 1.00	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.
<input type="radio"/> Use the elbow method.		
Total	1.00 / 1.00	


Question 5


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

Your Answer	Score	Explanation
<input checked="" type="checkbox"/> 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.	✓ 0.25	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.
<input checked="" type="checkbox"/> If we are worried	✓ 0.25	Since each run of K-means is independent, multiple

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

runs can find different optima, and some should avoid bad local optima.

<input type="checkbox"/> The standard way of initializing K-means is setting $\mu_1 = \dots = \mu_k$ to be equal to a vector of zeros.	 0.25	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.
--	--	---

<input type="checkbox"/> Once an example has been assigned to a particular centroid, it will never be reassigned to another different centroid	 0.25	Each iteration of K-means performs a cluster assignment step in which each example may be assigned to a different centroid.
--	--	---

Total	1.00 / 1.00
-------	----------------