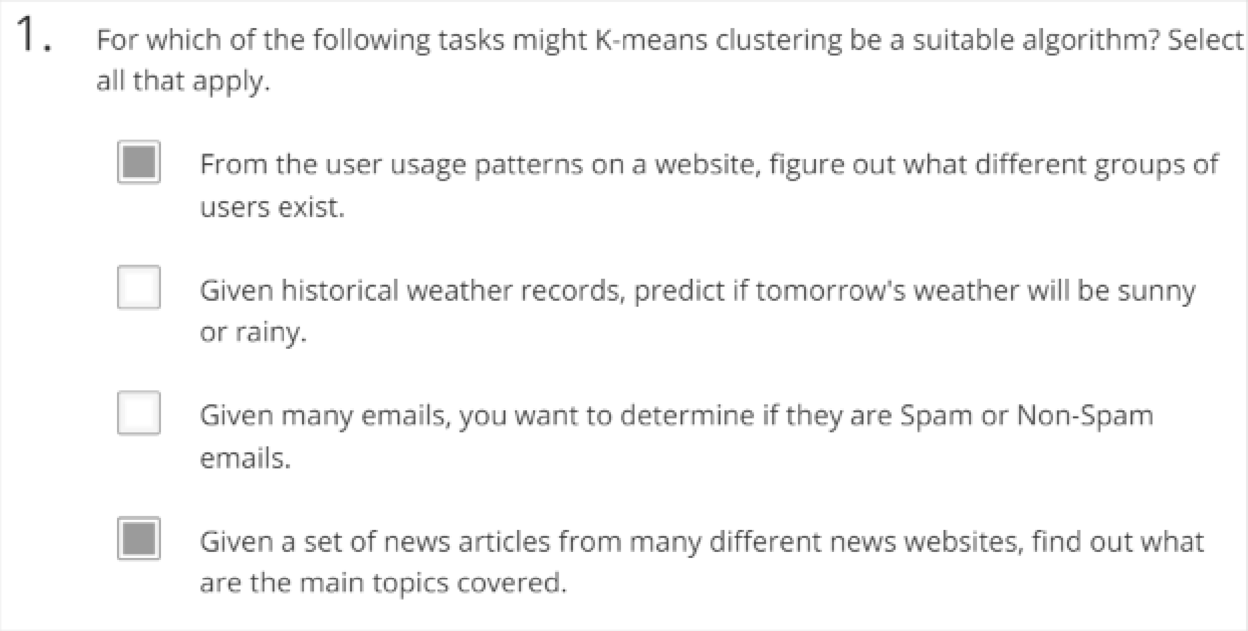
W8-1 Unsupervised Learning

Sunday, September 11, 2016 7:52 PM

Right: 1, 3, 4, 5



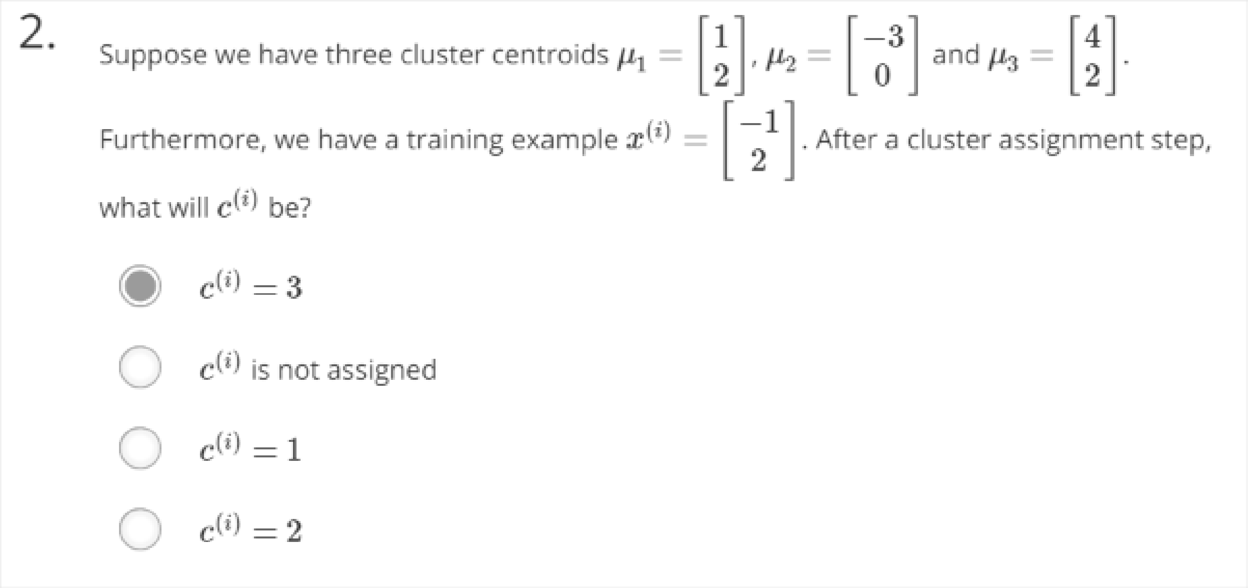
1. For which of the following tasks might K-means clustering be a suitable algorithm? Select all that apply.

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

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

□ Given many emails, you want to determine if they are Spam or Non-Spam emails.

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



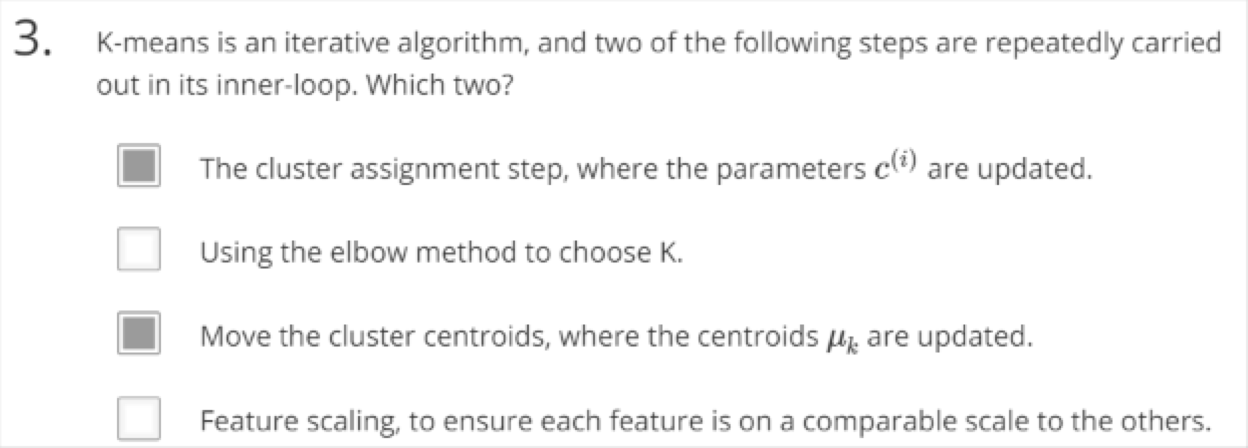
2. Suppose we have three cluster centroids $\mu_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$, $\mu_2 = \begin{bmatrix} -3 \\ 0 \end{bmatrix}$ and $\mu_3 = \begin{bmatrix} 4 \\ 2 \end{bmatrix}$. Furthermore, we have a training example $x^{(i)} = \begin{bmatrix} -1 \\ 2 \end{bmatrix}$. After a cluster assignment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

● $c^{(i)} = 3$

○ $c^{(i)}$ is not assigned

○ $c^{(i)} = 1$

○ $c^{(i)} = 2$



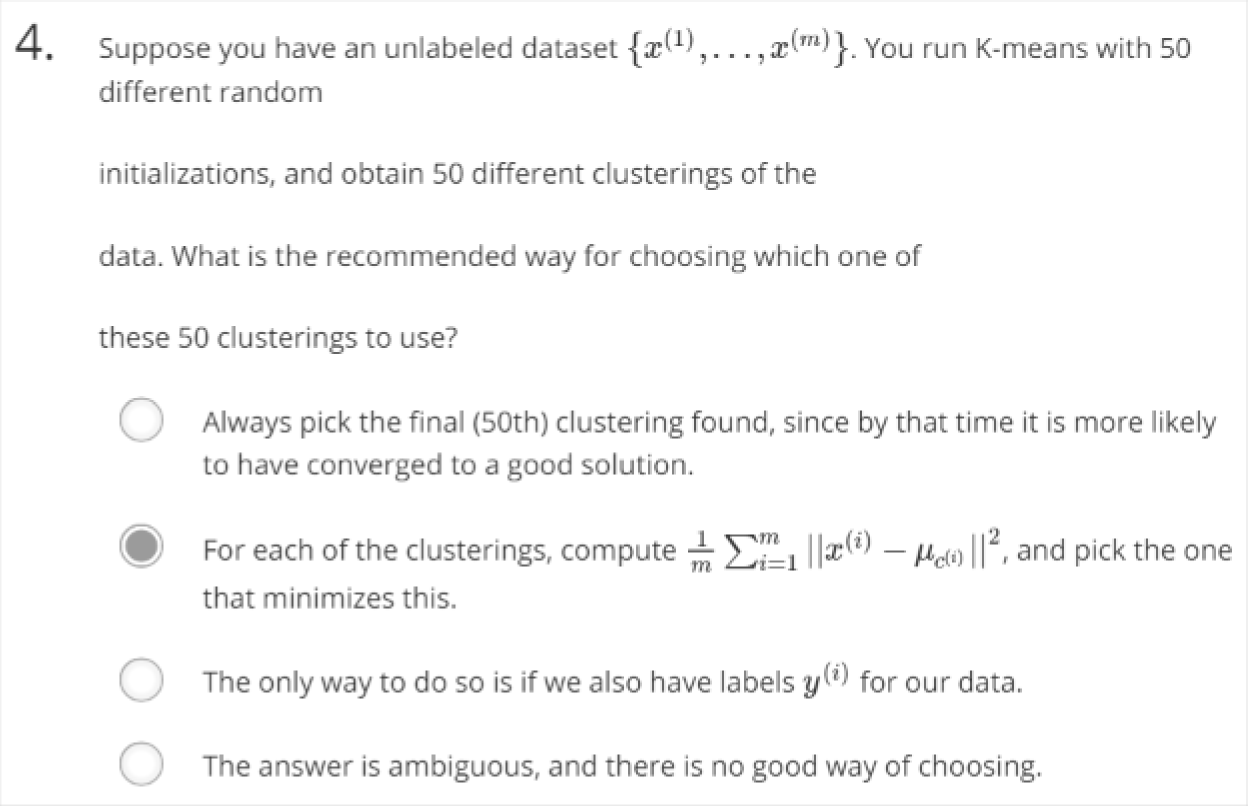
3. K-means is an iterative algorithm, and two of the following steps are repeatedly carried out in its inner-loop. Which two?

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

□ Using the elbow method to choose K.

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

□ 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

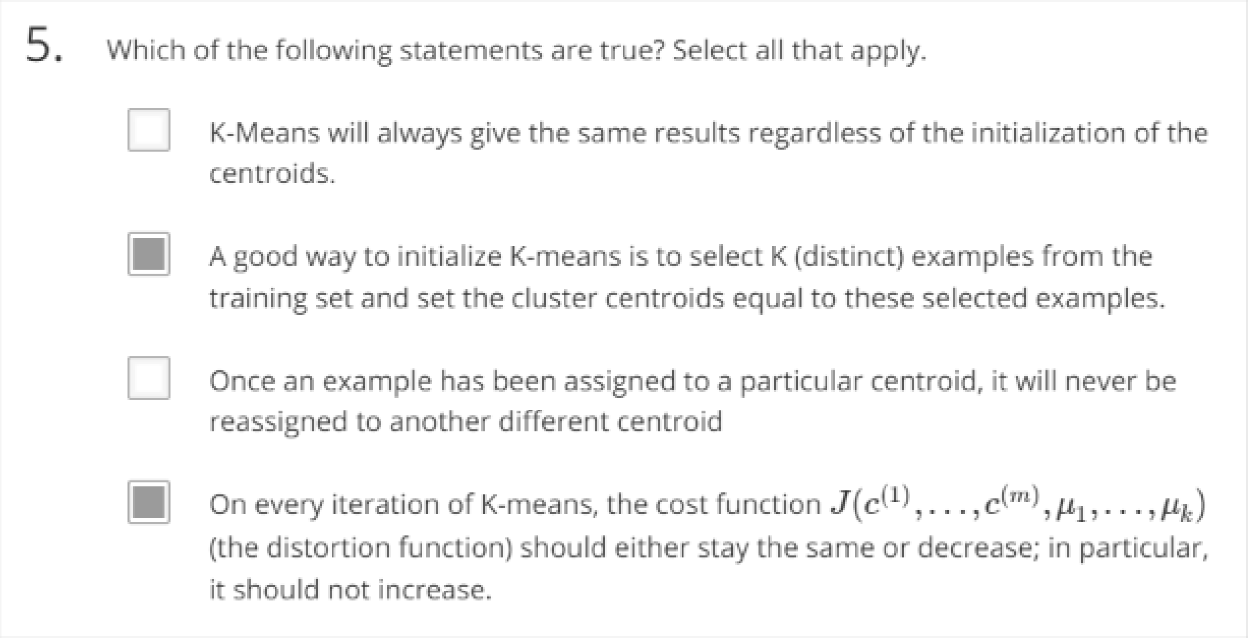
these 50 clusterings to use?

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

● 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.

○ The answer is ambiguous, and there is no good way of choosing.



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

□ K-Means will always give the same results regardless of the initialization of the centroids.

■ 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.

□ Once an example has been assigned to a particular centroid, it will never be reassigned to another different centroid

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