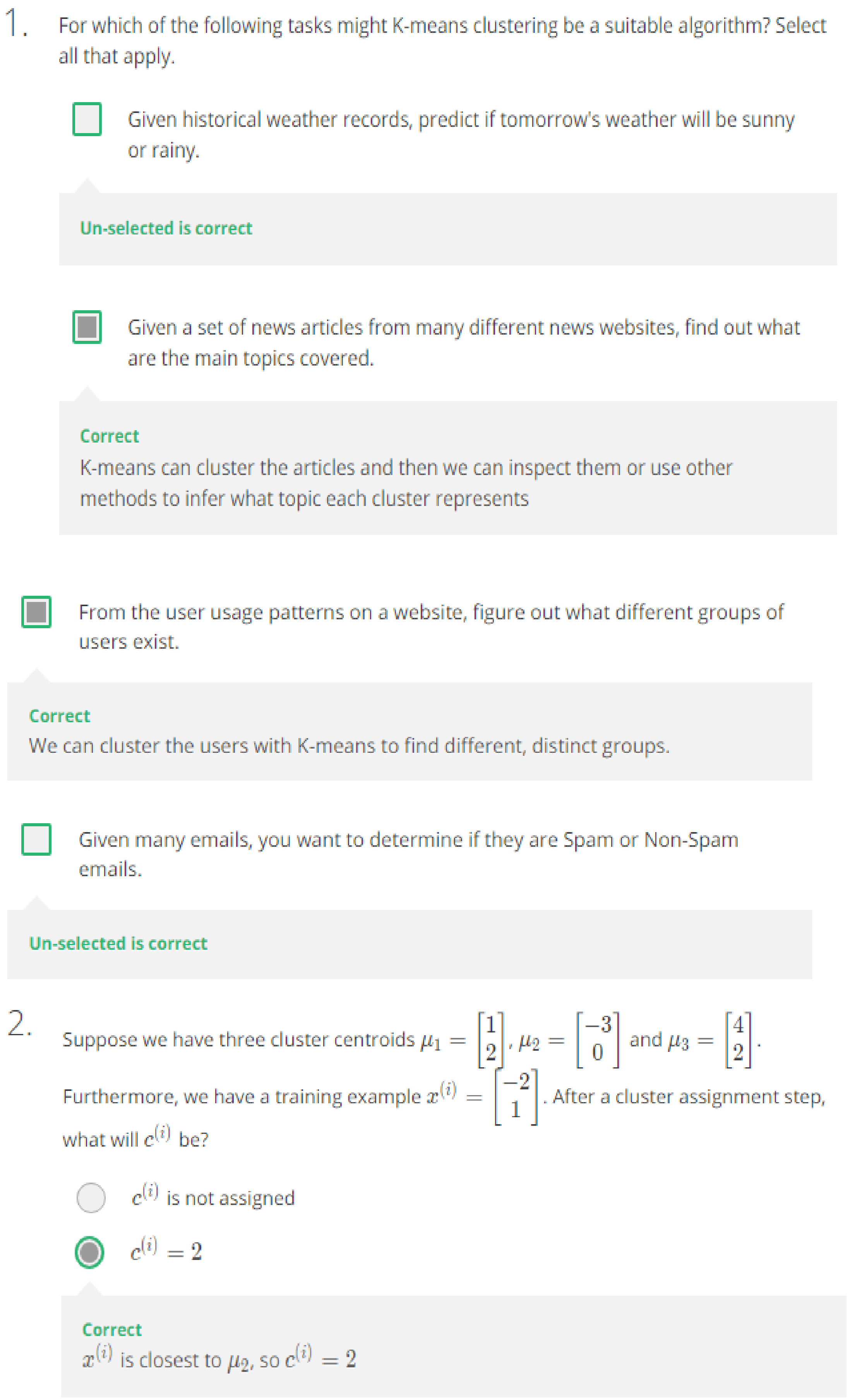


1. For which of the following tasks might K-means clustering be a suitable algorithm? Select all that apply.

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

Un-selected is correct

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

Correct
K-means can cluster the articles and then we can inspect them or use other methods to infer what topic each cluster represents

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

Correct
We can cluster the users with K-means to find different, distinct groups.

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

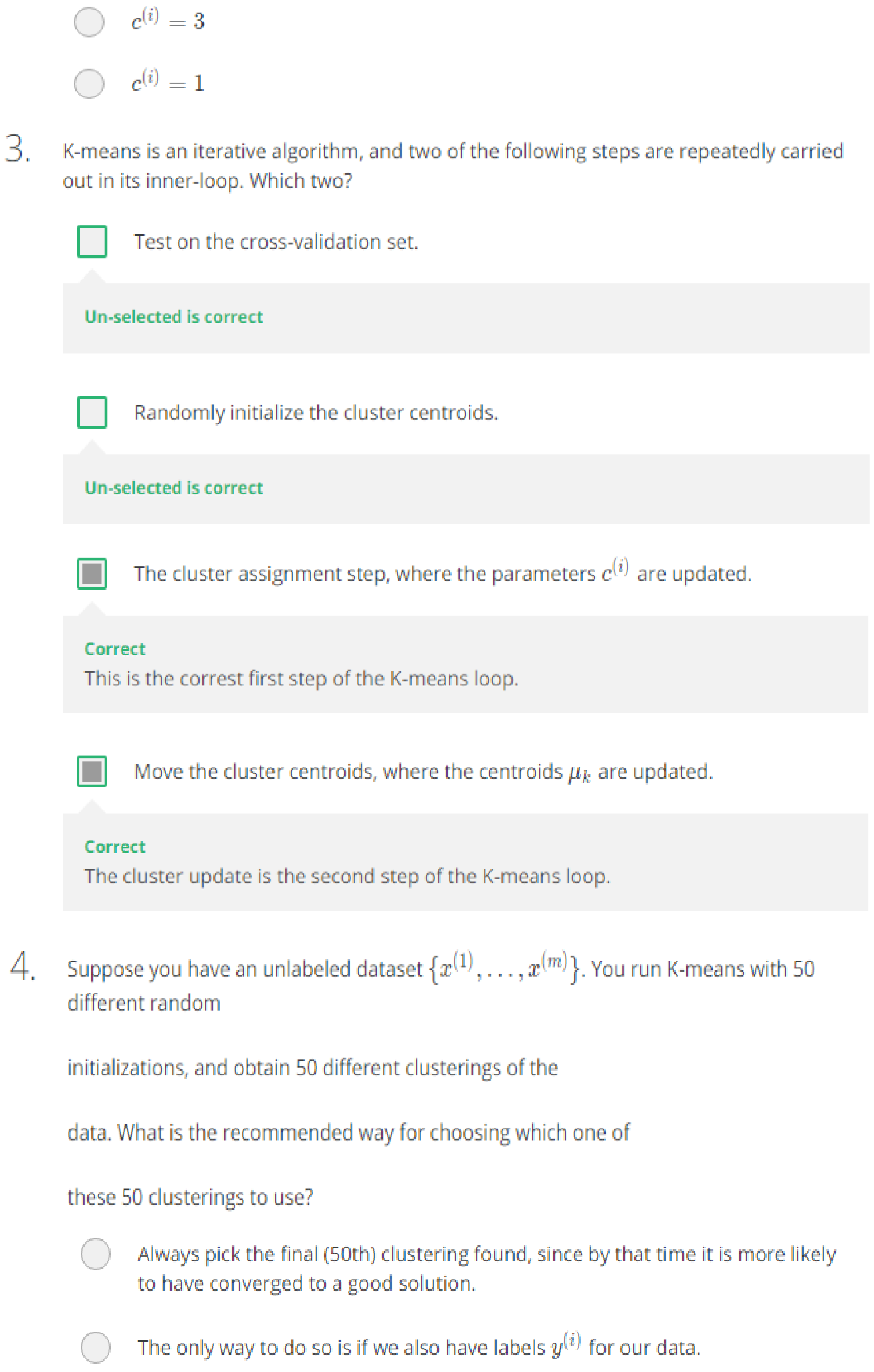
Un-selected is correct

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?

○ $c^{(i)}$ is not assigned

● $c^{(i)} = 2$

Correct
$x^{(i)}$ is closest to μ_2, so $c^{(i)} = 2$



○ $c^{(i)} = 3$

○ $c^{(i)} = 1$

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

□ Test on the cross-validation set.

Un-selected is correct

□ Randomly initialize the cluster centroids.

Un-selected is correct

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

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

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

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

4. Suppose you have an unlabeled dataset $\{x^{(1)}, \ldots, x^{(m)}\}$. You run K-means with 50 different random

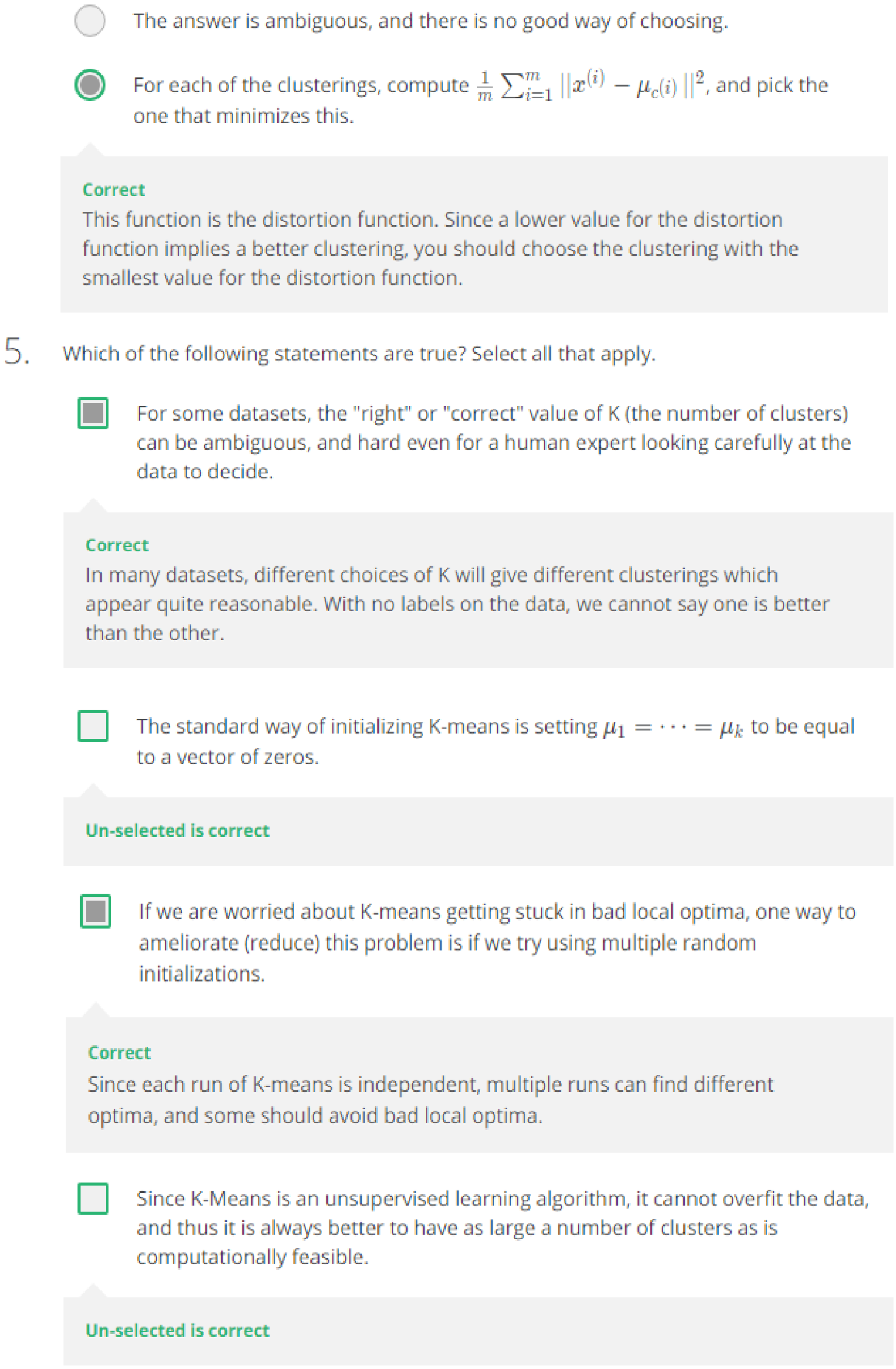
initializations, and obtain 50 different clusterings of the

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 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.

● 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.

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.

Correct
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.

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

Un-selected is correct

☑ 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.

Un-selected is correct