CS 4700: Foundations of Artificial Intelligence
Spring 2020
Review questions
Quiz 10

1. True/False: If k-means clustering is given a set of data and halts with some resulting final set of k clusters, you get the same set of clusters if each example instead has double the number of attributes, where each of the original attributes is redundantly duplicated as a second attribute. Assume in both cases you start with the same initial centroids, all ties are broken identically, and clusters are identified by which examples are in which clusters (in other words the fact that the examples in the second case look different – have double the number of attributes – is not important).

2. Consider the following data set of six items, each described by two real-valued attributes: (0,1), (1,0), (1,2), (3,0), (3,2), and (4,1). Simulate k-means clustering on this data set with k=2. Assume the first two points used as the initial centroids are (0,1) and (1,0). For each iteration show the two centroids that are generated and to which of the centroids each instance is assigned to.

3. Give an example where k-means clustering forms different clusters depending on which data points are randomly chosen as the initial centroids for each cluster. Your example should not rely on how you break ties if you have multiple points that are the same distant from multiple centroids. (Or, stated differently, you can make this happen even if the distance between every pair of points is distinct.)

4. k-means clustering can be formulated as an algorithm that assigns points to clusters so that the sum of the distances between each point and the centroid it's assigned to is minimized:

$$\underset{S_1,S_2,\ldots,S_k}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} (x - c_i)^2$$

where $\bigcup_{i=1}^{k} S_i = D$ (the sets collectively include all the data), $S_i \cap S_j = \emptyset \; i \neq j$ (the sets are disjoint), and c_i is the centroid for cluster *i*.

   a. Describe the k-means clustering algorithm using one of the search methods from the first part of the class. Make sure to specify what are the states, operators, etc.

   b. Give an example (a set of data, a value for k, and starting centroids) for which k-means clustering will not yield an optimal clustering. (*Not* that it won't find optimal clusterings in some cases; you want a data set and a value of k for which if you start with the wrong random centroids the result for that case is not optimal.)

   c. Give an example where k-means clustering *is* guaranteed to find an optimal clustering no matter what starting points it is given. Assume k=2 and that your example must have at least four examples.

5. Imagine k-means clustering were given the following set of data:

	x_1	x_2	x_3
1	1	1	1
2	0	0	0
3	0	1	1
4	1	1	0
5	0	0	1
6	0	1	0
7	1	0	0
8	0	0	1

If k=2 and where the initial starting points are the first two examples (1,1,1) and (0,0,0). Which points are in each cluster when it halts? How many iterations does it take?

1. True/False: If k-means clustering is given a set of data and halts with some resulting final set of k clusters, you get the same set of clusters if each example instead has double the number of attributes, where each of the original attributes is redundantly duplicated as a second attribute. Assume in both cases you start with the same initial centroids, all ties are broken identically, and clusters are identified by which examples are in which clusters (in other words the fact that the examples in the second case look different – have double the number of attributes – is not important).

   True. If point a is closer to b than c in the original data that remains the case when the attributes are doubled. Since the centroids each point is closest to is the same either way, the results turn out the same.

2. Consider the following data set of six items, each described by two real-valued attributes: (0,1), (1,0), (1,2), (3,0), (3,2), and (4,1). Simulate k-means clustering on this data set with k=2. Assume the first two points used as the initial centroids are (0,1) and (1,0). For each iteration show the two centroids that are generated and to which of the centroids each instance is assigned to.

Iteration	c_1	c_2	Which cluster this point is assigned to					
			(0,1)	(1,0)	(1,2)	(3,0)	(3,2)	(4,1)
0	(0,1)	(1,0)	1	2	1	2	2	2
1	(1/2,3/2)	(11/4,3/4)	1	1	1	2	2	2
2	(2/3,1)	(10/3,1)	1	1	1	2	2	2
No change to cluster assignments, so stop								

3. Give an example where k-means clustering forms different clusters depending on which data points are randomly chosen as the initial centroids for each cluster. Your example should not rely on how you break ties if you have multiple points that are the same distant from multiple centroids. (Or, stated differently, you can make this happen even if the distance between every pair of points is distinct.)

   If the data set contains:

   A: (0,0)
   B: (0,1)
   C: (1,0)
   D: (1,1)

   If you do k-means clustering with k=2 and start with A and B as the centroids you form a cluster containing A and C and a cluster containing B and D. If you start with A and C as the centroids you form a cluster containing A and B and a cluster containing C and D.

4. k-means clustering can be formulated as an algorithm that assigns points to clusters so that the sum of the distances between each point and the centroid it's assigned to is minimized:

$$\underset{S_1, S_2, \ldots, S_k}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{x \in S_i} (x - c_i)^2$$

   where $\bigcup_{i=1}^{k} S_i = D$ (the sets collectively include all the data), $S_i \cap S_j = \emptyset \; i \neq j$ (the sets are disjoint), and c_i is the centroid for cluster *i*.

a. Describe the k-means clustering algorithm using one of the search methods from the first part of the class. Make sure to specify what are the states, operators, etc.

It does hill climbing. There are two ways you could formulate this, depending on whether you consider (1) the centroids as the states or (2) the partition of the data into subsets as the states.

(1) You start with a random state that corresponds to picking a different data point as the centroid for each of the k clusters. The operators take the average of the data points closest to each centroid to generate a new state. You keep doing this as long as the formula above continues to decrease.

(2) You select k points as centroids and then assign each data point to the set defined by the centroid assigned to it. That's the initial state. Subsequent states are generated by computing as new centroids the average of all the points in each set and then assigning data points to sets based on the centroid it is closest to. You keep doing this as long as the formula above continues to decrease.

b. Give an example (a set of data, a value for k, and starting centroids) for which k-means clustering will not yield an optimal clustering. (*Not* that it won't find optimal clusterings in some cases; you want a data set and a value of k for which if you start with the wrong random centroids the result for that case is not optimal.)

If the data set contains:

A: (0,0)
B: (0,1)
C: (2,0)
D: (2,1)

If you start with A and B as initial centroids you wind up with A and C in one cluster and B and D in a second. The formula above gives you the value of 2. This is suboptimal. If you start with A and C as the initial centroids you wind up with A and B in one cluster and C and D in a second cluster. The formula for this gives you 1.

c. Give an example where k-means clustering *is* guaranteed to find an optimal clustering no matter what starting points it is given. Assume k=2 and that your example must have at least four examples.

If the data set contains:

A: (0,0)
B: (1,1)
C: (4,0)
D: (5,1)

There are six cases:

1. If you start with A and B as initial centroids c_1 and c_2 you get clusters S_1 = {A} and S_2 = {B,C,D}. The centroids on the next iteration would be (0,0) and (10/3,2/3), with S_1 = {A,B} and S_2 = {C,D}. On the next iteration the centroids are (1/2,1/2) and (9/2,1/2) , with S_1 = {A,B} and S_2 = {C,D}. Since the clusters haven't changed we stop.
2. If you start with A and C as initial centroids c_1 and c_2 you get clusters S_1 = {A,B} and S_2 = {C,D}. On the next iteration the centroids are (1/2,1/2) and (9/2,1/2) , with S_1 = {A,B} and S_2 = {C,D}. Since the clusters haven't changed we stop.

3. If you start with A and D as initial centroids c_1 and c_2 you get clusters S_1 = {A,B} and S_2 = {C,D}. On the next iteration the centroids are (1/2,1/2) and (9/2,1/2) , with S_1 = {A,B} and S_2 = {C,D}. Since the clusters haven't changed we stop.
4. If you start with B and C as initial centroids c_1 and c_2 you get clusters S_1 = {A,B} and S_2 = {C,D}. On the next iteration the centroids are (1/2,1/2) and (9/2,1/2) , with S_1 = {A,B} and S_2 = {C,D}. Since the clusters haven't changed we stop.
5. If you start with B and D as initial centroids c_1 and c_2 you get clusters S_1 = {A,B} and S_2 = {C,D}. On the next iteration the centroids are (1/2,1/2) and (9/2,1/2) , with S_1 = {A,B} and S_2 = {C,D}. Since the clusters haven't changed we stop.
6. If you start with C and D as initial centroids c_1 and c_2 you get clusters S_1 = {A,B,C} and S_2 = {D}. The centroids on the next iteration would be (5/3,1/3) and (5,1), with S_1 = {A,B} and S_2 = {C,D}. On the next iteration the centroids are (1/2,1/2) and (9/2,1/2) , with S_1 = {A,B} and S_2 = {C,D}. Since the clusters haven't changed we stop.

In all four cases we wind up with the same clusters.

5. Imagine k-means clustering were given the following set of data:

	x_1	x_2	x_3
1	1	1	1
2	0	0	0
3	0	1	1
4	1	1	0
5	0	0	1
6	0	1	0
7	1	0	0
8	0	0	1

If k=2 and where the initial starting points are the first two examples (1,1,1) and (0,0,0). Which points are in each cluster when it halts? How many iterations does it take?

Iteration	c_1	c_2	Which cluster this point is assigned to							
			1	2	3	4	5	6	7	8
0	(1,1,1)	(0,0,0)	1	2	1	1	2	2	2	2
1	(2/3,1,2/3)	(1/5,1/5,2/5)	1	2	1	1	2	2	2	2
Since the clusters didn't change, we stop										

One cluster has points 1, 3, and 4, and the other has points 2, 5, 6, 7, and 8. It took just one iteration to generate it.