Take Test: Quiz 6

Test Information

Description	Due Monday, April 10, 11:30pm.
Instructions	
Multiple Attempts	This test allows 3 attempts. This is attempt number 2.
Force Completion	This test can be saved and resumed later.

Question Completion Status:

1 2 3 4 5 6 7 8 9 10 11 12 13

QUESTION 1

2 points Saved

For which of the following tasks might clustering be a suitable approach?

- [ ] A. Given historical weather records, predict if tomorrow's weather will be sunny or rainy
- [ ] B. Given sales data from a large number of products in a supermarket, estimate future sales for each of these products
- [x] C. Given a database of information about your users, automatically group them into different market segments
- [x] D. From the user's usage patterns on a website, identify different user groups

QUESTION 2

3 points Saved

Which of the following statements about clustering are correct?

- [ ] 1. In order to perform cluster analysis, we need to have a similarity measure between data objects.
- [ ] 2. We must know the number of clusters a priori for all clustering algorithms.
- [ ] 3. The choice of an appropriate metric will influence the shape of the clusters.
- [ ] 4. When clustering, we want to put two dissimilar data objects into the same cluster.
- [ ] 5. Graphs, time-series data, text, and multimedia data are all examples of data types on which cluster analysis can be performed.
- [x] 6. We need to be able to handle a mixture of different types of attributes (e.g., numerical, categorical).
- [x] 7. Clustering analysis is unsupervised learning since it does not require labeled training data.

QUESTION 3

2 points Saved

Which of the following statements about the K-means algorithm are correct?

- [x] 1. The K-means algorithm can converge to different final clustering results, depending on initial choice of representatives.
- [x] 2. A standard way of initilizaing K-means is to set all the centroids, mu1 ro muk, to be a vector of zeros.
- [ ] 3. K-means will always give the same clustering result regardless of the initialization of the centroids.
- [ ] 4. The centroids in the K-means algorithm may not be any observed data points.
- [x] 5. To avoid K-means getting stuck at a bad local optima, we should try using multiple randon initialization.

QUESTION 4

2 points Saved

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

- [x] A. Assign each point to its nearest cluster
- [x] B. Update the cluster centroids based the current assignment
- [ ] C. Using the elbow method to choose K
- [ ] D. Test on the cross-validation set

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Save All An

Suppose you run K-means with 10 different random initizlizations and obtain 10 different clusterings of the data. Which is the recommended way for choosing which one of the 10 clusterings to use?

- [ ] Always pick the last (10th) clustering result, since by that time the algorithm is more likely to have converged to a good solution
- [x] Evaluate the objective function at the 10 clustering results and pick the one that gives rise to the smallest value of the objective function
- [ ] We have to obtain labels y_i for each observation
- [ ] Average the 10 sets of centroids and then reassign x_i to its nearest centroids

QUESTION 6

2 points Saved

Which of the following is required by K-means or K-medoids clustering?

- [x] initial guess of the cluster centroids
- [ ] the true label of the n data points
- [x] number of clusters
- [x] defined distance metric

QUESTION 7

1 points Saved

Which of the following is finally produced by Hierarchical Clustering?

- ( ) final estimate of cluster centroids
- (•) a dendrogram showing how close things are to each other
- ( ) assignment of each point to clusters
- ( ) All of the above mentioned

QUESTION 8

6 points Saved

In this problem, you will perform K-means clustering (using Euclidean distance) manually, with K = 2, on a small example with n = 6 observations and p = 2 features.

The observations are as follows.

(1,4), (1,3), (0,4), (5,1), (6,2), (4,0)

Set (1,4) as the centroid for cluster 1 and (1,3) as the centroid for cluster 2. Then

- (a) assign each observation to the nearest cluster.
  How many points will be assigned to cluster 1? Ans: 2
  How many points will be assigned to cluster 2? Ans: 4
- (b) update the centroids for the two clsuters.
  What's the x-coordinate of the new centroid for cluster 1? 0.5
  What's the x-coordinate of the new centroid for cluster 2? 4

Repeat (a) and (b) until convergence. After the algorithm converges,

- the x-coordinate of the cluster centroid to which (4,0) belongs is equal to 5, and
- the size of the cluster to which (4,0) belongs is 3, i.e., number of points in that cluster including (4,0).

QUESTION 9

2 points Saved

Suppose that we have four observations, whose pairwise dissimilarities are given below

- d(1,2) = 0.3
- d(1,3) = 0.4
- d(1,4) = 0.7
- d(2,3) = 0.5
- d(2,4) = 0.8
- d(3,4) = 0.45

For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

Apply hierarchical clustering on these four observations using **single linkage**. Suppose that we cut the dendrogram obtained from the hierarchical clustering such that two clusters result. Which observations are in each cluster? We use expressions like (1,2) to indicate the first and the second observations are in one cluster.

- ( ) (1,2) and (3,4)

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Save All An

(1,2,3) and (4)

QUESTION 10

2 points Saved

Continue with the previous question.

Now apply hierarchical clustering on these four observations using **complete linkage**. Suppose that we cut the dendrogram obtained from the hierarchical clustering such that two clusters result. Which observations are in each cluster?

- (1,2) and (3,4)
- (1,2,3) and (4)
- (1,3) and (2,4)
- (1) and (2,3, 4)

QUESTION 11

12 points Saved

Consider the Dishonest Casino example from the Rcode page for HMM: a dishonest casino uses two dice, one of them is fair and the other is loaded.

- The probabilities of the fair die are (1/6, ..., 1/6).
- The probabilities of the loaded die are (1/10, ..., 1/10,1/2)

Let X_t denote the observed outcome at time t, which takes values from 1 to 6. Let Z_t=1 denote that a fair die is used at time t, and Z_t=2 denote that a loaded ide is used at time t.

We model Z_t's by a Markov chain. Assume the initial distribution for Z_1 is (0.5, 0.5), and P(Z_{t+1} = k | Z_t = k) = 0.6 for k=1, 2.

Compute the following probabilities. Express your answers as fractions **reduced to the lowest terms.**

- P(Z_1 = 1, Z_2 = 1, Z_3=1) = 27 / 1000 (a1, an integer, is the numerator, and a2, an integer, is the denominator)
- P(X_1 = 1) = 2 / 15
- P(X_2 = 1) = 2 / 15
- P(Z_1 = 2 | X_1 = 1, X2 = 1) = 19 / 54
- P(Z_2 = 2 | X_1 = 1, X2 = 1) = 19 / 54
- P(Z_1 = 2, Z_2 = 2| X_1 = 1, X2 = 1) = 1 / 6

QUESTION 12

2 points Saved

Let Z be a random variable taking three different values, {1,2,3}, with probabilities w1, w2, and w3. Conditioning on Z=k, X follows a Poisson distribution with parameter lambda_k. Which of the following equations must always be true (where d is any non-negative integer)?

- [x] A. $P(Z = k \mid X = d) \geq P(X = d \mid Z = k)\, P(Z = k),\ k = 1, 2, 3.$
- [x] B. $\sum_{k=1}^{3} P(Z = k \mid X = d) = 1$
- [ ] C. $\sum_{k=1}^{3} P(X = d \mid Z = k)\, P(Z = k) = 1$
- [ ] D. $\sum_{k=1}^{3} P(X = d \mid Z = k) = 1$

QUESTION 13

8 points Saved

Suppose we are fitting a one-dimensional Gaussian mixture model on the following five observations: **5, 15, 25, 30, 40**

Use K=2 components. The parameters are the mixing proportions for the two components, **w1** and **w2**, and the means for the two Normal components, **mu1** and **mu2**. The standard deviations for the two components are fixed at 10.

Suppose at some point in the EM algorithm, the E-step found that the probabilities of r_{ik} = P(Z_i=k | X_i = x) for the five data points were given as follows:

- r_{11} = 0.2, r_{12}=0.8;
- r_{21} = 0.2, r_{22}=0.8;

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Save All An

Then at the M-step, the new estimates for (w1, w2, mu1, mu2) should be:

- w1 = 0.6, w2=0.4, (round to the 1st decimal point; express your answer as "**0.2**" instead of "**.2**".)
- mu1=29, mu2=14 (round to the nearest integer)

Click Save and Submit to save and submit. Click Save All Answers to save all answers.

Save All An