1.

Question 1

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

1 point



Given historical weather records, predict if tomorrow's weather will be sunny or rainy. (Not a grouping)



Given sales data from a large number of products in a supermarket, estimate future sales for each of these products. (No, Unless future sales are some groups of sales)



From the user's usage patterns on a website, identify different user groups. (**Groups being the keyword here)**



Given a database of information about your users, automatically group them into different market segments. **(Groups being the keyword here)**

2.

Question 2 (Partially Correct) **0.714 / 1 point**

Which **two** of the following statements about clustering are **not correct**?

1 point



Clustering analysis is unsupervised learning since it does not require labeled training data.



In order to perform cluster analysis, we need to have a similarity measure between data objects.



~~Graphs, time-series data, text, and multimedia data are all examples of data types on which cluster analysis can be performed~~. (This one was selected but I think its below two)



We must know the number of clusters a priori for all clustering algorithms. (Only for k-means and not for hierarchical)



When clustering, we want to put two dissimilar data objects into the same cluster.



The choice of an appropriate metric will influence the shape of the clusters.



We need to be able to handle a mixture of different types of attributes (e.g., numerical, categorical).

3.

Question 3

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

1 point



To avoid K-means getting stuck at a bad local optima, we should try using multiple randon initialization. **(This is how the algorithm works anyhow, by repeatedly getting different local optima on multiple iterations, and till a smallest value is repeatedly received)**



K-means will always give the same clustering result regardless of the initialization of the centroids. **(No this is just the opposite)**



The K-means algorithm can converge to different final clustering results, depending on initial choice of representatives. **(Page 388 ISLR)**



The centroids in the K-means algorithm may not be any observed data points. **(this could be because a mean is chosen normally and not a specific point but below q8 to 12 we did chose an observed data point)**



A standard way of initializing K-means is to set all the centroids, *μ*1​ to *μk*​, to be a vector of zeros. (**No the Vector is a Vector of p feature means for the observation in the kth cluster - NORMALLY)**

4.

Question 4

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

1 point



Update the cluster centroids based the current assignment **(Page 388 ISLR and also can be observed the same doing q8 to 12 in code)**



Assign each point to its nearest cluster **(Page 388 ISLR and also can be observed the same doing q8 to 12 in code)**



Using the elbow method to choose K



Test on the cross-validation set

5.

Question 5

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

1 point



Always pick the last (10th) clustering result, since by that time the algorithm is more likely to have converged to a good solution



Evaluate the objective function at the 10 clustering results and pick the one that gives rise to the smallest value of the objective function



Average the 10 sets of centroids and then reassign x\_i*xi*​ to its nearest centroids



We have to obtain labels *yi*​ for each observation

6.

Question 6

Which of the following is finally produced by hierarchical clustering?

1 point



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

7.

Question 7

**Question 7- Question 12 are related. (Follow Lecture Notes – much clear on steps)**

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. (<https://rpubs.com/ppaquay/65568> - Q3)

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? **[n1]**

How many points will be assigned to cluster 2? **[n2]**

* (b) update the centroids for the two clsuters.

What's the x-coordinate of the new centroid for cluster 1? **[a1]**

What's the x-coordinate of the new centroid for cluster 2? **[a2]**

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 **[c1]**, and
* the size of the cluster to which (4,0) belongs is **[c2]**, i.e., number of points in that cluster including (4,0).

**[n1]**=\_\_2\_\_\_\_\_\_\_\_\_\_\_\_\_

1 point



8.

Question 8

**[n2]**=\_\_4\_\_\_\_\_\_\_\_\_\_\_\_\_

1 point



9.

Question 9

**[a1]**=\_0.5\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1 point



10.

Question 10

**[a2]**=\_\_4\_\_\_\_\_\_\_\_\_\_\_\_\_

1 point



11.

Question 11

**[c1]**=\_\_\_\_\_\_5\_\_\_\_\_\_\_\_\_

1 point



12.

Question 12

**[c2]**=\_\_3\_\_\_\_\_\_\_\_\_\_\_\_\_

1 point



13.

Question 13 (**https://rpubs.com/ppaquay/65568**)

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 point



(1,3) and (2,4)



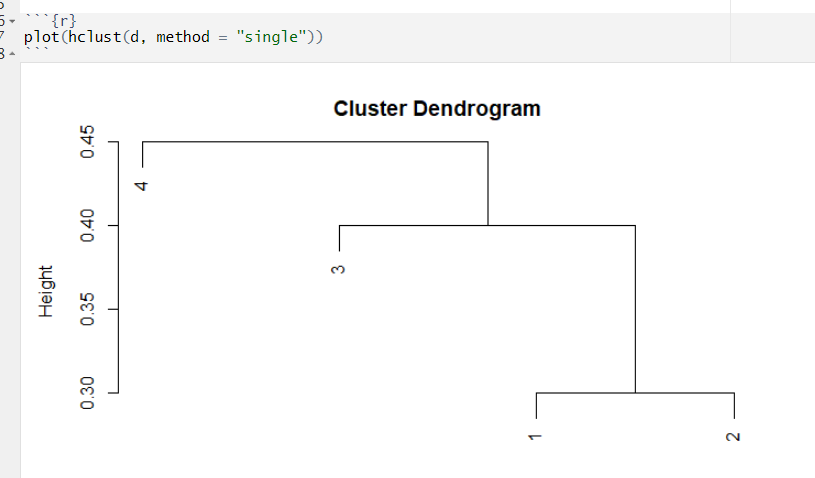
(1,2,3) and (4)



(1) and (2,3, 4)



(1,2) and (3,4)



14.

Question 14

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 point



(1) and (2,3, 4)



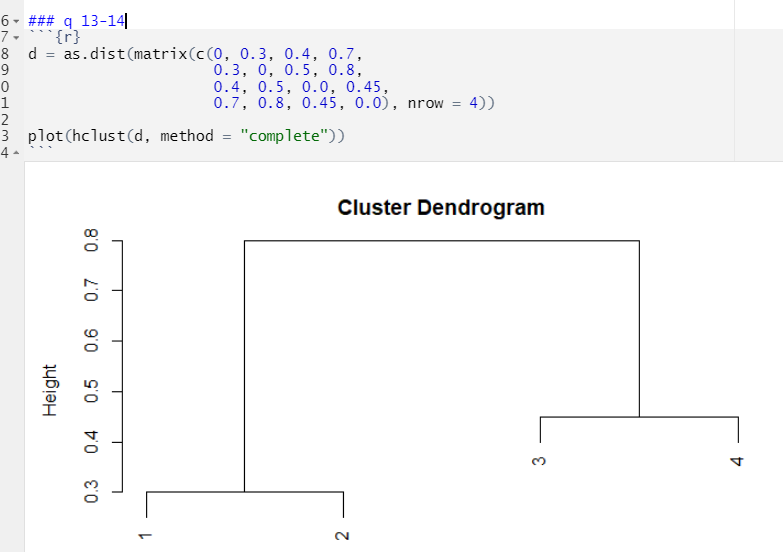
(1,2,3) and (4)



(1,3) and (2,4)



(1,2) and (3,4)



15.

Question 15

**Question 15-Question 26 are related.**

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 *Xt*​ denote the observed outcome at time t, which takes values from 1 to 6. Let Z\_t=1*Zt*​=1 denote that a fair die is used at time t*t*, and Z\_t=2*Zt*​=2 denote that a loaded die is used at time t*t*.

We model Z\_t*Zt*​'s by a Markov chain. Assume the initial distribution for Z\_1*Z*1​ is (0.5, 0.5), and P(Z\_{t+1} = k | Z\_t = k) = 0.6*P*(*Zt*+1​=*k*∣*Zt*​=*k*)=0.6 for k=1, 2*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) = [a1]/[a2]*P*(*Z*1​=1,*Z*2​=1,*Z*3​=1)=[*a*1]/[*a*2] (a1, an integer, is the numerator, and a2, an integer, is the denominator)
* P(X\_1 = 1) = [b1]/[b2]*P*(*X*1​=1)=[*b*1]/[*b*2]
* P(X\_2 = 1) = [c1]/[c2]*P*(*X*2​=1)=[*c*1]/[*c*2]
* P(Z\_1 = 2 | X\_1 = 1, X\_2 = 1) = [d1]/[d2]*P*(*Z*1​=2∣*X*1​=1,*X*2​=1)=[*d*1]/[*d*2]
* P(Z\_2 = 2 | X\_1 = 1, X\_2 = 1) = [e1]/[e2]*P*(*Z*2​=2∣*X*1​=1,*X*2​=1)=[*e*1]/[*e*2]
* P(Z\_1 = 2, Z\_2 = 2| X\_1 = 1, X\_2 = 1) = [f1]/[f2]*P*(*Z*1​=2,*Z*2​=2∣*X*1​=1,*X*2​=1)=[*f*1]/[*f*2]

a1*a*1=\_\_\_\_18\_\_\_\_\_\_\_\_

1 point



16.

Question 16

a2*a*2=\_\_\_\_100\_\_\_\_\_\_\_\_

1 point



17.

Question 17

b1*b*1=\_\_\_\_\_2\_\_\_\_\_\_\_

1 point



18.

Question 18

b2*b*2=\_\_\_\_\_15\_\_\_\_\_\_

1 point



19.

Question 19

c1*c*1=\_\_\_\_2\_\_\_\_\_\_\_\_

1 point



20.

Question 20

c2*c*2=\_\_\_\_15\_\_\_\_\_\_\_\_

1 point



21.

Question 21

d1*d*1=\_\_\_2\_\_\_\_\_\_\_\_\_

1 point



22.

Question 22

d2*d*2=\_\_15\_\_\_\_\_\_\_\_\_\_

1 point



23.

Question 23

e1*e*1=\_\_2\_\_\_\_\_\_\_\_\_\_

1 point



24.

Question 24

e2*e*2=\_\_\_15\_\_\_\_\_\_\_\_\_

1 point



25.

Question 25

f1*f*1=\_\_300\_\_\_\_\_\_\_\_\_\_

1 point



26.

Question 26

f2*f*2=\_\_\_648\_\_\_\_\_\_\_\_\_

1 point



27.

Question 27

**Questions 27-30 are related. (**[**https://www.youtube.com/watch?v=7e65vXZEv5Q**](https://www.youtube.com/watch?v=7e65vXZEv5Q)**, Solution:** [**http://www.utstat.utoronto.ca/~radford/sta414.S13/practice2-ans.pdf**](http://www.utstat.utoronto.ca/~radford/sta414.S13/practice2-ans.pdf) **)**

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**, *μ*1​ and *μ*2**​. 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 *rik*​=*P*(*Zi*​=*k*∣*Xi*​=*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;
* *r*31​=0.8,*r*32​=0.2;
* *r*41​=0.9,*r*42​=0.1;
* *r*51​=0.9,*r*52​=0.1.

Then at the M-step, the new estimates for (w1, w2, ,*μ*1,*μ*2) should be:

* *w*1=[*w*1],*w*2=[*w*2], (round to the 1st decimal point; express your answer as "**0.2**" instead of "**.2**".
* *μ*1​=[*μ*1​],*μ*2​=[*μ*2​] (round to the nearest integer)

*w*1= (0.2 + 0.2 + 0.8 + 0.9 + 0.9)/5 = 0.6



28.

Question 28

*w*2=\_(0.8 + 0.8 + 0.2 + 0.1 + 0.1)/5 = 0.4

1 point



29.

Question 29

*μ*1​=\_(0.2 × 5 + 0.2 × 15 + 0.8 × 25 + 0.9 × 30 + 0.9 × 40) / (0.2 + 0.2 + 0.8 + 0.9 + 0.9) = 29\_\_\_\_\_\_\_\_

1 point



30.

Question 30

*μ*2​=\_(0.8 × 5 + 0.8 × 15 + 0.2 × 25 + 0.1 × 30 + 0.1 × 40) / (0.8 + 0.8 + 0.2 + 0.1 + 0.1) = 14\_\_\_\_\_\_\_\_

1 point



31.

Question 31 (Partially Correct) – (0.5 / 1)

Which **two** of the following statements about Gaussian mixture models are correct?

1 point



Let K denote the number of components in a Gaussian mixture model. Increasing K will always lead to a strictly larger log-likelihood on the training data. (Lec: Gaussian Mixtures)



The estimates returned by the EM algorithm could be a local optimum. (**convergence happens for local optima only. Step 2 and 3 of EM is iterative till convergenc**



We can use BIC to select the number of components K. **(as per week 7, Lec 1, K can be chosen by AIC or BIC, so this seems to be right)**



The log-likelihood of the Gaussian mixture model is convex, therefore all local optima are global optima. (**Log-likelihood keeps it in Concave)**

32.

Question 32 (Partially Correct) – (0.5 / 1)

Which **two** of the following statements about the Latent Dirichlet Allocation model are correct? Assume there are K topics.

1 point



We can represent the i*i*-th document by a *K*-dim vector: (*ai*1​,...,*aiK*​) which indicates that the words in the *i*-th document can be modeled as a mixture of K topics with *aik*​'s being the mixing weights.



All the words in a document are generated by the same multinomial distribution once the topic label for the document is given.



Words in the same document can be generated from different multinomial distributions depending on the topic label for each word.



We can represent the *i*-th document by a K-dim vector: (*ai*1​,...,*aiK*​), with *aik*​ being the probability of the *i*-th document belonging to k*k*-th topic.