Feedback — Week 1 Quiz

Thank you. Your submission for this quiz was received.

You submitted this quiz on **Fri 1 May 2015 11:52 PM PDT**. You got a score of **11.00** out of **13.00**.

Question 1

Which of the following statements are true?

Your Answer		Score	Explanation
☑ When clustering, we want to put two data objects that are similar into the same cluster.	✔	0.25	
☑ Clustering analysis has a wide range of applications in tasks such as data summarization, dynamic trend detection, multimedia analysis, and biological network analysis.	✔	0.25	
☐ Cluster analysis is considered supervised learning.	✔	0.25	This is false because cluster analysis is by definition unsupervised learning.
☐ It is impossible to cluster objects in a data stream. We must have all the data objects that we need to cluster ready before clustering can be performed.	✔	0.25	This is false because clustering algorithms can be adapted to perform clustering in a streaming fashion.
Total		1.00 / 1.00	

Question Explanation

The correct answers are: "When clustering, we want to put two data objects that are similar into the same cluster." and "Clustering analysis has a wide range of applications in tasks such as data summarization, dynamic trend detection, multimedia analysis, and biological network analysis."

Question 2

What are some common considerations and requirements for cluster analysis?

Your Answer		Score	Explanation
☑ We need to be able to handle a mixture of different types of attributes (e.g., numerical, categorical).	✔	0.25	
☐ We must know the number of output clusters a priori for all clustering algorithms.	✔	0.25	This is false because there are hierarchical partitioning methods that do not require the number of clusters as an input parameter.
☑ We need to consider how to incorporate user preference for cluster size and shape into the clustering algorithm.	✔	0.25	
☑ In order to perform cluster analysis, we need to have a similarity measure between data objects.	✔	0.25	
Total		1.00 / 1.00	

Question Explanation

The correct answers are: "We need to be able to handle a mixture of different types of attributes (e.g., numerical, categorical).", "We need to consider how to incorporate user preference for cluster size and shape into the clustering algorithm.", and "In order to perform cluster analysis, we need to have a similarity measure between data objects."

Question 3

Which of the following statements are true?

Your Answer	Score	Explanation
☑ Dimensionality reduction helps make high-dimensional clustering more feasible and scalable.	✔ 0.25	
☐ There are no clustering algorithms that can handle time-series data since we always assume that the data points are	✔ 0.25	

temporally independent from each other.

☑ K-means is an example of a distance-based clustering method.	✔	0.25
☐ Since cluster analysis is unsupervised learning, there's no way to incorporate user preference or guidance into the clustering process.	✔	0.25
Total		1.00 / 1.00

Question Explanation

The correct answers are: "K-means is an example of a distance-based clustering method." and "Dimensionality reduction helps make high-dimensional clustering more feasible and scalable." We will provide further discussion on how to perform cluster analysis on time series data and how to incorporate user feedback in later lectures.

Question 4

The following real dataset contains information about Iris setosa and versicolor.

What is the Euclidean distance between these two objects?

Species	Sepal length	Sepal width	Petal length	Petal width
Iris setosa	4.9	3.0	1.4	0.2
Iris versicolor	5.6	2.5	3.9	1.1

Your Answer		Score	Explanation
○ 7.8			
○ 4.6			
○ 2.5			
⊙ 2.8	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

The Euclidean distance, corresponding to p = 2 Minkowski distance, between two objects is defined as follows:
$$d2(i,j) = \sqrt{{}^{l}\Sigma_{k=1} (x_{i,k} - x_{j,k})^2}$$

Question 5

The following real dataset contains information about Iris setosa and versicolor.

Cases	F1	F2	F3	F4	F5	F6	F7	F8	F9	F10
1	0	1	1	0	0	0	1	1	1	1
2	0	1	0	0	1	0	1	0	0	1

Assume all the activities are asymmetric binary variables. What is the distance between Case 1 and Case 2?

Your Answer	Score	Explanation
○ 4/10		
○ 3/10		
◉ 3/7	✖ 0.00	
○ 4/7		
Total	0.00 / 1.00	

Question Explanation

If the binary variables are asymmetric, we have

d(i,j) = r + s / q + r + s = 4/7

Question 6

The following real world dataset contains two samples from Car Evaluation Database, which was derived from a simple hierarchical decision model originally developed for the demonstration of DEX (M. Bohanec, V. Rajkovic: Expert System for Decision Making. Sistemica 1(1), pp. 145-157, 1990.). The model evaluates cars according to the following concept structure:

CAR   car acceptability
. PRICE   overall price
. . buying   buying price
. . maint   price of the maintenance
. TECH   technical characteristics

. . COMFORT   comfort

. . . doors   number of doors

. . . persons   capacity in terms of persons to carry

. . . lug_boot   the size of luggage boot

. . safety   estimated safety of the car

The attribute values are as follows:

Attrubute	Values (categorical)
buying	v-high, high, med, low
maint	v-high, high, med, low
doors	2, 3, 4, 5-more
persons	2, 4, more
lug_boot	small, med, big
safety	low, med, high

Case	buying	maint	doors	persons	lug_boot	safety
Car 1	med	v-high	3	more	small	med
Car 2	high	v-high	4	4	big	med

Assume we use simple match to estimate the proximity between objects. What is the distance between Car 1 and Car 2?

Your Answer		Score	Explanation
◉ 1/3	✖	0.00	
○ 8/10			
○ 2/3			
○ 8/21			
○ 8/17			
Total		0.00 / 1.00	

Question Explanation

Considering simple match, there are 2 out of 6 features between the two cars that are the same. Thus, the distance is:

d(i,j) = m-p/m = 6-2/6 = 2/3

Question 7

Given the following two short texts with punctuation removed, calculate the cosine similarity between them based on the bag of words model.

Text1: language is the source of misunderstandings

Text2: language is the soul of a nation

Your Answer		Score	Explanation
○ 0.095			
○ 0.44			
○ 0			
◉ 0.617	✔	1.00	
Total		1.00 / 1.00	

Question Explanation

We can get the vector representations for the two short texts,

T1 = (1, 1, 1, 1, 1, 1, 0, 0, 0)

T2 = (1, 1, 1, 0, 1, 0, 1, 1, 1),

where the dimensions correspond to: language, is, the, source, of, misunderstanding, soul, a, nation.

The cosine similarity can be obtained via

cos(T1,T2) = T1*T2/ ||T1|| ||T2|| = 4/ $\sqrt{6}$ $\sqrt{7}$

Question 8

With regard to the species of Iris setosa, we have sampled data on the features of sepal length and sepal width, as follows:

Feature	Sepal length	Sepal width
Case 1	5.1	3.5
Case 2	4.9	3.0
Case 3	4.7	3.2

| Case 4 | 4.6 | 3.1 |
| Case 5 | 5.0 | 5.4 |

What is the sample correlation coefficient between sepal length and sepal width?

Your Answer	Score	Explanation
○ 0.185		
○ 2.396		
● 0.479	✔	1.00
○ 0.895		
○ 0.398		
Total	1.00 / 1.00	

Question Explanation

The sample correlation coefficient can be calculated as:

$$\rho_{12} = \sigma_{12}/\sigma_1\,\sigma_2 = {}^n\Sigma_{i=I}\,(x_{iI} - \mu_I)(x_{i2} - \mu) / \sqrt{}\ {}^n\Sigma_{i=I}\,(x - \mu_2)^2$$

Question 9

Considering the K-means algorithm, after current iteration, we have 3 centroids (0, 1) (2, 1), (-1, 2). Will points (2, 3) and (-0.5, 0) be assigned to the same cluster in the next iteration?

Your Answer	Score	Explanation
○ Yes		
● No	✔	1.00
Total	1.00 / 1.00	

Question Explanation

(-0.5, 0) will be assigned to (0, 1) while (2, 3) will be assigned to (2, 1).

Question 10

Considering the K-means algorithm, if points (-1, 3), (-3, 1), and (-2, -1) are the only points which are assigned to the first cluster now, what is the new centroid for this cluster?

Your Answer		Score	Explanation
○ (0, 3)			
◉ (-2, 1)	✔	1.00	
○ (0, 2)			
○ (0, 0)			
Total		1.00 / 1.00	

Question Explanation

Calculate the average value for x and y separately. You will then find the answers are -2 and 1, and thus, the new centroid should be (-2, 1).

Question 11

K-means++ algorithm is designed for better initialization for K-means, which will take the farthest point from the currently selected centroids. Suppose k = 2 and we have selected the first centroid is (0, 0). Among the following points (these are all the remaining points), which one should we take for the second centroid? (Here, the distance is measured by Euclidean Distance).

Your Answer		Score	Explanation
◉ (4, 0)	✔	1.00	
○ (0, 3)			
○ (2, 0)			
○ (-2, 1)			
Total		1.00 / 1.00	

Question Explanation

K-means++ will take the farthest point from the currently selected centroids.

Question 12

Considering the K-median algorithm, if points (1, -3), (1, 1), and (-2, 2) are the only points which are assigned to the first cluster now, what is the new centroid for this cluster?

Your Answer		Score	Explanation
◉ (1, 1)	✔	1.00	
○ (0, 3)			
○ (0, 2)			
○ (-2, 1)			
Total		1.00 / 1.00	

Question Explanation

Calculate the median value for x and y separately. You will then find the answers are 1 and 1, and thus the new centroid should be (1, 1)

Question 13

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

Your Answer	Score	Explanation
☐ For different initializations, the K-means algorithm will definitely give the same clustering results.	✔ 0.25	
☐ The K-means algorithm can directly handle non-numerical (categorical) data.	✔ 0.25	
☑ The centroids in the K-means algorithm may not be any observed data points.	✔ 0.25	
☑ The K-means algorithm is sensitive to outliers.	✔ 0.25	

Total   1.00 /
                                                                           1.00

Question Explanation

K-Means cannot handle non-numerical (categorical) data, while K-modes can. Different
initializations may generate rather different clustering results (some could be far from optimal).