The "Data Mining" Specialization Learn More ×

Feedback — Week 1 Quiz

Help Center

Thank you. Your submission for this quiz was received.

You submitted this quiz on **Sat 2 May 2015 12:15 AM PDT**. You got a score of **12.00** 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	
□ 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.
☑ 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	
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 consider the desired shape and size of clusters.	✓	0.25	
☑ In order to perform cluster analysis, we need to have a similarity measure between data objects.	✓	0.25	
☑ We need to consider the space on which the clustering is performed. In other words, we need to decide what subset of the available features we are going to consider in the similarity measure.	✓	0.25	
☐ Cluster analysis requires a large amount of labeled training data.	✓	0.25	This is false since cluster analysis is unsupervised learning, which by definition does not require training data.
Total		1.00 / 1.00	

Question Explanation

The correct answers are: "In order to perform cluster analysis, we need to have a similarity measure between data objects." , "We need to consider the desired shape and size of clusters." and "We need to consider the space on which the clustering is performed. In other words, we need to decide what subset of the available features we are going to consider in the similarity measure."

Question 3

Which of the following statements are true?

Your Answer		Score	Explanation
☑ 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
☐ There are no clustering algorithms that can handle time-series data since we always assume that the data points are temporally independent from each other.	✓	0.25
☑ Dimensionality reduction helps make high-dimensional clustering more feasible and scalable.	✓	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
● 2.8	✓	1.00	
○ 2.5			
○ 4.6			
○ 7.8			
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 Jaccard coefficient between Case 1 and Case 2?

Your Answer		Score	Explanation
○ 4/7			
○ 4/10			
● 3/7	✓	1.00	
○ 3/10			
Total		1.00 / 1.00	

Question Explanation

The Jaccard coefficient defined for asymmetric binary variables is as follows:

d (i,j) = q / q + r + s = 3/7

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

To calculate the distance between objects with categorical attributes, we use a set of binary attributes to represent each categorical attribute. Assume all the binary attributes are asymmetric. What is the distance between Car 1 and Car 2?

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

Question Explanation

Considering converting the categorical random variables into binary random variables, there will be (4 + 4 + 4 + 3 + 3 + 3 = 21) binary random variables in total. Moreover, we have the following contingency table:

	Car 1			
Car 2		1	0	sum
	1	2	4	6
	0	4	11	15
	sum	6	15	21

If all the binary attributes are asymmetric, we have the distance between Car 1 and Car 2 as:

d(i,j) = s+r/q+s+r = 8/10

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: all grown-ups were once children but only few of them remember it

Text2: all children should be very understanding of grown-ups

Your Answer		**Score**	**Explanation**
● 0.408	✓	1.00	
○ 0			
○ 0.22			
○ 0.042			
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, 1, 1, 1, 1,1, 1, 0, 0, 0, 0)

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

where the dimensions correspond to: all, grown-ups, were, once, children, but, only, few, of, them, remember, it, should, be, very, understanding.

The cosine similarity can be obtained via

cos(T1,T2) = T1*T2/ ||T1|| ||T2|| = 4/ √12 √ 8

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
○ 0.398			
○ 0.895			
● 0.479	✓	1.00	
○ 2.396			
○ 0.185			
Total		1.00 / 1.00	

Question Explanation

The sample correlation coefficient can be calculated as:

$\rho_{12} = \sigma_{12}/\sigma_1\ \sigma_2 = {}^{n}\Sigma_{i=l}\ (x_{il} - {}_{\mu l})(x_{i2} - \mu)\ /\ \sqrt{}\ {}^{n}\Sigma_{i=l}\ (x - \mu_2)^2$

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 (0, 3), (2, 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
○ (0, 0)		
● (0, 2) ✓	1.00	
○ (-2, 1)		
○ (0, 3)		
Total	1.00 / 1.00	

Question Explanation

Calculate the average value for x and y separately. You will then find the answers are 0 and 2, and thus, the new centroid should be (0, 2).

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
● (3, 0) ✓	1.00	

○ (2, 0)			
○ (-2, 1)			
○ (0, 2)			
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), (-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, 2)			
○ (0, 0)			
○ (0, 3)			
● (-2, 1)	✓	1.00	
Total		1.00 / 1.00	

Question Explanation

Calculate the median 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 13

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

Your Answer		Score	Explanation
☑ 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
□ For different initializations, the K-means algorithm will definitely give the same clustering results.	✓	0.25
□ The K-means algorithm can detect non-convex clusters.	✓	0.25
Total		1.00 / 1.00

Question Explanation

K-Means can only detect clusters that are linearly separable, while kernel K-means can detect some non-convex clusters. Different initializations may generate rather different clustering results (some could be far from optimal).