Question 1 (1 point)

Using the PCA algorithm as described in lectures, what does the

$$\alpha$$

parameter represent?

- ◯ the largest eigenvalue of the covariance matrix
- ◯ the minimum fraction of total variance to be preserved
- ◯ the minimum number of principal components to use
- ◯ the minimum number of new attributes to create

Question 2 (1 point)

What is the product SRx, where S, R and x are defined as below:

$$S = \begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix}$$

$$R = \begin{pmatrix} \cos\frac{\pi}{6} & -\sin\frac{\pi}{6} \\ \sin\frac{\pi}{6} & \cos\frac{\pi}{6} \end{pmatrix}$$

$$x = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

◯ $$\begin{pmatrix} \frac{3-3\sqrt{3}}{2} \\ 2+2\sqrt{3} \end{pmatrix}$$

◯ $$\begin{pmatrix} \frac{3+3\sqrt{3}}{2} \\ \sqrt{3}-1 \end{pmatrix}$$

◯ $$\begin{pmatrix} 2+2\sqrt{3} \\ \frac{3-3\sqrt{3}}{2} \end{pmatrix}$$

◯ $$\begin{pmatrix} \sqrt{3}-1 \\ \frac{3+3\sqrt{3}}{2} \end{pmatrix}$$

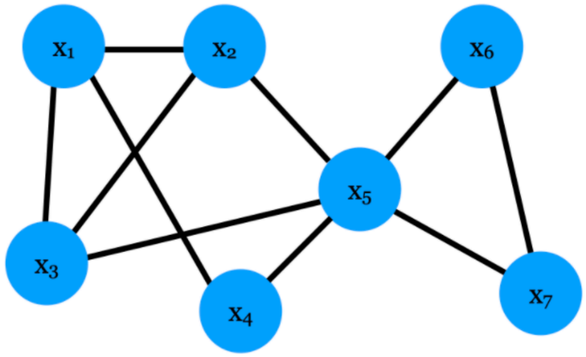
Question 3 (1 point)

Which of the following is a disadvantage of the PCA algorithm?

- ◯ The new attributes produced by PCA can be correlated with one another, making it difficult to determine which new attributes contribute most to the the observed variance in the data.
- ◯ PCA can reduce the dimensionality of a data set to two or three dimensions, but not 4 or more.
- ◯ PCA cannot project data onto nonlinear subspaces, and thus fails to capture nonlinear relationships in the attributes of a data set.
- ◯ PCA cannot be applied to data sets of very high dimensionality (e.g., 1000 or more attributes)

Question 5 (1 point)

Consider the following graph:



What is the clustering coefficient of node

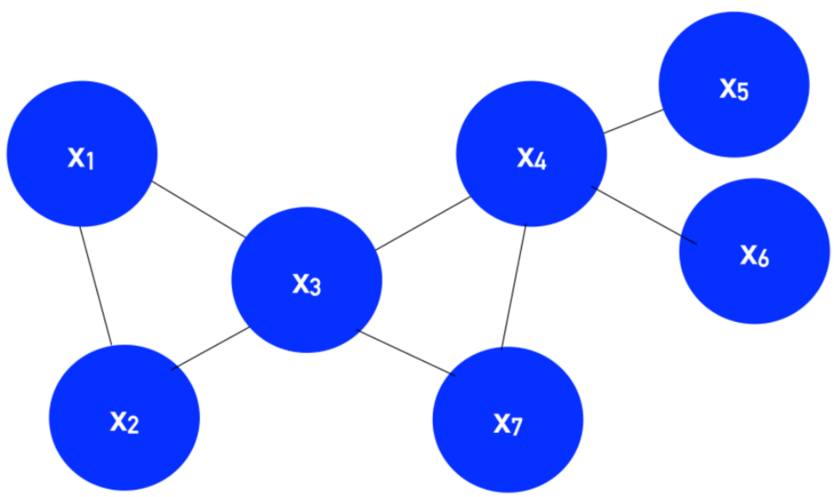
$$x_3$$

?

- 0
- $$\frac{5}{6}$$
- $$\frac{2}{3}$$
- 1

Question 6 (1 point)

Consider the following graph:



What is the closeness centrality of vertex

$$x_4$$

?

- ( ) $$\frac{1}{9}$$
- ( ) 8
- ( ) $$\frac{1}{8}$$
- ( ) 9

Question 7 (1 point)

Consider the following data matrix:

$$D = \begin{array}{cccccc} & X_1 & X_2 & X_3 & X_4 \\ x_1 & 0.2 & 23 & 5.7 & A \\ x_2 & 0.4 & 1 & 5.4 & C \\ x_3 & 1.8 & 0.5 & 5.2 & C \\ x_4 & 5.6 & 0.8 & 5.1 & A \\ x_5 & -0.5 & 34 & 5.3 & B \\ x_6 & 0.4 & 19 & 5.4 & C \\ x_7 & 1.1 & 11 & 5.5 & C \end{array}$$

What is the Euclidean distance between

$$x_3$$

and

$$x_4$$

after label-encoding attribute

$$X_4$$

with the labels:

$$A = 1, B = 2, C = 3$$

?

- 4.31
- 4.97
- 3.94
- 4.85

Question 8 (1 point)

Consider the following contingency table, showing the overlap between a ground-truth clustering with two clusters

$$T_1$$

and

$$T_2$$

and the clustering output of some clustering algorithm that produced three clusters,

$$C_1, C_2, C_3$$

:

	T_1	T_2
C_1	5	0
C_2	1	9
C_3	0	13

What is the precision of cluster

$$C_2$$

?

- 0.90
- 0.83
- 0.17
- 1.00

Question 9 (1 point)

Consider the two vectors

$$a$$

and

$$b$$

below:

$$a = \begin{pmatrix} 1 & -1 & -2 & 4 \end{pmatrix}$$

$$b = \begin{pmatrix} 2 & -1 & -1 & 3 \end{pmatrix}$$

What is the Euclidean distance between the two vectors (what is

$$||a - b||_2$$

?

- ○ $\sqrt{3}$
- ○ $\sqrt{2}$
- ○ 1
- ○ $\sqrt{7}$

Question 12 (1 point)

Consider the following contingency table, showing the overlap between a ground-truth clustering with two clusters

$$T_1$$

and

$$T_2$$

and the clustering output of some clustering algorithm that produced three clusters,

$$C_1, C_2, C_3$$

:

	T_1	T_2
C_1	5	0
C_2	1	9
C_3	0	13

What is the recall of cluster

$$C_2$$

?

- 0.17
- 0.59
- 0.83
- 0.41

Question 13 (1 point)

Consider the two vectors

$$a$$

and

$$b$$

below:

$$a = \begin{pmatrix} 1 & -1 & -2 & 4 \end{pmatrix}$$
$$b = \begin{pmatrix} 2 & -1 & -1 & 3 \end{pmatrix}$$

What is the dot product

$$a^T b$$

?

- ○ 11
- ○ 17
- ○ 22
- ○ 15

Question 14 (1 point)

The DBSCAN algorithm requires the number of clusters to discover as an input parameter.

- True
- False

Question 15 (1 point)

What is the volume of a sphere with radius 1 in 6 dimensions?

- 4.059
- 2.550
- 5.168
- 1.335

Question 16 (1 point)

Consider the following data matrix that we want to convert into graph data:

$$D = \begin{array}{cccc} & X_1 & X_2 & X_3 \\ x_1 & 0.2 & 23 & 5.7 \\ x_2 & 0.4 & 1 & 5.4 \\ x_3 & 1.8 & 0.5 & 5.2 \\ x_4 & 5.6 & 50 & 5.1 \\ x_5 & -0.5 & 34 & 5.3 \\ x_6 & 0.4 & 19 & 5.4 \\ x_7 & 1.1 & 11 & 5.5 \end{array}$$

Using the similarity function

$$sim(x_i, x_j) = e^{\frac{-||x_i - x_j||^2}{2\sigma^2}}$$

,

What would be the similarity between

$$x_1$$

and

$$x_5$$

after standard-normalizing the data matrix, when

$$\sigma = 1$$

?

- 0.1
- 0.0
- 0.6
- 0.98

Question 17 (1 point)

Consider the data matrix below, with instances in rows and attributes in columns.

$$D = \begin{array}{ccc} & X_1 & X_2 \\ x_1 & 0.2 & 2 \\ x_2 & 0.3 & 4 \\ x_3 & 0.5 & -1 \\ x_3 & 0.7 & 6 \end{array}$$

What is the sample covariance between

$$X_1 \text{ and } X_2$$

?

- ○ 8.92
- ○ 0.05
- ○ 3.33
- ○ 0.21

Question 18 (1 point)

Consider the data matrix below, with instances in rows and attributes in columns. What is the mean of the data?

$$
D = \begin{array}{c|cc}
 & X_1 & X_2 \\
x_1 & 0.2 & 2 \\
x_2 & 0.3 & 4 \\
x_3 & 0.5 & -1
\end{array}
$$

- ○ $\begin{pmatrix} 0.2 & 0.3 & 0.5 \end{pmatrix}$
- ○ $\begin{pmatrix} 2.18 & 0.66 \end{pmatrix}$
- ○ $\begin{pmatrix} 0.33 & 1.67 \end{pmatrix}$
- ○ $\begin{pmatrix} 1.1 & 2.15 & -0.25 \end{pmatrix}$

Question 19 (1 point)

Which of the following are valid reasons for reducing the dimensionality of a data set?

- ☐ Visualizing the data (in two or three dimensions)
- ☐ Eliminating noise in the data (by focusing on important attributes)
- ☐ Improving computational efficiency of algorithms applied to the data (by requiring less operations for distance or similarity computations)
- ☐ All of the above
- ☐ None of the above

Question 20 (1 point)

Let D be a data matrix. Let Z be the matrix that represents the mean-centered D.

True or false: Using the PCA algorithm as described in lectures, if the matrix D is passed as input to the PCA algorithm, the output will differ from the output produced when using Z as the input in place of D (keeping the

$$\alpha$$

parameter set to the same value).

- ◯ True
- ◯ False

Question 21 (1 point)

Suppose we have the following data matrix, and wish to find 2 clusters in the data using the k-means algorithm.

$$D = \begin{pmatrix} & X_1 & X_2 \\ x_1 & 5 & 6 \\ x_2 & 4.9 & 5.1 \\ x_3 & -2 & 2 \\ x_4 & -3 & 1 \\ x_5 & 4.5 & 4 \\ x_6 & 4 & 4.5 \\ x_7 & -1.1 & 1.8 \\ x_8 & -1 & 0.7 \\ x_9 & 5.3 & 4.2 \\ x_{10} & -2 & 0.9 \\ x_{11} & 5.7 & 3.8 \end{pmatrix}$$

Suppose also that our initial means are set to

$$\mu_1 = (3.9, 4)$$

and

$$\mu_2 = (6.2, 6)$$

.

After the first pass through the cluster assignment step in k-means, which set of points will constitute cluster 2?

- $\{x_1\}$
- $\{x_1, x_2\}$
- $\{x_1, x_2, x_9\}$
- $\{x_1, x_2, x_9, x_{11}\}$

Question 22 (1 point)

Consider the following data matrix:

$$D = \begin{array}{ccc} & X_1 & X_2 \\ x_1 & A & H \\ x_2 & B & L \\ x_3 & C & L \\ x_4 & A & L \\ x_5 & B & H \\ x_6 & C & L \\ x_7 & C & H \end{array}$$

What is the Hamming distance between

$$x_6$$

and

$$x_7$$

(assume one-hot encoding is reasonable to use; that is, the data is categorical, and not ordinal).

- 1
- 2
- 0
- -1

Question 23 (1 point)

Consider the following data matrix, where dashes (-) indicate missing values:

$$D = \begin{array}{c|ccc} & X_1 & X_2 & X_3 \\ \hline x_1 & - & - & 5.7 \\ x_2 & 0.4 & 1 & - \\ x_3 & 1.8 & - & 5.2 \\ x_4 & - & 50 & 5.1 \\ x_5 & - & 34 & - \\ x_6 & 0.4 & - & 5.4 \\ x_7 & 1.1 & 11 & - \end{array}$$

If we use forward fill to fill in missing entries, what would be the vector representing the data instance

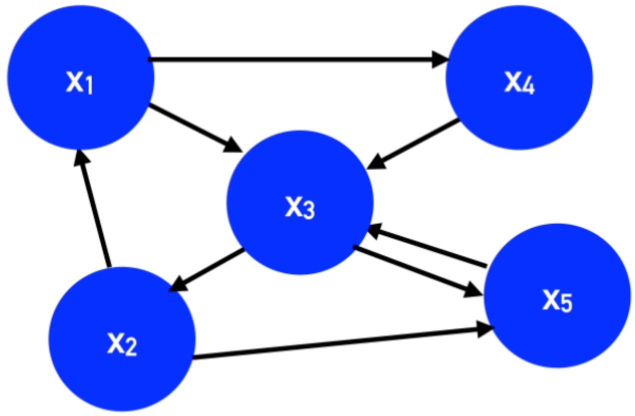
$$x_5$$

?

- $\begin{pmatrix} 0.4 & 34 & 5.4 \end{pmatrix}$
- $\begin{pmatrix} 1.8 & 34 & 5.1 \end{pmatrix}$
- $\begin{pmatrix} 1.8 & 50 & 5.1 \end{pmatrix}$
- $\begin{pmatrix} 0.4 & 11 & 5.4 \end{pmatrix}$

Question 24 (1 point)

Consider the following graph:



Suppose we wish to find the prestige (eigenvector centrality) of each node in the network, and we are using Power Iteration. If we set the initial prestige vector

$$p_0$$

to be a vector of all 1's, what is the prestige vector going to be after the third iteration (what is

$$\frac{p_3}{\max(p_3)}$$

going to be)?

- $\begin{pmatrix} 1 \\ 1 \\ 3 \\ 1 \\ 2 \end{pmatrix}$
- $\begin{pmatrix} 0.33 \\ 0.33 \\ 1.00 \\ 0.33 \\ 0.67 \end{pmatrix}$
- $\begin{pmatrix} 0.43 \\ 0.57 \\ 0.86 \\ 0.14 \\ 1.00 \end{pmatrix}$
- $\begin{pmatrix} 0.25 \\ 0.75 \\ 1.00 \\ 0.25 \\ 0.75 \end{pmatrix}$

Question 25 (1 point)

Consider the following data set D:

$$D = \begin{array}{cc} & X1 \\ x1 & 4 \\ x2 & 1.1 \\ x3 & 12 \\ x4 & 16.4 \\ x5 & 2.3 \\ x6 & 5 \\ x7 & 15 \\ x8 & 13.7 \\ x9 & 3.5 \end{array}$$

Suppose a clustering algorithm returned the clusters:

$$C_1 = \{x_2, x_5\}$$

and

$$C_2 = \{x_1, x_3, x_4, x_6, x_7, x_8, x_9\}$$

.

What is the silhouette score

$$s_1$$

of point

$$x_1$$

?

- -0.76
- -0.50
- 0.86
- -0.67