Attempt #1

Jun 19, 11:55 AM

Marks: 27

**Question 1**

Correct Answer

Marks: 3/3

Which of the below statements is/are true about corrections for multiple testing?

1. Bonferroni correction implies that the family-wise error rate is less than or equal to the significance level
2. Holm-Bonferroni correction implies that the family-wise error rate is less than or equal to the significance level
3. Benjamini-Hochberg correction implies that the false discovery rate is less than or equal to the significance level

Only 1 and 2

Only 2

1, 2, and 3

You Selected

Only 2 and 3

1. Bonferroni correction implies FWER ≤ α

2. Holm-Bonferroni correction implies FWER ≤ α

3. Benjamini-Hochberg correction implies FDR ≤ α

**Note:** Refer to “dav\_lecture\_slides\_LVC1\_GL\_MIT\_prof\_Caroline\_Uhler.pdf” corrections for multiple testing

**Question 2**

Correct Answer

Marks: 3/3

Suppose that PC1 and PC2 are the first and the second principal components respectively, then which of the following statements is true?

PC1 will always be orthogonal to PC2

PC1 and PC2 are the eigenvectors of the covariance matrix

PC1 and PC2 will create a plane with the smallest orthogonal distance to all the points

All the statements are true

You Selected

All the Principal Components are perpendicular to each other. The Principal Components are the EigenVectors which represent the direction of vectors. The Principal Components will make a new plane in such a way that it has the smallest orthogonal distance to all the points.

**Question 3**

Incorrect Answer

Marks: 0/3

State whether the following statement is True or False:

"t-SNE gives rise to non-linear embeddings where close-by points remain close-by and far away points remain far away"

True

You Selected

False

Correct Option

t-SNE is a non-linear dimensionality reduction technique that focuses on preserving local structure, meaning that **close-by points will remain close** in the lower-dimensional embedding. However, **far away points do not necessarily remain far away**. t-SNE does not guarantee the preservation of global distances or overall distances between far-away points. Instead, it aims to preserve local relationships and often distorts the global structure in the process.

The primary and most widely used Python package that includes a TSNE class is:

scikit-learn:

Class: sklearn.manifold.TSNE

Description: This is the standard and most commonly used implementation of t-SNE in Python. It's part of the scikit-learn library, which is a comprehensive machine learning library. It includes options for various parameters like n\_components, perplexity, early\_exaggeration, learning\_rate, n\_iter, metric, etc. It supports both exact and Barnes-Hut approximation methods.

Beyond scikit-learn, there are a few other implementations that aim for performance improvements, especially for large datasets:

Multicore-TSNE:

Class: MulticoreTSNE.MulticoreTSNE

Description: This package (often found on GitHub as DmitryUlyanov/Multicore-TSNE) provides a parallelized version of t-SNE that can leverage multiple CPU cores. It's designed to be a near drop-in replacement for sklearn.manifold.TSNE, but with significant speed-ups for larger datasets.

tsne-cuda:

Class: tsnecuda.TSNE

Description: This is a GPU-accelerated implementation of t-SNE, which can provide even more dramatic speed-ups, particularly for very large datasets, by utilizing NVIDIA CUDA-enabled GPUs. It's also often found on GitHub (e.g., CannyLab/tsne-cuda) and aims for an API similar to scikit-learn's TSNE.

When choosing which TSNE class to use, consider:

Dataset Size: For smaller to medium-sized datasets, sklearn.manifold.TSNE is generally sufficient and easy to use.

Performance Needs: For larger datasets where computation time is a concern, Multicore-TSNE (for CPU parallelization) or tsne-cuda (for GPU acceleration) would be better choices if your system supports them.

Yes, **perplexity is a key parameter across the TSNE class in all the packages you mentioned (scikit-learn, Multicore-TSNE, and tsne-cuda).**

In essence, **perplexity is a crucial knob that dictates the "scale" at which t-SNE views your data's neighborhoods**, and adjusting it allows you to uncover different aspects of your data's structure in the lower-dimensional embedding.

Important : Lower the T-SNE value

**Question 4**

Correct Answer

Marks: 3/3

Netflix is an example of which type of network?

Acyclic Network

Bipartite Network

You Selected

Cyclic Network

Simple Network

A movie is associated with a certain rating and a rating does refer to a certain movie or set of movies. But a movie does not refer to another movie or a rating count does not refer to another rating data.

**Question 5**

Correct Answer

Marks: 3/3

Data Scientists tend to make a network with other Data Scientists, this property is called:

Degree Centrality

Closeness Centrality

Homophily

You Selected

Betweenness Centrality

Homophily is the tendency of people to associate with others that are similar.

**Question 6**

Correct Answer

Marks: 3/3

Suppose the police are trying to break the chain of a criminal gang. Which centrality measure they should look to break the network of the criminal gang?

Degree Centrality

Eigenvector Centrality

Betweenness Centrality

You Selected

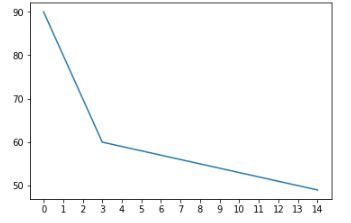
Closeness Centrality

To identify the person that should be removed in order to break the criminal gang Betweenness centrality should be used. The higher the centrality value, the more the likelihood the gang will break on their removal.

**Question 7**

Correct Answer

Marks: 3/3



Y-axis denotes the within-group sum of squared errors

X-axis denotes the number of clusters

According to the above elbow plot, while performing K-Means clustering, what is the ideal value of K to choose?

1

2

3

You Selected

4

As the slope decreases drastically from 2 to 3 and at K=3, the graph takes a sharp turn. Therefore, 3 is considered to be the ideal number of clusters.

**Question 8**

Correct Answer

Marks: 3/3

Which of the below statements is/are true about K-Medoids?

1. In this algorithm, cluster center must be an actual observation from the data.
2. It is a distance based algorithm like K-Means.
3. It is more sensitive to outliers as compared to K-Means.

Only 1 and 2

You Selected

Only 3

Only 2

Only 1 and 3

K-Medoids is a clustering algorithm resembling the K-Means clustering technique. It majorly differs from the K-Means algorithm in terms of the way it selects the clusters’ centers. The former selects the average of a cluster’s points as its center (which may or may not be one of the data points) while the latter always picks the actual data points from the clusters as their centers. K-Medoids is robust to outliers as compared to K-Means.

**Question 9**

Correct Answer

Marks: 3/3

Which of the following statements can be used to define complete linkage in hierarchical clustering?

It is the distance between centroids of two clusters

It is the shortest distance measured between any pair of observations in two clusters

It is the average distance measured between any pair of observations in two clusters

It is the longest distance measured between any pair of observations in two clusters

You Selected

In the case of maximum linkage used as a distance measure in clustering, the distance between 2 clusters is measured as the maximum possible distance between the points in two different clusters

**Question 10**

Correct Answer

Marks: 3/3

Which of the following are the parameters of the DBSCAN algorithm?

A) eps - Defines the neighborhood around a data point

B) MinPts - The minimum number of data points required to form a dense region

Only A

Only B

Both A and B

You Selected

Neither A nor B

The Parameters of DBSCAN are

MinPts - The minimum number of points required to form a dense region.

Eps - Distance measured which is used to locate points in the neighborhood of any point.