Hierarchical Clustering and PCA

Questions to discuss

1. How the clusters are formed in connective based clustering?
2. What are dendrograms, and how to choose different dendrograms?
3. What is Principal Component Analysis?
4. How to use PCA for dimensionality reduction?

How the clusters are formed in connective based clustering?

- Two techniques for cluster formation in connective based clustering, i.e, divisive and agglomerative
  - **Divisive** - Start with one cluster and divide into different clusters
  - **Agglomerative** - Start with different clusters and ultimately clubbing them to form one cluster
- Once a cluster is formed we wish to ‘agglomerate it with another cluster’ in order to reach to one cluster.
- That again is achieved by calculating the distance between these new clusters, ‘closer’ clusters are more probable to be part of the same cluster.
- This process is repeated till we get one cluster containing all our other sub clusters.

What are dendrograms, and how to choose different dendrograms?

- What are dendrograms?
    - Dendrograms are used to represent the distances at which the the different clusters meet.
    - They provide us an idea as to how the clustering looks like diagrammatically .
- Different dendrograms for the same dataset
    - Based on the method chosen to calculate distance between the clusters, the same dataset may result in different dendrograms.
    - Which dendrogram to choose?

What are dendrograms, and how to choose different dendrograms?

- The right choice of dendrogram is done by considering a value known as a cophenetic correlation.
- Dendrogram Distance: the distance between two points/clusters as described by that dendrogram.
- Cophenetic correlation computes the correlation between the euclidean distance and the dendrogram distance for a particular dendrogram of all possible pair of points.
- Performance measure - The dendrogram corresponding to highest correlation coefficient is considered to be better representative of the clustered data and is used to produce labels/ clusters for the dataset.

What is Principal Component Analysis?

- Principal Component Analysis, or PCA, is a method for reducing the dimensionality of data.
- It can be thought of as a projection method where data with m-columns (features) is projected into a subspace with m or fewer columns, whilst retaining the essence of the original data.
- Steps Involved:
  - Begin by standardizing the data. I.e. bring...
  - Generate the covariance matrix
  - Perform eigen decomposition
  - Sort the eigen pairs in descending order and select the largest one.

Covariance Matrix

- Variance is measured within the dimensions and covariance is among the dimensions.

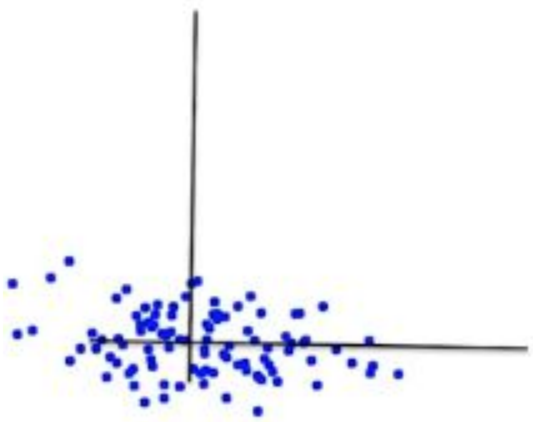
$$\text{var}(X) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})}{(n-1)}$$

$$\text{cov}(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{(n-1)}$$

- In the covariance matrix
    - The diagonal elements represent the variance of the individual attributes
    - The non-diagonal elements represent the covariance between pairs of attributes

Improving SNR through PCA

- The mean is subtracted from all the points on both dimensions.
- The dimensions are transformed using algebra into new set of dimensions.
- The transformation is a rotation of axes in mathematical space.



How to use PCA for dimensionality reduction?

- PCA can also be used to reduce the dimensionality of a dataset.
- Arrange all eigen vectors along with corresponding eigenvalues in descending order of eigenvalues.
- Plot a cumulative eigen value graph.
- Eigenvectors with insignificant contribution to total eigenvalues can be removed from analysis.

greatlearning

Power Ahead

Happy Learning !