Hierarchical Clustering and PCA

Contents

- Hierarchical clustering
  - Distance calculation between data points
  - Cluster formation and dendrograms
  - Cophenetic correlation
- Principal Component Analysis
  - Covariance matrix
  - PCA for dimensionality reduction

Hierarchical Clustering

- **Need to cluster data?**
  - Clustering gives us insights into the distribution of customers, it helps to understand the customers, create segmentation and formulate precise advertising, marketing, logistics mechanisms.

- **Clustering as an unsupervised technique**
  - Clustering is an unsupervised learning technique, which essentially means that there is no label/target value we have with our training data. There is no right and wrong answer and the the groups/clusters depend upon the methods used to reach to that clustering.

Connectivity based Clustering

- **Considers the distance between two data points**.
    - Nearer points are more similar/connected are more probable to be a part of the same cluster.
- **Distance Calculation**
    - Different process to calculate distance between data points as discussed previous week - Euclidean, Manhattan, Chebyshev

Distance Calculation

- The distances between points are calculated the same way it is calculated in a two-dimensional space, i.e considering all the different features/columns as different dimensions.
- Need to scale features/columns before bringing them into distance calculation.
    - To bring all the columns in the same scale so that distance calculation isn't skewed towards one particular feature.
- Finally, distances are calculated as per the scaled features.

Cluster Formation

- Two techniques for cluster formation, 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.

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?

Cophenetic Correlation

- 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.

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.
    - 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.

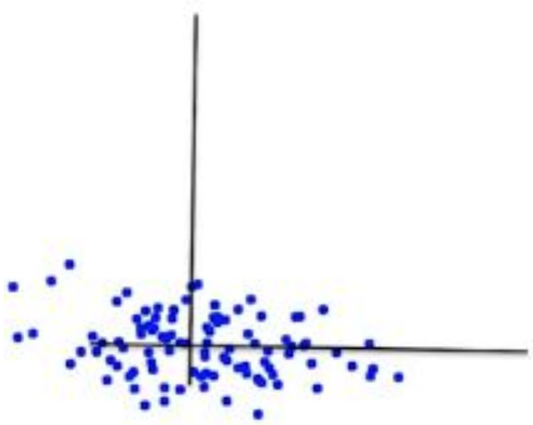
$$\operatorname{var}(X) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(X_i - \bar{X}\right)}{(n-1)}$$

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

- 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.



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 !