greatlearning
Learning for Life

Hierarchical Clustering and PCA

Unsupervised Learning - Topics

- Hierarchical clustering
  - Distance calculation between data points
  - Cluster and dendograms formation
  - Cophenetic correlation

- Principal Component Analysis
  - Principal component Co variance matrix
  - PCA for dimensionality reduction
  - Hands on Exercise

- Case study

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.

greatlearning
Learning for Life

Cluster Formation - underlying algorithm

- 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.

Dendograms

- What are dendograms?
  - Dendograms 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 dendograms for the same dataset -
  - Based on the method chosen to calculate distance between the clusters, the same dataset may result in different dendograms.
  - Which dendogram to choose?

Cophenetic Correlation

The right choice of dendogram is done by considering a value known as a cophenetic correlation.

- Dendogram Distance - distance between two points/clusters as described by that dendogram.

Cophenetic correlation computes the correlation between the euclidean distance and the dendogram distance for a particular dendogram of all possible pair of points.

Performance measure -
The dendogram corresponding to highest correlation coefficient is considered to be better representative of the clustered data and is used to produce labels/ clusters for the data set.

Case Study on hierarchical clustering

Problem Statement -

The data set has information about features of silhouette extracted from the images of different cars Four "Corgie" model vehicles were used for the experiment: a double decker bus, Chevrolet van, Saab 9000 and an Opel Manta 400 cars. This particular combination of vehicles was chosen with the expectation that the bus, van and either one of the cars would be readily distinguishable, but it would be more difficult to distinguish between the cars.

Here, we seek to apply Hierarchical Clustering

Hands On - Contd.

Steps to follow -

1. Importing necessary libraries and the reading the data in dataframe.
2. Using pairplot to visually inspect the number of clusters.
3. Declaring a clustering model, here we use agglomerative clustering.
4. Fitting our data on the model
5. Getting model labels based on the fit and retrieving them in a separate column in our dataframe.
6. Calculating dendograms and cophenetic correlation.
7. Plotting dendograms.
8. Plot the clusters

Principal Component Analysis

Principal Component Analysis, or **PCA** for short, 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:

Begin by standardizing the data.

- Generate the covariance matrix.
- Perform eigen decomposition
- Sort the eigen pairs in descending order and select the largest one.

Principal Component Co variance Matrix

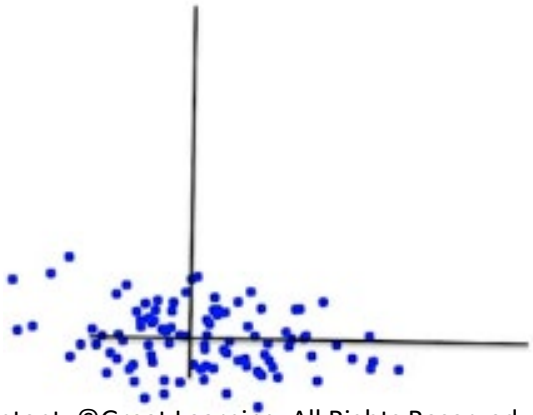
- Variance is measured within the dimensions and co-variance is among the dimensions.
- Express total variance (variance and cross variance between dimensions as a matrix)
- Covariance matrix is a mathematical representation of the total variance of individual dimension and ac

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

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

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.



greatlearning
Learning for Life

PCA for dimensionality reduction

- PCA can also be used to reduce dimensions.
- Arrange all eigen vectors along with corresponding eigen values in descending order of eigen values.
- Plot a cumulative eigen_value graph.
- Eigen vectors with insignificant contribution to total eigen values can be removed from analysis.

Case study on PCA

greatlearning
Learning for Life

Bank Note Authentication (Source: https://archive.ics.uci.edu/ml/datasets/banknote+authentication)

Abstract:

Data were extracted from images that were taken for the evaluation of an authentication procedure for banknotes.

Data Set Information:

Data were extracted from images that were taken from genuine and forged banknote-like specimens. For digitization, an industrial camera usually used for print inspection was used. The final images have 400x 400 pixels. Due to the object lens and distance to the investigated object gray-scale pictures with a resolution of about 660 dpi were gained. Wavelet Transform tool were used to extract features from images.

Case Study - Contd

Dataset information

1. variance of Wavelet Transformed image (continuous)
2. skewness of Wavelet Transformed image (continuous)
3. curtosis of Wavelet Transformed image (continuous)
4. entropy of image (continuous)
5. class (integer)

Case Study - Contd

Steps to follow

1. Import libraries
2. Get data
3. Plot and compare each variables.
4. Split the dataset into train and test set
5. Perform covariance matrix
6. Perform Cumulative Variance Explained
7. Plot explained variance ratio and principal component
8. Calculate eigen pairs and values

Case Study - Contd.

9. Calculate matrix
10. Import SVC library
11. Find out the model
12. Import logistic regression library
13. Perform model using logistic regression
14. Import naive bayes library
15. Perform model using Naive bayes

greatlearning
Learning for Life

Questions if any...