SNS COLLEGE OF ENGINEERING

Kurumbapalayam(Po), Coimbatore – 641 107
Accredited by NAAC-UGC with 'A' Grade
Approved by AICTE, Recognized by UGC & Affiliated to Anna University, Chennai

Department of Information Technology

19IT601– Data Science and Analytics

III Year / VI Semester

Unit 2 – DESCRIPTIVE ANALYTICS USING STATISTICS

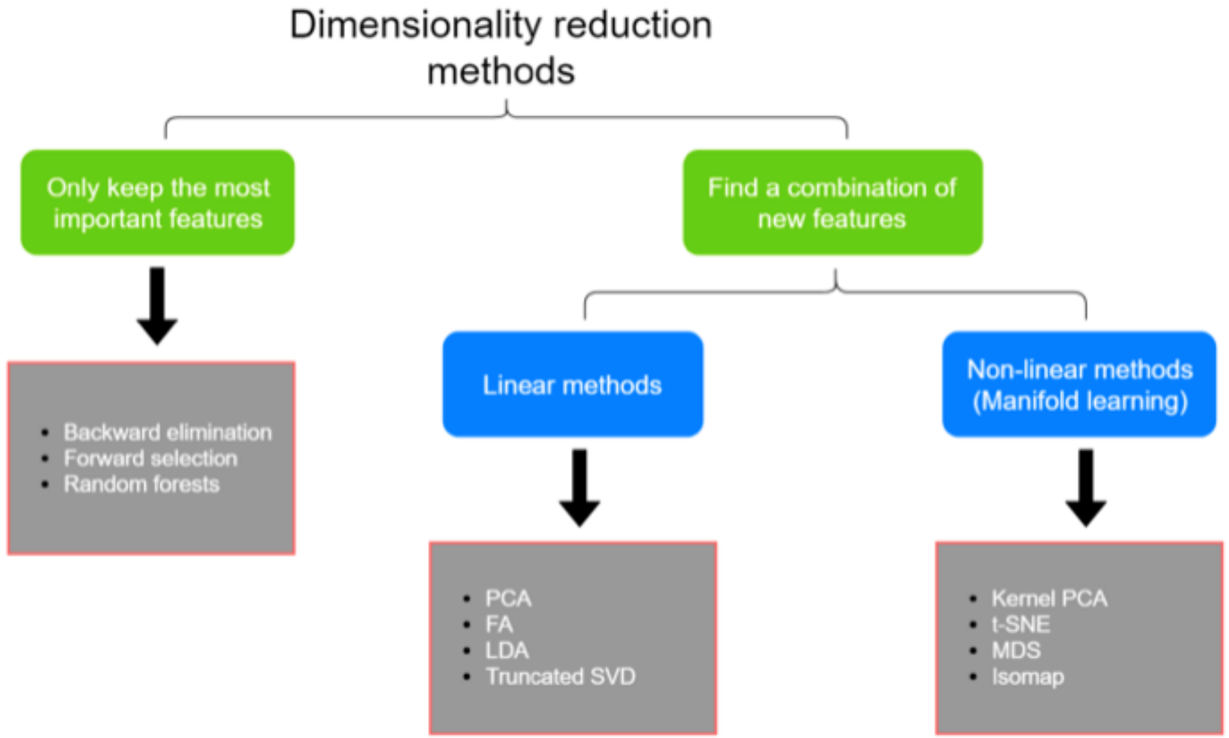
Topic 9: PCA,LDA

Dimensionality reduction

Methods of Dimensionality Reduction

The various methods used for dimensionality reduction include:

- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)



Dimensionality reduction

Principal Component Analysis (PCA)

- This method was introduced by Karl Pearson.

- PCA is a linear dimensionality reduction technique (algorithm) that transforms a set of correlated variables (p) into a smaller k (k<p) number of uncorrelated variables called principal components while retaining as much of the variation in the original dataset as possible.

- In the context of Machine Learning (ML), PCA is an unsupervised machine learning algorithm that is used for dimensionality reduction.

- It works on a condition that while the data in a higher dimensional space is mapped to data in a lower dimension space, the variance of the data in the lower dimensional space should be maximum.

Dimensionality reduction

Principal Component Analysis (PCA)

PCA implementation is quite straightforward. We can define the whole process into just four steps:

1. **Standardization**: The data has to be transformed to a common scale by taking the difference between the original dataset with the mean of the whole dataset. This will make the distribution 0 centered.

2. **Finding covariance**: Covariance will help us to understand the relationship between the mean and original data.

3. **Determining the principal components**: Principal components can be determined by calculating the **eigenvectors and eigenvalues**.

4. **Final output**: It is the dot product of the standardized matrix and the eigenvector.

Dimensionality reduction

Goals of PCA

1.Extract the most important information from the data table.
2.Compress the size of the data set by keeping only this important information
3.Simplify the description of data set
4.Analyze the structure of the observations and variables.

In order to achieve these goals, PCA computes new variables called principal components, which are obtained as linear combinations of the original variables.

Applications of PCA

- Image compression
- Facial Recognition
- Data Visualisation

These applications are most commonly used in Healthcare and Financial Industries.

LDA

Linear Discriminant Analysis

Linear Discriminant Analysis is a dimensionality reduction technique that is commonly used for supervised classification problems. It is used for modelling differences in groups i.e. separating two or more classes.

It is used to project the features in higher dimension space into a lower dimension space.

In 1936, Ronald A.Fisher formulated Linear Discriminant first time and showed some practical uses as a classifier, it was described for a 2-class problem, and later generalized as 'Multi-class Linear Discriminant Analysis' or 'Multiple Discriminant Analysis' by C.R.Rao in the year 1948.

It projects the dataset into moderate dimensional-space with a genuine class of separable features that minimize overfitting and computational costs.

Two criteria are used by LDA to create a new axis:

- Maximize the distance between means of the two classes.
- Minimize the variation within each class.

LDA

Linear Discriminant Analysis

Linear Discriminant Analysis is a dimensionality reduction technique that is commonly used for supervised classification problems. It is used for modelling differences in groups i.e. separating two or more classes.

It is used to project the features in higher dimension space into a lower dimension space.

In 1936, Ronald A.Fisher formulated Linear Discriminant first time and showed some practical uses as a classifier, it was described for a 2-class problem, and later generalized as 'Multi-class Linear Discriminant Analysis' or 'Multiple Discriminant Analysis' by C.R.Rao in the year 1948.

It projects the dataset into moderate dimensional-space with a genuine class of separable features that minimize overfitting and computational costs.

Two criteria are used by LDA to create a new axis:

- Maximize the distance between means of the two classes.
- Minimize the variation within each class.

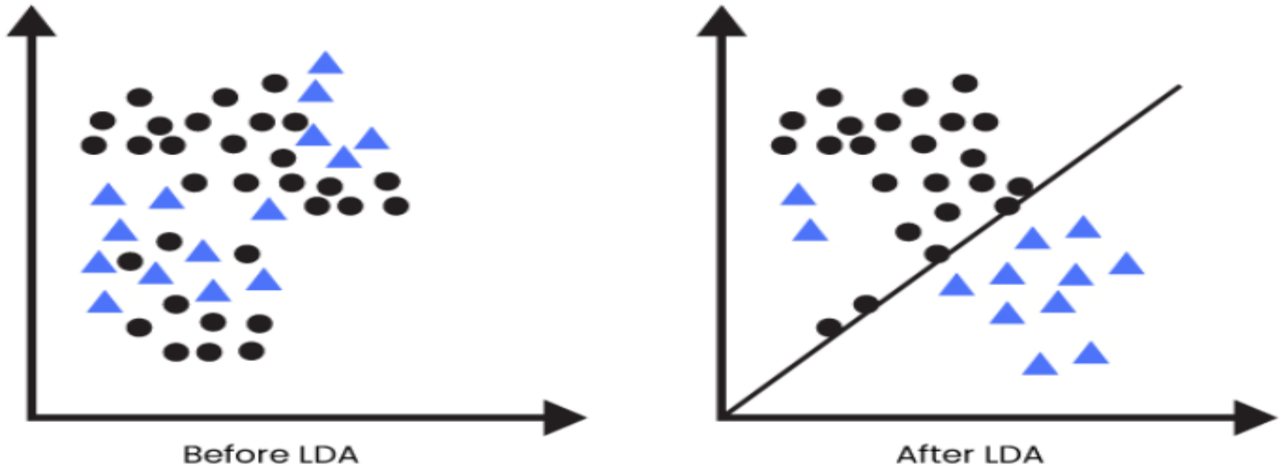
LDA

Steps in LDA

- First step: To compute the separate ability amid various classes,i.e, the distance between the mean of different classes, that is also known as between-class variance.

- Second Step: To compute the distance among the mean and sample of each class,that is also known as the within class variance.

- Third step: To create the lower dimensional space that maximizes the between class variance and minimizes the within class variance.



LDA

Applications

LDA is used in Marketing, Finance, and other areas to perform a number of classification tasks such as customer profiling and fraud detection.

LDA can be used as a classification task for speech recognition, microarray data classification, face recognition, image retrieval, bioinformatics, biometrics, chemistry, etc. below are other applications of LDA.

In medical: LDA is used here to classify the state of patients' diseases as mild, moderate or severe based on the various parameters and the medical treatment the patient is going through in order to decrease the movement of treatment.

THANK YOU