greatlearning
Power Ahead

LINEAR DISCRIMINAT ANALYSIS

An in-depth look at LDA with emphasis on Key Concepts & Working Procedure

greatlearning
Power Ahead

Contents -

Overview-

- In 1936, statistician *Ronald Fisher* presented dichotomous discriminant analysis in "The use of multiple measurements in taxonomic problems." which was later generalized into Linear Discriminant Analysis. LDA became a common method to be used in pattern recognition and machine learning. The core idea behind LDA is to determine a linear combination of features that are able to discriminate between two (or more) classes. This linear combination can also be used for dimensionality reduction.

- This leads to the conclusion that LDA has similar use to both Logistic regression (for classification) and Principal Component Analysis (for dimensionality reduction). Let us have a brief discussion on the comparison of these two techniques.

LDA vs Logistic regression

- LDA and logistic regression are both multivariate statistical methods which are used to determine relationships between different independent variables to the categorical dependent variable.

- In logistic regression, the probability of the data point belonging to a class is obtained or to be more precise, the odds of the plausible outcome are determined (ratio of the probability of the event happening to the probability of the non-occurrence of the event).

- In LDA, the orthogonal (perpendicular to each other) discriminant functions are estimated such that it maximizes the difference of means between the existing groups (class labels) while minimizing the standard deviation within the groups. Thus, the predicted class for a data point will be the one that has the highest value for its corresponding linear function.

Limitations of Logistic Regression

Logistic regression though a very powerful classification algorithm, has certain limitations.

I Multi-class classification –

Logistic regression is primarily intended to be used as binary classifier, although it can be extended to multi-class classification.

II Poor performance with small sample size –

If the sample size is small, then the parameters estimated can be highly unreliable. A good amount of observations (decent sample size) is necessary for a stable logistic regression model.

III Poor performance with well separated classes –

Although it sounds a bit weird, but logistic regression performs terribly when the two classes are well separated. The reason being, if you have features that separate the classes perfectly, the coefficients go off to infinity.

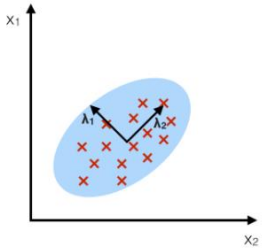
For further clarification on this, please do refer to the following article on "Quasi Separation on Logistic Regression"

All the above three limitations of logistic regression are usually handled by LDA, and thus can be used as an alternate classifier in certain situations.

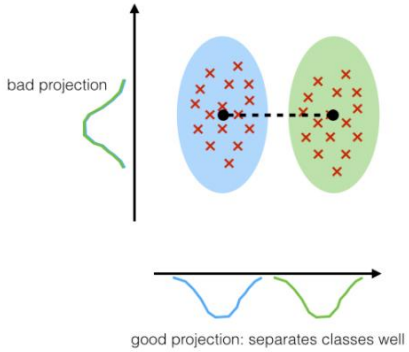
LDA versus PCA -

Both Linear Discriminant Analysis (LDA) and Principal Component Analysis (PCA) are linear transformation methods which closely relate to each other.

PCA:
component axes that maximize the variance



LDA:
maximizing the component axes for class-separation



Source - *Linear Discriminant Analysis – Bit by Bit* By Sebastian Raschka

- The major difference between LDA and PCA lies in the fact that LDA is a supervised learning technique whereas PCA is an unsupervised machine learning algorithm. What it essentially means is that PCA aims towards finding the dimensions which captures the maximum variance irrespective of the class labels (this can be intuitively thought of as treating the entire data as a single class).
- On the other hand, LDA takes into account the different classes present in the data and finds the dimensions which captures the maximum variance & maximum separability among the classes.
- Although we might be under the impression that LDA should always outperform PCA since it directly deals with the class labels, but various studies (PCA versus LDA - Aleix M. Martinez et. al. 2001) have proved this does not necessarily hold true when we have very few samples of certain classes.

Key Concepts and a few terminologies –

Before diving headfirst into the working of the LDA algorithm, let us discuss few key concepts that are very crucial for a holistic understanding of LDA.

1. Bayes' Theorem –

Bayes' theorem relates the conditional and marginal probabilities of Events A and B:

$$P(A|B) = \frac{P(B|A).P(A)}{P(B)}$$

Where ,

- **P(A)** is the prior probability or marginal probability of A. It is "prior" in the sense that it does not take into account any information about B.
- **P(A|B)** is the conditional probability of A, given B. It is also called the posterior probability because it is derived from or depends upon the specified value of B.
- **P(B|A)** is the conditional probability of B given A.
- **P(B)** is the prior or marginal probability of B, and acts as a normalizing constant.

2. Discriminant functions -

- Prediction using Bayes' theorem assumes that the forms for the underlying probability are known, and the training samples are used to estimate the values of their parameters. Discriminant functions on the other hand does not require knowledge of the forms of the underlying probability distributions.
- To find the linear discriminant functions, minimization of a criterion function or training error - the average loss incurred in classifying the set of training samples is done.

$$g(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + w_0$$

Where **g(x)** = Discriminant function

w = Weight vector

w0 = Bias or threshold weight

The bias and weight vectors are initiated with small random numbers (similar to the Weight initialization for neural networks) and these values are updated during minimization of the training error.

The linear discriminant function divides the feature space by a hyperplane decision surface. The orientation of the surface is determined by the normal vector w, and the location of the surface is determined by the bias wo.

How predictions are made using the discriminant functions -

For a discriminant function of the form presented before, a binary classifier implements the following decision rule:

1. If value of discriminant function at x i.e. $g(x) > 0$, then assign x to class I .

2. Similarly if $g(x) < 0$, then assign it to class II.

In the event that $g(x) = 0$, x can be ordinally assigned to either of the two classes, or can be left undefined.

3. Mahalanobis Distance –

The Mahalanobis distance (MD) is the distance between two points in multivariate space. In a regular Euclidean space, variables (e.g. x, y, z) are represented by axes drawn at right angles to each other; The distance between any two points can be measured with a ruler.

For uncorrelated variables, the Euclidean distance equals the MD. However, if two or more variables are correlated, the axes are no longer at right angles, and the measurements become impossible with a ruler. In addition, if you have more than three variables, you cannot plot them in regular 3D space at all. The MD solves this measurement problem, as it measures distances between points, even correlated points for multiple variables.

$MD(\mu_1,\mu_2,\sum) = (\mu_1 - \mu_2)^t \sum^{-1} (\mu_1 - \mu_2)$ [where μ_1 and μ_2 are two population means with a common dispersion matrix $\sum$]

LDA Implementation

Assumptions in LDA –

- Multivariate normality: The independent variables must follow normal distribution.
- Homogeneity of variance/covariance (homoscedasticity): Variances among group variables are the same across levels of predictors.
- Multicollinearity: Predictive power can decrease with an increased correlation between predictor variables.

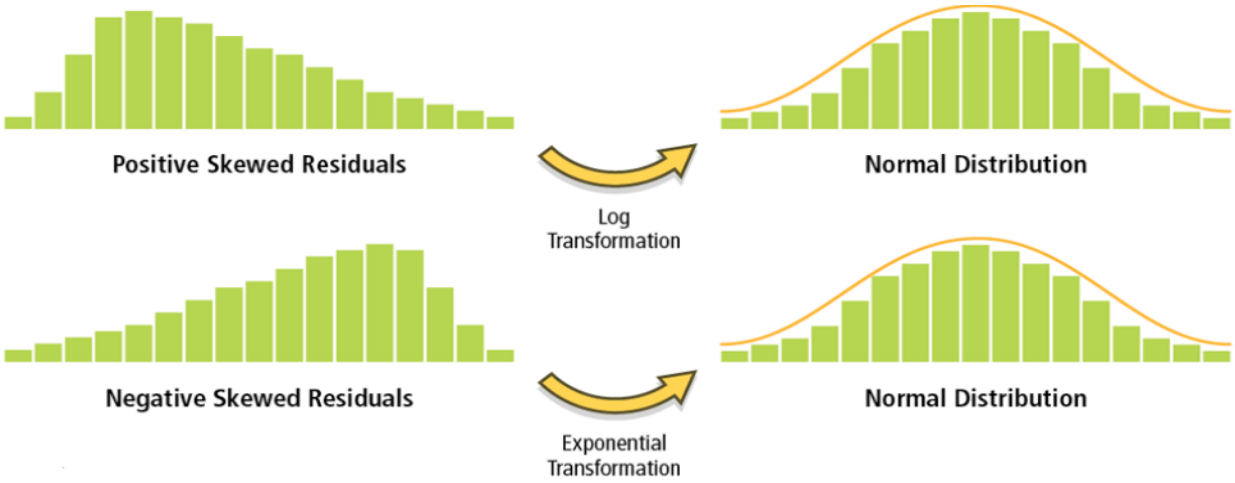
Preparation of Data for LDA –

LDA algorithm has various underlying assumptions about the data discussed above, hence it can be a good idea to prepare the data before applying LDA, whether for classification or for dimensionality reduction. Although past studies suggest that discriminant analysis is relatively robust to slight violations of these assumptions.

Some of the preparatory steps are as follows –

1. Variable Transformation -

LDA assumes a Normal distribution of the input variables. Hence different transformations (e.g log and square root) can be applied to the data to make it more – normal (or near normal).



2. Standardization –

Another LDA assumption is that the input variables have same variance. Thus,

Standardizing the variables to have mean of 0 and a standard deviation of 1 is advised. This can be done when we are looking for the standardized coefficients in the Linear Discriminant Model.

3. Outlier Treatment –

Although linear models are very sensitive to outlier values, it greatly depends on the particular use case whether a high values observation is treated as an outlier or not. After the decision is made, outliers can be treated to avoid skew in the basic statistics.

Mathematical formulation of the LDA

LDA can be derived from simple probabilistic models which model the class conditional distribution of the data for each class . Predictions can then be obtained by using Bayes' rule, for each training sample :

$$P(y=k|x) = \frac{P(x|y=k)\,P(y=k)}{P(x)} = \frac{P(x|y=k)\,P(y=k)}{\sum_l P(x|y=l)\cdot P(y=l)} \quad \text{(eq. 1)}$$

all the expressions have there usual meaning (as discussed earlier)

And we select the class k which maximizes this posterior probability.

Now, P(x|y) is modeled as a multivariate Gaussian distribution with density:

$$P(x|y=k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_k)^t \Sigma_k^{-1}(x-\mu_k)\right) \quad \text{(eq. 2)}$$

Substituting the Probability distribution (eq. 2) in Bayes Rule (eq. 1) , and further simplification, we get :

(eq. 3)

$$\log P(y=k|x) = -\frac{1}{2}(x-\mu_k)^t \Sigma^{-1}(x-\mu_k) + \log P(y=k) + Cst.$$

- The term $(x-\mu_k)^t \sum^{-1} (x-\mu_k)$ corresponds to the Mahalanobis Distance between the sample and the mean. The Mahalanobis distance tells how close x is from the mean μ_k , while also accounting for the variance of each feature.

- We can thus interpret LDA as assigning to the class whose mean is the closest in terms of Mahalanobis distance, while also accounting for the class prior probabilities.

Let's take a look at how P(k | x) can be obtained for continuous variable x.

P(k) is a prior probability that the native class for x is k; and has to be specified by the user. Usually by default all classes receive equal P(k) = 1/number_of_classes. Or, P(k) can be the count of the occurrence of one class divided by the total number of occurrences of all the classes.

P(x|k) is probability of presence of point x in class k, if class being dealt with is k.

The main issue in finding value for this term is that the variables are continuous and not discrete. Hence, we need to compute the Probability Density Function (PDF).

Once we have PDF (x | k) for each of the classes, we can sum it up and normalize it as below -

Consider we have two classes **k and m** –

So,

$$P(k|x)=P(k)*PDF(x|k)/[P(k)*PDF(x|k) + P(m)*PDF(x|m)]$$

And ,

$$P(m|x)= P(m)*PDF(x|m)/[P(k)*PDF(x|k) +P(m)*PDF(x|m)]$$

Now each x will be substituted in the above two equations and the point will be classified to the class for which P(Class | Data) is the highest.

Limitations of LDA –

1. LDA is a parametric method since it assumes unimodal Gaussian likelihoods. If the distributions are significantly non-Gaussian, the LDA projections will not be able to preserve any complex structure of the data, which may be needed for classification

2. LDA will fail when the discriminatory information is not in the mean but rather in the variance of the data.

3. LDA produces at most C-1 feature projections. If the classification error estimates establish that more features are needed, some other method must be employed to provide those additional features

Conclusion:

As we have discussed, LDA is a very useful linear algorithm, which can be used for both Dimensionality reduction and Classification problems. LDA can be seen as an alternative to PCA for dimensionality reduction, and Logistic regression for classification at situations where the aforementioned algorithms do not perform satisfactorily. The procedure of LDA is very straightforward and can be summarized as follows –

1) Find the variance (also called scatter) within and between the classes.

2) Find the linear combinations which maximize the between class variance and minimize the within class variance.

3) Transform the data as per the New Linear combinations (hyperplanes). We have achieved Dimension Reduction till this step.

4) Predict class for each data point using Bayesian approach for the reduced dimensions (top k dimensions)

Use Cases of LDA:

1. Altman's Z-score model

In 1968 an American finance professor Edward Altman developed a model to predict the chances of a business going bankrupt in the next two years. The coefficients used in the model were calculated by using Fisher's Discriminant Analysis.

The formula for the Z-score bankruptcy model is as follows:

$$Z = 0.012X1 + 0.014X2 + 0.033X3 + 0.006X4 + 0.999X5$$

Where , X1, X2, X3, X4 are in percentage points.

X1 = working capital / total assets

X2 = retained earnings / total assets

X3 = earnings before interest and taxes / total assets

X4 = market value of equity / total liabilities

X5 = sales / total assets

2. Facial Recognition:

The aim of a typical facial recognition task is to identify the faces represented by a very large number of pixel values, with each pixel serving as a feature. LDA is often used to reduce the number of features to a more manageable number for further classification. The linear combinations obtained using Fisher's linear discriminant are called Fisher faces.

3. Medical Field:

Linear discriminant analysis (LDA) is used to classify the patient disease state as mild, moderate or severe based upon the patient various parameters and the medical treatment he is going through. This helps the doctors to intensify or reduce the pace of their treatment.

Further Reading :

- An Introduction to Statistical Learning: with Applications in R, Chapter 4, Page 138.
- Modern Multivariate Statistical Techniques: Regression, Classification, and Manifold Learning, chapter 8
- Applied Predictive Modeling, Chapter 12, Page 287
- Linear Discriminant Analysis bit by bit (examples with Python)
- Linear Discriminant Analysis (includes a link to an interactive LDA interface)
- The mathematics behind how sklearn performs Linear Discriminant Analysis.