What is the introduction to Linear Discriminant Analysis?

Authored by

stats writer

April 21, 2024

The introduction to Linear Discriminant Analysis (LDA) is a statistical technique used to classify and differentiate between two or more groups based on a set of variables. It aims to find a linear combination of the variables that maximally separates the groups and reduces the dimensionality of the data. LDA is commonly used in data mining, pattern recognition, and machine learning to analyze and interpret high-dimensional data sets. It is a powerful tool for understanding the relationships between variables and identifying the key factors that contribute to group differences. LDA has numerous applications in various fields, such as biology, finance, psychology, and marketing. It is an essential tool for researchers and data analysts seeking to gain insights and make accurate predictions from complex data sets.

Introduction to Linear Discriminant Analysis

When we have a set of predictor variables and we'd like to classify a response variable into one of two classes, we typically use logistic regression.

For example, we may use logistic regression in the following scenario:

We want to use *credit score* and *bank balance* to predict whether or not a given customer will default on a loan. (Response variable = "Default" or "No default")

However, when a response variable has more than two possible classes then we typically prefer to use a method known as linear discriminant analysis, often referred to as LDA.

For example, we may use LDA in the following scenario:

We want to use *points per game* and *rebounds per game* to predict whether a given high school basketball player will get accepted into one of three schools: Division 1, Division 2, or Division 3.

Although LDA and logistic regression models are both used for **classification**, it turns out that LDA is far more stable than logistic regression when it comes to making predictions for multiple classes and is therefore the preferred algorithm to use when the response variable can take on more than two classes.

LDA also performs better when sample sizes are small compared to logistic regression, which makes it a preferred method to use when you're unable to gather large samples.

How to Build LDA Models

LDA makes the following assumptions about a given dataset:

(1) The values of each predictor variable are **normally distributed**. That is, if we made a histogram to visualize the distribution of values for a given predictor, it would roughly have a "bell shape."

(2) Each predictor variable has the same variance. This is almost never the case in real-world data, so we typically scale each variable to have the same mean and variance before actually fitting a LDA model.

Once these assumptions are met, LDA then estimates the following values:

μk: The mean of all training observations from the kth class.σ2: The weighted average of the sample variances for each of the *k* classes.πk: The proportion of the training observations that belong to the kth class.

LDA then plugs these numbers into the following formula and assigns each observation X = x to the class for which the formula produces the largest value:

Dk(x) = x * (μk/σ2) - (μk2/2σ2) + log(πk)

Note that LDA has *linear* in its name because the value produced by the function above comes from a result of *linear functions* of x.

How to Prepare Data for LDA

Make sure your data meets the following requirements

before applying a LDA model to it:

1. The response variable is categorical. LDA models are designed to be used for classification problems, i.e. when the response variable can be placed into classes or categories.

2. The predictor variables follow a normal distribution. First, check that each predictor variable is roughly normally distributed. If this is not the case, you may choose to first transform the data to make the distribution more normal.

3. Each predictor variable has the same variance. As mentioned earlier, LDA assumes that each predictor variable has the same variance. Since this is rarely the case in practice, it's a good idea to scale each variable in the dataset such that it has a mean of 0 and a standard deviation of 1.

4. Account for extreme outliers. Be sure to check for extreme outliers in the dataset before applying LDA. Typically you can check for outliers visually by simply using boxplots or scatterplots.

Examples of Using Linear Discriminant Analysis

LDA models are applied in a wide variety of fields in real life. Some examples include:

1. Marketing. Retail companies often use LDA to classify shoppers into one of several categories. For example, they may build an LDA model to predict whether or not a given shopper will be a low spender, medium spender, or high spender using predictor variables like *income*, *total annual spending*, and household size.

2. Medical. Hospitals and medical research teams often use LDA to predict whether or not a given group of abnormal cells is likely to lead to a mild, moderate, or severe illness.

3. Product development. Companies may build LDA models to predict whether a certain consumer will use their product daily, weekly, monthly, or yearly based on a variety of predictor variables like *gender*, *annual income*, and *frequency of similar product usage*.

4. Ecology. Researchers may build LDA models to predict whether or not a given coral reef will have an

overall health of good, moderate, bad, or endangered based on a variety of predictor variables like *size*, *yearly contamination*, and *age*.

LDA in R & Python

The following tutorials provide step-by-step examples of how to perform linear discriminant analysis in R and Python:

Linear Discriminant Analysis in R (Step-by-Step)
Linear Discriminant Analysis in Python (Step-by-Step)