

Intro to Classifiers 1: Linear Discriminant Analysis

Ramesh Srinivasan

November 17, 2022

Overview

The problem of Classification

In conventional statistics courses, and experimental psychology or neuroscience courses, emphasis is placed on the notion of finding a *significant* difference between two (or more) subject groups or experimental conditions. For example in clinical research we ask questions such as

Is the patient data different than the control data?

But in our minds (and definitely in the patient and in the physicians mind) perhaps we should ask a different question

Based on the characteristics of the data, can we determine if the data comes from a patient or a control?

The first approach is built around hypothesis testing for differences, the second approach is classification of data.

What is a Classifier

At the simplest level a Classifier is a decision rule, that allows us to categorize data.

For example, when you go to the doctor they take your blood pressure, and you get a pair of numbers like 120 80 for the systolic and diastolic pressure.

The doctor has a decision rule. If the systolic pressure is above 130, the patient receives a stern lecture about diet and exercise, and if the systolic pressure is above 140, medication is prescribed to lower blood pressure.

Thus, there are 3 classes of patients based on the systolic blood pressure reading,

< 130 - healthy -> Pat on Back

130-140 - at risk -> Stern Lecture

> 140 - hypertension -> Medication and (possibly) Stern Lecture.

This is a classifier

How were these critical values found? Huge amounts of data are collected to look at patient cardiovascular health and blood pressure, and the data says if blood pressure remains above 140, the heart walls thicken and secondary cardiovascular diseases can emerge. (There are other bad effects too).

Learning, as in Machine Learning

Machine Learning

- In Cognitive Science and Cognitive Neuroscience research there is much interest in using statistical algorithms to infer patterns from complex data.
- One motivation for this is that data has gotten BIG – there is so much of it that without statistical algorithms we would not be able to make much sense of it.
- But perhaps even more importantly, these algorithms have the potential to give us insight into how the brain and/or mind actually operate.
- Humans are confronted with big data all the time and are very efficient at processing this information and learning about how to interpret sensory information coming from the environment and how to act on the environment to achieve goals. If we can learn how to do this in a computer, we may understand something about intelligence.

Supervised versus Unsupervised Learning







- There are two classes of learning algorithms – supervised and unsupervised learning. The difference between the two has to do with whether OTHER information is brought to bear on the learning process.
- For example, when learning a foreign language, supervised learning would be analogous to having someone repeatedly point to objects and name them until the associations are learned by you.
- Unsupervised learning would be analogous to simply listening to a conversation in a foreign language and using only consistencies in that language to infer aspects of the structure of the language, e.g., what the pronouns, nouns, and verbs are.
- You have already learned about one form of unsupervised learning – Principal Components Analysis, which makes use of the correlation or covariance in data to infer “latent” variables.

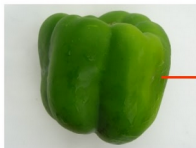
Classification

- Classification, is an area of supervised learning that addresses the problem of how to systematically assign unlabeled (classes unknown) novel data to their labels (classes or groups or types) by using knowledge of their features (characteristics or attributes) that are obtained from observation and/or measurement.
- A classifier algorithm is a specific technique or method for performing classification.
- To classify new data, the classifier algorithm first uses labeled (classes are known) training data to train a model (i.e., fit parameters), and then it uses a function known as its classification rule (or for short, the classifier) to assign a label to each new data input after feeding the input's known feature values into the model to determine how much the input belongs to each class.

The grocery store problem

Objects Features (X) Labels (Y)

	→ (Green, 6, 4, 4.5) →	Green Pepper
	→ (Green, 7, 4.5, 5) →	Green Pepper
	→ (Red, 6, 3, 3.5) →	Red Pepper
	→ (Red, 4.5, 4, 4.5) →	Red Pepper
	→ (Yellow, 1.5, 8, 2) →	Hot Pepper
	→ (Yellow, 1.5, 7, 2.5) →	Hot Pepper



→ (Green, 6, 4, 4.5) $\xrightarrow{h(\text{Green}, 6, 4, 4.5)}$?

Examples in Cognitive Science/Cognitive Neuroscience

- Categorization
- Automatic Speech Recognition
- Face Recognition
- Brain-Computer Interfaces
- Biomarkers
- Single-trial analysis

In data science/machine learning applications, we are solely interested in making classifiers work as accurately as possible.

In scientific applications, we want to know how the classifier was able to work. We want to know what features of the data were useful and what transformations of the data produced the accurate classification.

Linear Discriminant Analysis

Linear Discriminant Analysis

- Linear discriminant analysis (LDA), also called Fisher's linear discriminant and closely related to Logistic Regression, is a method used in statistics, pattern recognition and machine learning to find a linear combination of features which characterize or separate two or more classes of objects or events.
- The idea is to find a function $f(x)$ which maps each input vectors of features directly onto a class or category.
- LDA is also closely related to principal component analysis (PCA) in that both look for linear combinations of variables which best explain the data. LDA explicitly attempts to find the linear combination of variables that best **separates** the classes of data.

Linear versus Quadratic Discriminant Analysis

- Linear discriminant analysis is a statistical method used to find the linear combination of features which best separate two or more classes of objects or events. It is widely applied in classifying diseases, positioning, product management, and marketing research. LDA assumes that the different classes have the same covariance matrix (among the features).
- Quadratic Discriminant Analysis, on the other hand, aims to find the quadratic combination of features. It is more general than linear discriminant analysis. Unlike LDA, QDA does not make the assumption that the different classes have the same covariance matrix (among the features), but works well even when that assumption is violated.

Two class discrimination problem

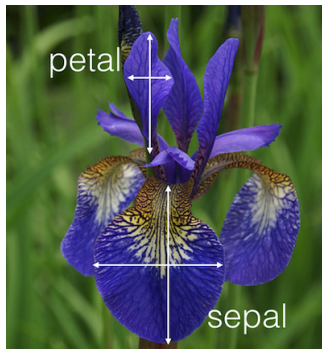
- Let x be a vector of features (this just means a sample containing multiple variables that want to use to do the classification).
- There is data from two classes $y=1$ and $y=2$, and the mean of the feature vectors for each class are μ_1 and μ_2 and the covariance matrices are Σ_1 and Σ_2 .
- We have to find a transformation that satisfies two goals: 1. To make the means of the classes as far apart as possible. We are going to project the data onto one direction and we want to find the direction that makes the means as far apart as possible. 2. To make the variance of the data after projection as small as possible. Ideally we would like to project the data onto two small clusters of points as far away from each other as possible so they are easy to tell apart.
- If we can satisfy the above, we can make a simple decision rule to separate the two classes

Suspiciously like an ANOVA model

- Assume w is the projection vector (an orthonormal vector like an eigenvector) we want to apply to a row vector of data x that maximizes the discrimination of the classes. Then the projection is $w \cdot x$. Our goal is to find w and to make a decision rule.
- The mean of the data x for each class is a vector μ_1 and μ_2 . Thus the mean can also be projected as $w \cdot \mu_1$ and $w \cdot \mu_2$.
- In the transformed space the vector difference between the means is $w \cdot \mu_1$ and $w \cdot \mu_2$. We want to maximize the length of this vector.
- In addition the covariance of the data in each class is transformed as $w \Sigma_1 w^T$ and $w \Sigma_2 w^T$. These need to be minimized.

You might notice that we are maximizing differences between groups and minimizing differences within. In standard statistics (e.g., ANOVA analysis, we compare the difference between means of groups to the variability within groups

Fisher Iris Data



Data Structure

In these data there are three different flower varieties - Setosa, Versicolor, and Virginica.

$$X = \begin{bmatrix} X_{1\text{sepal length}} & X_{1\text{sepal width}} & X_{1\text{petal length}} & X_{1\text{petal width}} \\ X_{2\text{sepal length}} & X_{2\text{sepal width}} & X_{2\text{petal length}} & X_{2\text{petal width}} \\ \dots & & & \\ X_{150\text{sepal length}} & X_{150\text{sepal width}} & X_{150\text{petal length}} & X_{150\text{petal width}} \end{bmatrix} \quad y = \begin{bmatrix} \omega_{\text{setosa}} \\ \omega_{\text{setosa}} \\ \dots \\ \omega_{\text{virginica}} \end{bmatrix}$$

Instead of only 1 mean, there are three different means corresponding to each flower.

$$m_i = [\mu_{\omega_i}(\text{sepal length}) \quad \mu_{\omega_i}(\text{sepal width}) \quad \mu_{\omega_i}(\text{petal length}) \quad \mu_{\omega_i}(\text{petal width})]$$

with $i = 1, 2, 3$ corresponding to the 3 flower - Setosa, Versicolor, and Virginica

Covariance Matrices

The within-class covariance matrix Σ_W is computed by the following equation:

$$\Sigma_W = \sum_{i=1}^c \Sigma_i$$

where

$$\Sigma_i = \frac{1}{n} (x_j - m_i)(x - m_i), x \in D_i$$

The between-class covariance matrix Σ_B is computed by the following equation:

$$\Sigma_B = \sum_{j=1}^c \Sigma_j$$

where

$$\Sigma_j = \frac{1}{n_{class}} (m_j - m)(m_j - m), m_j \in class$$
 The eigenvalues and eigenvectors we are

interested in are the eigenvalues of

$$\Sigma_W^{-1} \Sigma_B$$

Solution - Generalized Eigenvalue Problem

$$\Sigma_W^{-1} \Sigma_B$$

- The matrix Σ_W is the covariance matrix WITHIN CLASSES. You calculate this separately for each class and add them together, i.e., for two classes $\Sigma_W = \Sigma_1 + \Sigma_2$.
- The matrix Σ_B is a measure of distance between the classes. It is the covariance matrix you would get if the data were exactly the mean of each class.