Pattern Recognition

Lecture 09. Fisher's Linear Discriminant Analysis

Dr. Shanshan ZHAO

School of AI and Advanced Computing
Xi'an Jiaotong-Liverpool University

Academic Year 2023-2024

Table of Contents

1. Overview

2. Fisher's Linear Discriminant

1 Overview

2 Fisher's Linear Discriminant

Overview

- **Classification based on Bayes Decision Theory**
  - Bayes Decision Theory
  - Minimizing the classification error probability
  - Minimizing the Average Risk
  - Discriminant functions for Normal densities and decision surfaces
    - a linear function when all classes share same covariance matrix
- **Parametric Estimation (parameters estimation for known probability)**
  - MLE
  - MAP
- **Non-Parametric Estimation (parameters esitmation for unknown probability)**
  - Kernel density estimation
  - KNN (from the density to classification)

Introduction

Generative methods

The major concern was to design classifiers based on probability density or probability functions.

- Parametric Methods
- non-Parametric Methods

In some cases, we saw that the resulting classifiers were equivalent to a set of linear discriminant functions.

Discriminative methods

In this part, we will focus on the design of linear classifiers, regardless of the underlying distributions describing the training data.
Assumption: all feature vectors from the available classes can be classified correctly using a linear classifier

- Linear Discriminant Functions; Artificial Neural Networks; Support Vector Machines; etc.

1. Overview

2. Fisher's Linear Discriminant

- Discriminant analysis is a generative approach to classification, which requires fitting an multi-variate normals to the features. this can be problematic in high dimensions.
- An alternative approach is to reduce the dimensionality of the features $x \in \mathscr{R}^D$ and then fit an MVN to the resulting low-dimensional features $z \in \mathscr{R}^K$.
- The simplest approach is to use a linear projection matrix, $z = Wx$, where W is a $K \times D$ matrix.

Fisher's linear discriminant attempts to find the vector that maximizes the separation between classes of the projected data.

Fisher's Linear Discriminant

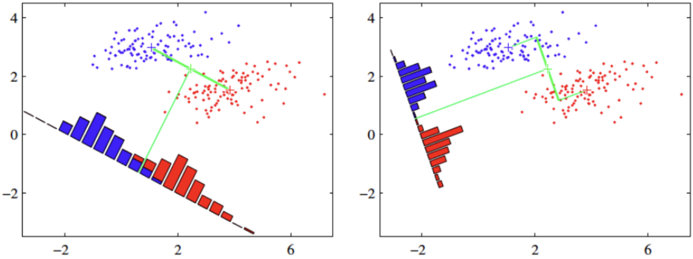


Figure 1: The left plot shows samples from two classes (depicted in red and blue) along with the histograms resulting from projection onto the line joining the class means. Note that there is considerable class overlap in the projected space. The right plot shows the corresponding projection based on the Fisher linear discriminant, showing the greatly improved class separation[**bishop2006pattern**][**LDA**].

Fisher's Linear Discriminant

- to find a linear combination of features which characterizes or separates two or more classes of objects or events
- used as a linear classifier or dimensionality reduction

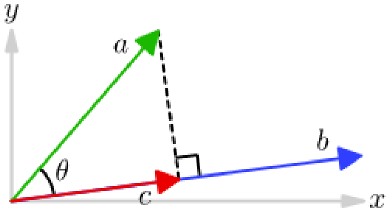
Fisher Linear Discriminant project to a line which preserves direction useful for data classification.

The transformation is based on maximizing the ratio of **"between-class variance"** to **"within-class variance"**, aiming at reducing data variation in the same class and in increasing the separation between class.

Fisher's Linear Discriminant

Preliminary 1. **Projection from vector a to vector b**

- Scalar $|c|$ is the projection from a to b.
- If b is a unit vector which means $|b| = 1$, then the scalar projection $|c|$ can be represented as $a \cdot b$.



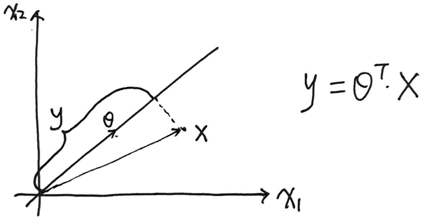
$$a.b = |a||b|cos(\theta)$$

$$a.b = a_x b_x + a_y b_y$$

$$|c| = |a|cos(\theta)$$

Fisher's Linear Discriminant

Our goal is seeking to obtain a **scalar** y by projecting the samples X onto a line:



Then try to find the θ^* to maximize the ratio of "between-class variance" to "within-class variance".

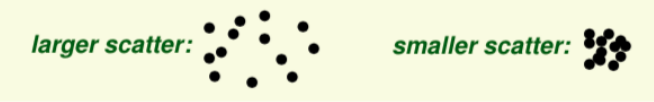
Fisher's Linear Discriminant

Preliminary 2. **Introduce scatter**

- There are samples $z_1, z_2, ..., z_n$, the sample mean is $\mu_z = \frac{1}{n}\sum_{i=1}^{n} z_i$
- Define samples **scatter** as

$$s = \sum_{i=1}^{n}(z_i - \mu_z)^2$$

- Scatter is just sample variance multiplied by n, it measures the spread of data around the mean.
- Scatter measures the same thing as variance, but on different scale.



Fisher's Linear Discriminant

Let's see how to use mathematical way to present this problem.

- Assume we have a set of D-dimensional samples $X = \{x_1, x_2, ...x_m\}$, N_1 of which belong to class C_1, and N_2 of which belong to class C_2. We also assume the mean vector of two classes in X-space:

$$\mu_k = \frac{1}{N_k} \sum_{i \in C_k} x_i, \quad k = 1, 2.$$

- and in y-space:

$$\hat{\mu}_k = \frac{1}{N_k} \sum_{i \in C_k} \theta^T x_i = \theta^T \mu_k, \quad k = 1, 2$$

Fisher's Linear Discriminant

The **between-class variance** is

- to define a **measure of separation** between two classes is to choose the distance between the projected means

$$\hat{\mu}_2 - \hat{\mu}_1 = \theta^T(\mu_2 - \mu_1)$$

The **within-class variance** for each class C_k is

- use scatter

$$\hat{s}_k^2 = \sum_{i \in C_k} (y_i - \hat{\mu}_k)^2 \quad k = 1, 2$$

Now that we get the between-class variance and within-class variance, we can define our **objective function** $J(\theta)$ as:

$$J(\theta) = \frac{(\hat{\mu}_2 - \hat{\mu}_1)^2}{\hat{s}_1^2 + \hat{s}_2^2}$$

Fisher's Linear Discriminant

Objective function

$$J(\theta) = \frac{(\hat{\mu}_2 - \hat{\mu}_1)^2}{\hat{s}_1^2 + \hat{s}_2^2}$$

If maximizing the objective function $J(\theta)$, we are looking for a projection where

- examples from the class are projected very close to each other (small scatter, which is the denominator)
- the projected means are as further apart as possible (large difference between two mean value, which is in the numerator)

Fisher's Linear Discriminant

To find the optimum θ^*, we must express $J(\theta)$ as a function of θ.

- 1. The scatter of the projection y can then be expressed as

$$\begin{aligned}\hat{s}_k^2 &= \sum_{i \in C_k} (y_i - \hat{\mu}_k)^2 \\ &= \sum_{i \in C_k} (\theta^T x_i - \theta^T \mu_k)^2 \\ &= \sum_{i \in C_k} \theta^T (x_i - \mu_k)(x_i - \mu_k)^T \theta \\ &= \theta^T S_k \theta\end{aligned}$$

So we can get,

$$\hat{s}_1^2 + \hat{s}_2^2 = \theta^T S_1 \theta + \theta^T S_2 \theta = \theta^T S_W \theta$$

Fisher's Linear Discriminant

where we denote the class k scatter in feature space-x as:
$S_k = \sum_{i \in C_k} (x_i - \mu_k)(x_i - \mu_k)^T$
within-class scatter matrix: $S_W = S_1 + S_2$

Fisher's Linear Discriminant

- 2. Similarly, the difference between the projected means can be expressed in terms of the means in the original feature space:

$$\begin{aligned}(\hat{\mu}_2 - \hat{\mu}_1)^2 &= (\theta^T \mu_2 - \theta^T \mu_1)^2 \\ &= \theta^T (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T \theta \\ &= \theta^T S_B \theta\end{aligned}$$

where we denote the between-class scatter matrix:
$S_B = (\mu_2 - \mu_1)(\mu_2 - \mu_1)^T$

We can finally express the Fisher criterion in terms of S_W and S_B as:

$$J(\theta) = \frac{\theta^T S_B \theta}{\theta^T S_W \theta}$$

Fisher's Linear Discriminant

Solve the Problem

The easiest way to maximize the object function J is to derive it and set it to zero.
For now, the problem has been solved and we just want to get the direction of the θ, which is the optimum θ^*

$$\theta^* \propto S_W^{-1}(\mu_2 - \mu_1)$$

This is known as Fisher's linear discriminant(1936), although it is not a discriminant but rather a specific choice of direction for the projection of the data down to one dimension, which is $y = {\theta^*}^T X$. **[LDA]**

Conclusion

- **Gaussian LDA**

$$\boldsymbol{w} = \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

- **Fisher's Linear Discriminant**

$$\theta^* \propto S_W^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$$

where $S_W = S_1 + S_2$

extension

Find a Fisher's Linear Discriminant solution, and apply it to *Iris datset* , taking a two-class problem as example.

Thank You !

$\mathscr{Q}$ & $\mathscr{A}$