Linear models for classification

Blanca Cano Camarero

Universidad Autónoma de Madrid

25 November

Presentation Overview

1. Classification introduction
    - Problem contextualization
    - Math formulation
2. Discriminant Function
3. Multiple classes
4. Least squares for classification
5. Fisher's linear discriminant
    - Relation between Fisher discriminant and least squares
    - Fisher's discriminant for multiple classes
6. The perceptron algorithm

Goal of classification

The goal in classification is to take an input vector x an assign it to one of the the $K \in \mathbb{N}$ discrete classes $\mathscr{C}_k$ where $k \in \{1, \dots, K\}$.
Some properties:

- In the most common scenario, the classes are taken to be disjoint.
- The input space is thereby dividen into *decision regions* whose boundaries are called *decision boundaries* or surfaces.

If we consider linear models for classification, the decisión surface are linear functions of the input vector $x \in \mathbb{R}^D$, and hence are defined by $(D-1)$-dimensional hyperplanes.

Problem formulation

- Probabilistic models: $K = 2$, two-class problems where:
  - There is a single target variable
$$t \in \{0, 1\}$$
  - Class C_1 is represented by $t = 1$,
  - Class C_2 is represented by $t = 0$.
  - We can interpret the value of t as the probability of class C_1 taking only extreme values $\{0, 1\}$.
- For $K > 2$ classes : **1 - of- K** coding scheme:

$$t \in \{0, 1\}^K \text{where one only appears one time.}$$

Distinct appraches to the classification problem

- Constructing a ***discriminant function*** that directly assigns each vector x to a specific class.
- Using $p(C_k|x)$ *parametric function*.
- Using $p(C_k|x)$ with a generative approach.

Math formulation

We consider a generalization of lineal model, for which we transform the linear function of w using a nonlinear function f

$$y(x) = f(w^t x + w_0). \tag{1}$$

- In machine learning literature f is known as *activation function*.
- In statistics literature. The *link function* provides the relationship between the linear predictor and the mean of the distribution function [1], its inverse.

[1] Read **[?]**

Discriminant functio a geometrical understanding

A discriminant is a function that takes an input vector x and assigns it to one of K classes denoted by C_k.
We shall restrict attention to *linear discriminats* (decision surface are hyperplanes).
Two classes

$$y(x) = w^T x + w_0 \tag{2}$$

where

- w is called *weight vector*.
- w_0 is a *bias*

The decision boundary is defined by the relation

$$y(x) = 0 \tag{3}$$

.

w determines the orientation of the decision surface

Theorem

The vector w is orthogonal to every vector lying within the decision surface.

Proof.

Let v a vector lying within the decision surface, it can be write as the difference of two point

$$v = v_1 - v2. \tag{4}$$

Those points verify that

$$y(v_i) = w^t v_i + w_0 = 0, \tag{5}$$

Hence

$$0 = y(v_1) - y(v_2) = w^t(v_1 - v_2). \tag{6}$$

□

The bias parameter w_0 determines the location of the deision surface

Theorem

The normal distance from the origin to the decision surface is given by

$$\frac{w^T x}{\|w\|} = -\frac{w_0}{\|w\|}. \tag{7}$$

Multiple classes: By two class discriminant functions

One versus the rest classifier has a problem: leads to regions of input space that are ambiguously classified.

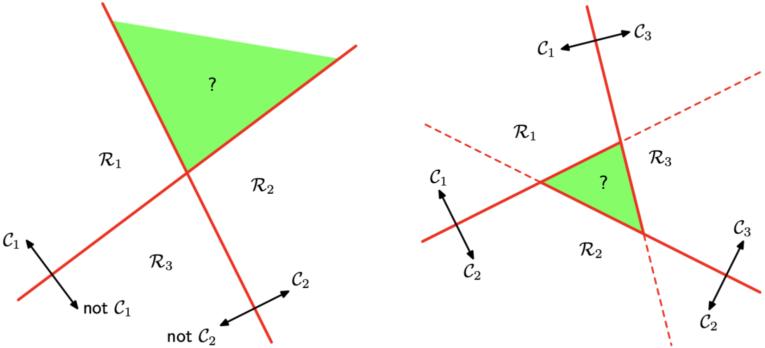

Figure 4.2 Attempting to construct a K class discriminant from a set of two class discriminants leads to ambiguous regions, shown in green. On the left is an example involving the use of two discriminants designed to distinguish points in class $\mathcal{C}_k$ from points not in class $\mathcal{C}_k$. On the right is an example involving three discriminant functions each of which is used to separate a pair of classes $\mathcal{C}_k$ and $\mathcal{C}_j$.

Alternative: *One versus one classifier*

Each point is classified according to a majory amongst the discriminant function. One for ever $k(k-1)/2$ binary discriminat functions .
Problem: This too runs into the problem of ambiguous regions.

K- Class discriminant

Comprising K linear functions of the form

$$y_k(x) = w_k^T x + w_{k0} \tag{8}$$

and the assigning a point x to class C_k if

$$y_k(x) > y_j(x) \text{ for all } j \neq k. \tag{9}$$

The decision boundary between class C_k and class C_j is therefore given by $y_k(x) = y_j(x)$ and hence correspond to a $(D-1)$dimensional hyperplane

$$(w_k - w_j)^T x + (w_{k0} - w_{j0}) = 0. \tag{10}$$

Properties

- Regions of such a discriminant are always singly connected and convex.

Let x_a, x_b two points both of which lie inside decision region. Any point $\hat{x}$ that lies on the line can be expressed in the form

$$\hat{x} = \lambda x_a + (1-\lambda)x_b \tag{11}$$

where $0 \leq \lambda \leq 1$.
It follows that

$$y_k(\hat{x}) = \lambda y_k(x_a) + (1-\lambda)y_k(x_b). \tag{12}$$

Introduction

Motivation: Approximates the conditional expectation $\mathbb{E}[t|x]$
These probabilities are typically approximated rather poorly, can have values outsides the range $(0, 1)$.

Matrix equivalence

$$y_k(x) = y_k^T x + w_{k0} \tag{13}$$

where $k \in \{1, \ldots, K\}$. Using vector notation

$$y(x) = \tilde{W}^T \tilde{x} \tag{14}$$

where $\tilde{W}$ is a matrix whose k^{th} column comprise the $D + 1$-dimensiona vector $\tilde{w}_k = (w_{k0}, w_k^T)^T$ and $\tilde{x}$ is the corresponding augmented input vector $(1, x^T)^T$.

Determining $\tilde{W}$ by minimization error function

The sum of squared error function can be written as

$$E_D(\tilde{W}) = \frac{1}{2} Tr\{(\tilde{X}\tilde{W} - T)^T(\tilde{X}\tilde{W} - T)\}. \tag{15}$$

Setting the derivative with respect to $\tilde{W}$ to zero we obtain the discriminant function in the form

$$\tilde{W} = (\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T T = \tilde{X}^\dagger T. \tag{16}$$

The discriminat function in the form

$$y(x) = \tilde{W}^T \tilde{x} = T^T \left(\tilde{X}^{\dagger} \right)^T \tilde{x}. \tag{17}$$

Properties

If every target vector in the training set satisfies some linear constraint

$$a^T t_n + b = 0 \tag{18}$$

for some constant a and b, then the model for any value prediction will satisfy the same constraint so that

$$a^T y(x) + b = 0. \tag{19}$$

Thus if we use a $1-$of$-K$ coding scheme for K classes, then the prediction made by the model will have the property that the elements of $y(x)$ will sum to one for any value of x.
This summation constraint alone is not sufficient to allow the model outputs to be interpreted as probabilities because they are not constrained to lie within the interval $(0, 1)$.

Problems

- Lack robustness to outliers.

Introduction to Fisher's linear discriminant

Consider first the case of two classes, and suppose we take the $D-$dimensional input vecto x and project it down to one dimension using

$$y = w^T x. \tag{20}$$

Place a threshold on y and classify $y \geq -w_0$ as class C_1 otherwise class C_2.
We can select a projection that maximizes the class separation

Math formulation: first idea

The mean vectors of the two classes are given by

$$m_i = \frac{1}{N_i} \sum_{n \in C_i} x_n \tag{21}$$

where $i \in \{1,2\}$ and there are N_i points of class C_i.
The simplest measure of the separation of the classes, when projected onto w, is the separation of the projected class mean. This suggests that we might choose w so as to maximize

$$c_2 - c_2 = w^T(m_2 - m_1) \tag{22}$$

where

$$c_i = w^T m_i \tag{23}$$

is the mean of the projected data form class C_i.

Problem of this approach

This expression can be made arbitrarily large simply by increasing the magnitude of w. Solution: constrain w to have a unit length so that

$$\sum_i w_i^2 = 1. \tag{24}$$

Use a Lagrange multiplier to perform the constrained maximization, we then find that

$$w \propto (m_2 - m_1). \tag{25}$$

Problems: Overlap when projected

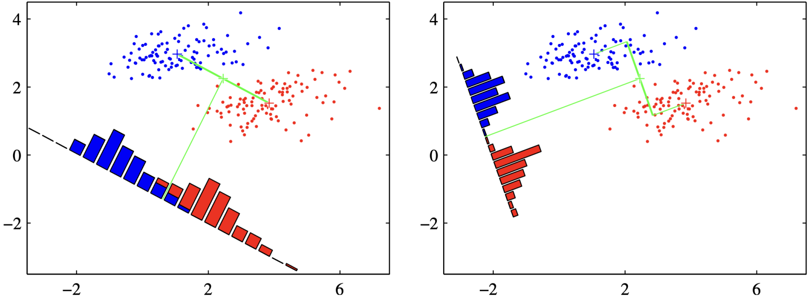

Figure 4.6 The left plot shows samples from two classes (depicted in red and blue) along with the histograms resulting from projection onto the line joining the class means. Note that there is considerable class overlap in the projected space. The right plot shows the corresponding projection based on the Fisher linear discriminant, showing the greatly improved class separation.

Fisher's solution

Maximize a function a function that will give a large separation between the projected class means while also giving a small variance within each class, **thereby minimizing the class overlap**. the within-class variance of the transformed data from class C_k is therefore given by

$$s_k^2 = \sum_{n \in C_k} (y_n - c_k)^2 \tag{26}$$

where $y_n = w^T x_n$.
We can define the total within-class variance for the whole data set to be simply $s_1^2 + s_2^2$.

Fisher criterion definition

The Fisher criterion is defined to be the ration of the between-class class variance to the within class variance and is given by

$$J(w) = \frac{(m_2 - m_1)^2}{s_1^2 + s_2^2}. \tag{27}$$

We can make the dependence on w explicit by using (20) (23) (26) to rewrite the Fisher criterion in the form

Fisher explicit form

$$J(w) = \frac{w^T S_B w}{w^T S_W s} \tag{28}$$

where S_B is the *between class* covariance matrix and is given by

$$S_B = (m_2 - m_1)(m_2 - m_1)^T \tag{29}$$

and S_W is the total *within class* covariance matrix, given by

$$S_W = \sum_{k \in \{1,2\}} \sum_{n \in C_i} (x_n - m_k)(x_n - m_k)^T . \tag{30}$$

Maximizing

Differentiating 28 with respect to w, we find that $J(w)$ is maximized when

$$(w^T S_B w) S_W w = (W^T S_W w) S_B w. \tag{31}$$

Some consideration:

- From 29 we see that $S_B w$ is always in the direction of $(m_2 - m_1)$.
- We do not care about the magnitude of w, only its direction, and so we can drop the escalar factors $(w^T S_B w)$ and $(w^T S_W w)$.

Multiplying both side of 31 by S_W^{-1} we obtain the **Fisher's linear discriminant**

$$w \propto S_W^{-1}(m_2 - m_1). \tag{32}$$

Note that if the within class covariance is isotropic (proportional to unit matrix), we find that S_W is proportional to the difference of the class mean.

Some observation

It is not a discriminant but rather a specific choice of direction for projection to one dimension.
In order to construct a discriminant we can choose a threshold y_0 so that we classify a new point as belonging to C_1 if $y(x) \geq y_0$.

Fisher criterion can be obtained as a special case of least squares

We shall take the targets for class C_1 to be $\frac{N}{N_1}$ and for class $-\frac{N}{N_2}$. The sum if squares error function can be written

$$E = \frac{1}{2}\sum_{n=1}^{N}(w^T x_n + w_0 - t_n)^2. \tag{33}$$

Derivatives of E with respect to w_0 and w to zero

erivatives of E with respect to w_0 and w to zero

$$\sum_{n=1}^{N} (w^T x_n + w_0 - t_n) = 0, \tag{34}$$

$$\sum_{n=1}^{N} (w^T x_n + w_0 - t_n) x_n = 0. \tag{35}$$

Working with 34

Since

$$\sum_{n=1}^{N} t_n = N_1 \frac{N}{N_1} - N_2 \frac{N}{N_2} = 0 \tag{36}$$

and

$$\sum_{n=1}^{N} x_n = Nm = N_1 m_1 + N_2 m_2. \tag{37}$$

Substituting 37 and 36 in 34 we obtain an expression for the bias in form

$$w_0 = -w^T m. \tag{38}$$

Working with 35

$$\left(S_W + \frac{N_1 N_2}{N} S_B \right) w = N(m_1 - m_2) \tag{39}$$

Fisher's discriminant for multiple classes

$$S_W = \sum_{k=1}^{K} S_k \tag{40}$$

where

$$S_k = \sum_{n \in C_k} (x_n - m_k)(x_n - m_k)^T \tag{41}$$

$$m_k = \frac{1}{N_k} \sum_{n \in C_k} x_n \tag{42}$$

Introduction

Is another importan example of a linear discriminant model. Was introduced by Rosenblatt in 1962. Two class model

$$y(x) = f(w^T \phi(x)) \tag{43}$$

where the nonlinear activation function f is given by a step function of the form

$$f(x) = \begin{cases} +1, & a \geq 0 \\ -1, & a < 0 \end{cases} \tag{44}$$

The vector $\phi(x)$ will typically include a bias component $\phi_0(x) = 1$.