Linear Discriminative Models

Ying Wu

Electrical Engineering and Computer Science
Northwestern University
Evanston, IL 60208

http://www.eecs.northwestern.edu/~yingwu
Outline

Linear Discriminant Functions

Fisher Criterion

Perceptron Criterion

Minimum Misclassification Criterion

Minimum Squared-Error Criterion

Minimum Mean-Squared-Error Criterion
The linear discriminative model is

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

where \( \mathbf{w} \) is the weight vector and \( w_0 \) is the bias.

- The sign of \( g(\mathbf{x}) \) can be used to discriminate two classes
- The decision boundary is a hyperplane
- \( g(\mathbf{x}) = 0 \) gives an algebraic form of the hyperplane \( H \)
- The distance \( r \) from \( \mathbf{x} \) to \( H \) is

\[ r = \frac{g(\mathbf{x})}{||\mathbf{w}||} \]

- The orientation of \( H \) depends on \( \mathbf{w} \) and the location on \( w_0 \).
- For multiple-classes, one such linear function can be used for per class.
Generalized Linear Discriminant Function

- The linear form can be generalized

\[ g(x) = \sum_{i=1}^{d'} a_i \phi_i(x) = a^T \Phi(x) = a^T y \]

where \( \phi_i(x) \) can be an arbitrary function of \( x \)

- Example

\[ g(x) = a_1 + a_2 x + a_3 x^2 \]

- The order of \( \phi_i(x) \) can be very high, even infinity

- \( \Phi(x) : \mathbb{R}^d \rightarrow \mathbb{R}^{d'} \), maps the original feature space to another feature space

- The new feature space can have a much higher dimensionality

- A hyperplane in \( \mathbb{R}^{d'} \) corresponds to a nonlinear surface in \( \mathbb{R}^d \)

- We’ll see this in Kernel Machines.
Linearly Separable and Separating Vector

- If a linear discriminant function can classify all the samples correctly, this set of samples are said to be *linearly separable*
- \( y_i \rightarrow \omega_1 \) if \( a^T y_i > 0 \), and \( y_i \rightarrow \omega_2 \) if \( a^T y_i < 0 \)
- If we use \(-y_i\) to replace \( y_i \) for \( \omega_2 \) class, then
  \[
  a^T y_i > 0, \quad \forall i
  \]
- We call this *normalization*, and \( a \) the *separating vector*
- This only gives a set of inequality constraints on \( a \)
- The solution to \( a \) is of course not unique
- Unless we impose its optimality \( J(a) \)
- We can design a gradient-based iterative procedure
  \[
  a(k + 1) = a(k) - \eta(k) \nabla J(a(k))
  \]
  where \( \eta > 0 \) is call *learning rate* or step size.
As said, the solution to $\mathbf{a}$ is not unique

Some $\mathbf{a}$ are good, some are not so good

A margin $b$ can be introduced

$$\mathbf{a}^T \mathbf{y}_i > b, \quad \forall i$$

where $b > 0$.

This gives a subset of the original solution region

Insulated from the old boundary by $\frac{b}{\|\mathbf{y}_i\|}$

It avoids the boundary
Designing Linear Discriminative Models

- A labelled training data set $\mathcal{X} = \{x_1, \ldots, x_n\}$, or its generalized set $\mathcal{Y} = \{y_1, \ldots, y_n\} = \{\Phi(x_1), \ldots, \Phi(x_n)\}$
- A criterion to be optimized $J(w, w_0)$ or $J(a)$
- A good procedure to find the best parameters $w$ and $w_0$, or $a$
Outline

Linear Discriminant Functions

Fisher Criterion

Perceptron Criterion

Minimum Misclassification Criterion

Minimum Squared-Error Criterion

Minimum Mean-Squared-Error Criterion
Fisher Linear Discriminative Model

We have discussed this before in LDA.

\[
\max_w J_F(w) = \frac{w^T S_b w}{w^T S_w w}
\]

where

\[
S_b = (m_1 - m_2)(m_1 - m_2)^T
\]

and

\[
S_w = \sum_{x \in \mathcal{X}_1} (x - m_1)(x - m_1)^T + \sum_{x \in \mathcal{X}_2} (x - m_2)(x - m_2)^T
\]

The solution is given by

\[
w = S_w^{-1}(m_1 - m_2)
\]

Please refer to my previous lectures for details.
Outline

Linear Discriminant Functions

Fisher Criterion

Perceptron Criterion

Minimum Misclassification Criterion

Minimum Squared-Error Criterion

Minimum Mean-Squared-Error Criterion
We have normalized set \( \mathcal{Y} = \{y_1, \ldots, y_n\} \)

Define
\[
J_p(a) = \sum_{y \in M} (-a^T y)
\]

where \( M(a) \) is the set of samples misclassified by \( a \)

\( J_p(a) \) is proportional to the sum of distances from the misclassified samples to the hyperplane

It is clear that \( J_p(a) \geq 0 \)

Its derivative is
\[
\nabla J_p = \sum_{y \in M} (-y)
\]

Iterative procedures can be designed based on this gradient
Batch Updating vs. Single-Sample Updating

- Batch updating uses all the training samples for each update

\[ a(k + 1) = a(k) + \eta(k) \sum_{y \in M_k} y \]

- Single-sample updating uses only one sample for each update
- Each misclassified sample gives a correction to \( a \)
- **Fixed-increment Single-sample Perceptron**

\[ a(k + 1) = a(k) + y^k \]

where \( y^k \) is a misclassified sample

- In the linearly separable case, this procedure converges to a separating vector
Some Generalizations

- **Variable-increment Single-sample Perceptron**
  \[ a(k + 1) = a(k) + \eta(k)y^k \]

- **Variable-increment Perceptron with Margin**
  \[ a(k + 1) = a(k) + \eta(k)y^k, \quad \text{if } a^Ty^k \leq b \]

- **Batch Variable-increment Perceptron**
  \[ a(k + 1) = a(k) + \eta(k) \sum_{y \in M_k} y \]
Outline

Linear Discriminant Functions

Fisher Criterion

Perceptron Criterion

Minimum Misclassification Criterion

Minimum Squared-Error Criterion

Minimum Mean-Squared-Error Criterion
Motivation and Idea

- Perceptron is limited, as it only applies to the linearly separable cases
- Its iterative never ends for nonseparable cases
- We need a criterion for both cases
- One idea:
  - In the linearly separable case, it acts like Perceptron
  - In the nonseparable case, minimize the number of misclassified samples
Given a set of normalized samples \( \{y_1, \ldots, y_n\} \)

Denote by \( Y = [y_1, \ldots, y_n]^T \)

To have a robust solution, a margin can be introduced \( Ya \geq b \)

Define the criterion

\[
J_{q1}(a) = \| (Ya - b) - |Ya - b| \|^2
\]

Or another one

\[
J_{q2}(a) = \sum_{i=1}^{n} \frac{1 + sgn(a^T y_i - b)}{2}
\]
Outline

Linear Discriminant Functions

Fisher Criterion

Perceptron Criterion

Minimum Misclassification Criterion

Minimum Squared-Error Criterion

Minimum Mean-Squared-Error Criterion
From Inequalities to Equalities

- In our previous formulations, we attempted to solve a set of inequality constraints.
- Things will be easier if looking at the set of equality constraints:

  \[ a^T y_i = b_i > 0, \quad \forall i \]

- Denote by \( Y = [y_1, \ldots, y_n]^T \).
- We have its matrix form:

  \[ Ya = b \]

- We introduce an objective function:

  \[ J_s(a) = \|Ya - b\|^2 \]

- It is well known that the solution is:

  \[ a^* = Y^\dagger b \]

where \( Y^\dagger = (Y^TY)^{-1}Y^T \) is the pseudo-inverse of \( Y \).
The Widrow-Hoff Procedure

- Problems in pseudo-inverse
  - We are in trouble if $\mathbf{Y}^T\mathbf{Y}$ is singular
  - We are in trouble if $\text{dim}(\mathbf{Y}^T\mathbf{Y})$ is large
- An iterative method is preferred
- It is clear the gradient is

$$\nabla J_s(\mathbf{a}) = 2\mathbf{Y}^T(\mathbf{Y}\mathbf{a} - \mathbf{b})$$

- We can easily design a batch gradient-based updating

$$\mathbf{a}(k + 1) = \mathbf{a}(k) + \eta(k)\mathbf{Y}^T(\mathbf{Ya} - \mathbf{b})$$

- Or the one-sample updating (Widrow-Hoff Procedure)

$$\mathbf{a}(k + 1) = \mathbf{a}(k) + \eta(k)[b_k - \mathbf{y}_k^T\mathbf{a}(k)]\mathbf{y}_k$$
Relation to Fisher Linear Discriminant

- This is closely related to Fisher discriminant.
- The dataset is \( \mathcal{X} = \{x_1, \ldots, x_n\} \) that consists of samples from two classes (\( \omega_1 \) and \( \omega_2 \)).
- We partition \( Y, a \) and \( b \) to be

\[
Y = \begin{bmatrix} 1_1 & X_1 \\ -1_2 & -X_2 \end{bmatrix}, \quad a = \begin{bmatrix} w_0 \\ w \end{bmatrix}, \quad b = \begin{bmatrix} n_1 1_1 \\ n_2 1_2 \end{bmatrix}
\]

- \( Y^T Ya = Y^T b \) becomes

\[
\begin{bmatrix} 1_1^T & -1_2^T \\ X_1^T & -X_2^T \end{bmatrix} \begin{bmatrix} 1_1 & X_1 \\ -1_2 & -X_2 \end{bmatrix} \begin{bmatrix} w_0 \\ w \end{bmatrix} = \begin{bmatrix} 1_1^T & -1_2^T \end{bmatrix} \begin{bmatrix} n_1 1_1 \\ n_2 1_2 \end{bmatrix}
\]

- Using the notation \( m_i \) and \( S_w \), we have

\[
\begin{bmatrix} n \\ (n_1 m_1 + n_2 m_2) \end{bmatrix} S_w + n_1 m_1 m_1^T + n_2 m_2 m_2^T \begin{bmatrix} w_0 \\ w \end{bmatrix} = \begin{bmatrix} 0 \\ n(m_1 - m_2) \end{bmatrix}
\]
Relation to Bayesian Classification (cont.)

- From the first equation, we have $w_0 = -m^T w$
- Plug it into the second equation, we have
  \[
  \left[ \frac{1}{n} S_w + \frac{n_1 n_2}{n^2} (m_1 - m_2)(m_1 - m_2)^T \right] w = m_1 - m_2
  \]
- As $(m_1 - m_2)(m_1 - m_2)^T w \propto (m_1 - m_2)$, we have
  \[
  S_w w = c (m_1 - m_2)
  \]
  where $c$ is a constant.
- In other words
  \[
  w = c S_w^{-1} (m_1 - m_2)
  \]
- This is Fisher linear discriminant!
Relation to Bayesian Classification

- If we let $b = 1_n$, i.e., 1 for every normalized sample
- Then the MSE solution approaches asymptotically to the Bayes discriminant function

$$g_0(x) = P(\omega_1|x) - P(\omega_2|x)$$

as we have an infinite number of samples

- In other words, the mean-squared-error between the MSE linear discriminant $g(x) = a^T y$ and the Bayes discriminant $g_0(x)$

$$e^2 = \int [a^T y - g_0(x)]^2 p(x) dx$$

is minimized by

$$a^* = Y^\dagger b = Y^\dagger 1_n$$
Proof

For unnormalized data (i.e., \(\{y, 1\} \rightarrow \omega_1\) and \(\{y, -1\} \rightarrow \omega_2\))

\[
J_s(a) = n \left[ \frac{n_1}{n} \frac{1}{n_1} \sum_{y \in Y_1} (a^T y - 1)^2 + \frac{n_2}{n} \frac{1}{n_2} \sum_{y \in Y_2} (a^T y + 1)^2 \right]
\]

By the law of large numbers, as \(n \rightarrow \infty\), \(J_s(a) \xrightarrow{p=1} \bar{J}(a)\)

\[
\bar{J}(a) = \int (a^T y - 1)^2 p(x, \omega_1) dx + \int (a^T y + 1)^2 p(x, \omega_2) dx
\]

After simple manipulations

\[
\bar{J}(a) = \int (a^T y)^2 p(x) dx - 2 \int a^T y g_0(x) p(x) dx + 1
\]

\[
= \int [a^T y - g_0(x)]^2 p(x) dx + \left[ 1 - \int g_0^2(x) p(x) dx \right]
\]

Minimizing \(\bar{J}(a)\) → minimizing \(e^2\)
- $a^T y$ gives direct information about $P(\omega_1|\mathbf{x})$ and $P(\omega_2|\mathbf{x})$
- The quality of the approximation depends on the $y_i(\mathbf{x})$
- But MSE criterion emphasizes on samples where $p(\mathbf{x})$ is larger (i.e., dense regions)
- instead of those that are close to the decision surface
- Thus, it does not necessarily minimize the probability of error
Outline

Linear Discriminant Functions

Fisher Criterion

Perceptron Criterion

Minimum Misclassification Criterion

Minimum Squared-Error Criterion

Minimum Mean-Squared-Error Criterion
Minimum Mean-Squared-Error Criterion

- We talked about deterministic sample set. Now we treat all samples as random variables, and have stochastic sample set.

- Suppose a generative model for the random samples: select a class according to $P(\omega_i)$, and then select a sample $x$ according to $p(x|\omega_i)$.

- Each sample is associated with a label $z$, where
  \[
  z = \begin{cases} 
  +1 & x \in \omega_1 \\ 
  -1 & x \in \omega_2 
  \end{cases}
  \]

- we have $P(z = 1|x) = P(\omega_1|x)$ and $P(z = -1|x) = P(\omega_2|x)$

- and $E[z(x)] = P(\omega_1|x) - P(\omega_2|x) = g_0(x)$

- Define the stochastic MMSE criterion
  \[
  J_m(a) = E[(a^T y - z)^2]
  \]

- Let’s compare this and the MSE criterion
Similarly we can prove that $g_m(x) = a^T y$ approximates the Bayes discriminant $g_0(x)$ asymptotically, i.e.,

$$e^2 = E[(a^T y - g_0(x))^2]$$

is also minimized by the MMSE solution.

Let’s find the MMSE solution. See the gradient

$$\nabla J_m(a) = 2E[(a^T y - z(x))y]$$

Then we have $a^* = E[yy^T]^{-1}E[z(x)y]$

Let’s compare this and the pseudo-inverse solution at the deterministic case.
Iterative Solutions

- We can have the Widrow-Hoff procedure

\[ a(k + 1) = a(k) + \eta(k)(z_k - a^T(k)y_k)y_k \]

as long as an appropriate \( \eta(k) \) is used

- As this converges slowly, we can do better by using Newtonian method

\[ a(k + 1) = a(k) + E[yy^T]^{-1}E[(z - a^Ty)y] \]

- This can be done stochastically

\[ a(k + 1) = a(k) + R_{k+1}(z_k - a^T(k)y_k)y_k \]

with

\[ R_{k+1}^{-1} = R_k^{-1} + y_ky_k^T \]

or equivalently

\[ R_{k+1} = R_k - \frac{R_ky_k(R_ky_k)^T}{1 + y_k^TR_ky_k} \]