Xi'an Jiaotong-Liverpool University

Pattern Recognition

Lecture 06. Linear and quadratic discriminant analysis: Gaussian densities

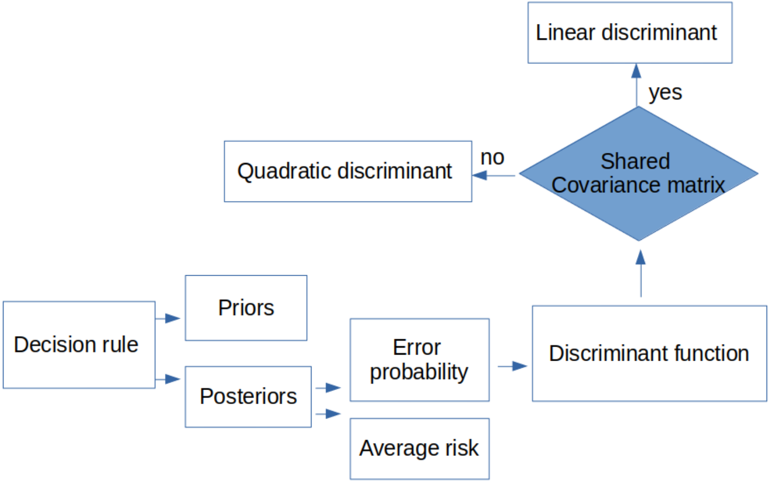
Dr. Shanshan ZHAO

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

Academic Year 2023-2024

Table of Contents

recap


Linear discriminant
yes
Quadratic discriminant
no
Shared Covariance matrix
Decision rule
Priors
Posteriors
Error probability
Average risk
Discriminant function

Discriminant Functions

Minimizing either the risk or the error probability is equivalent to partitioning the feature space into M regions, for a task with M classes. For the minimum error probability case, this described by the equation

$$P(\omega_i|x) - P(\omega_j|x) = 0$$

Sometimes, it may be more convenient to work with equivalent function of them, for example $g_i(x) \equiv f(P(\omega_i|x))$, where $f(\cdot)$ is a monotonically increasing function. $g_i(x)$ is known as a *discriminant function*. The decision rule is now stated as

$$\text{Decide } x \text{ in } \omega_i \text{ if} \quad g_i(x) > g_j(x) \quad \forall j \neq i$$

The decision boundaries, separating regions are described by

$$g_{ij}(x) \equiv g_i(x) - g_j(x) = 0, \quad i,j = 1,2,...,M, \quad i \neq j$$

Discriminant Functions

This is precisely what we mentioned in previous lectures when classifying based on the values of the log posterior probability. Thus the log posterior probability of class ω_k given a data point x is a possible discriminant function.

$$g_k(x) = \ln P(\omega_k|x) = \ln p(x|\omega_k) + \ln P(\omega_k) + const.$$

Decision boundaries are not changed by monotonic transformation(such as taking the log) of the discriminant functions.

Discriminant Function for Normal density

What is the form of the discriminant function when using a Gaussian *pdf*? As before, we take the discriminant function as the log posterior probability:

$$
\begin{aligned}
g_k(x) = \ln P(\omega_k|x) &\equiv \ln p(x|\omega_k) + \ln P(\omega_k). \\
&= -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\ln|\Sigma_k| + \ln P(\omega_k)
\end{aligned}
$$

- We have dropped the term $-1/2\ln(2\pi)$, since it is a constant that occurs in the discriminant function for each class.
- The first term on the right hand side of the equation is quadratic in the elements of x (i.e., if you multiply out the elements, there will be some terms containing x_i^2 or $x_i x_j$).

Linear Discriminants

Take the discriminant function as the log posteriori probability

$$
\begin{aligned}
g_k(x) &= \ln P(\omega_k|x) \equiv \ln p(x|\omega_k) + \ln P(\omega_k). \\
&= -\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\ln|\Sigma_k| + \ln P(\omega_k)
\end{aligned}
$$

Consider the case in which the Gaussian pdfs for each class all share the same covariance matrix.
That is, for all classes ω_k, $\Sigma_k = \Sigma$. Therefore, the discriminant function can be:

$$
\begin{aligned}
g_k(x) &= -\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) + \ln P(\omega_k) \\
&= -\frac{1}{2}(x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_k - \mu_k^T\Sigma^{-1}x + \mu_k^T\Sigma^{-1}\mu_k) + \ln P(\omega_k) \\
&= \mu_k^T\Sigma^{-1}x - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \ln P(\omega_k)
\end{aligned}
$$

Linear Discriminants

The linear discriminant function:

$$
\begin{aligned}
g_k(x) &= \mu_k^T \Sigma^{-1} x - \frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \ln P(\omega_k) \\
&= w_k^T x + w_{k0}
\end{aligned}
$$

where,

$$
\begin{aligned}
w_k^T &= \mu_k^T \Sigma^{-1} \\
w_{k0} &= -\frac{1}{2}\mu_k^T \Sigma^{-1} \mu_k + \ln P(\omega_k)
\end{aligned}
$$

This is a linear equation in D dimensions, We refer to w_k as the weight vector and w_{k0} as the bias for class ω_k.

Case 1

Diagonal covariance matrix with **equal** elements, *which means* $\Sigma = \sigma^2 \mathrm{I}$, I *is the D-dimensional indenty matrix*, and

$$g_k(x) = w_k^T x + w_{k0} = \mu_k^T \Sigma^{-1} x + w_{k0}$$
$$= \frac{1}{\sigma^2} \mu_k^T x + \ln P(\omega_k) - \frac{1}{2} \frac{||\mu_k||^2}{\sigma^2}$$

the corresponding decision hyperplanes can now be written as

$$g_{ij}(x) = g_i(x) - g_j(x) = w^T (x - x_0) = 0$$

where,

$$w = \mu_i - \mu_j$$

and

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \sigma^2 \ln(\frac{P(\omega_i)}{P(\omega_j)}) \frac{\mu_i - \mu_j}{||\mu_i - \mu_j||^2}$$

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \sigma^2 \ln(\frac{P(\omega_i)}{P(\omega_j)}) \frac{\mu_i - \mu_j}{||\mu_i - \mu_j||^2}$$

- the decision surface is a hyperplane passing through the point x_0.
- if $P(\omega_i) = P(\omega_j)$, then $x_0 = \frac{1}{2}(\mu_i + \mu_j)$,and the hyperplane passes through the average of μ_i and μ_j.
- On the other hand, if $P(\omega_j) > P(\omega_i)$, the hyperplane is located closer to μ_i. *In other words, the area of the region where we decide in favor of the more probable of the two classes is increased.*
- the decision hyperplane (straight line) is orthogonal to $(\mu_i - \mu_j)$.

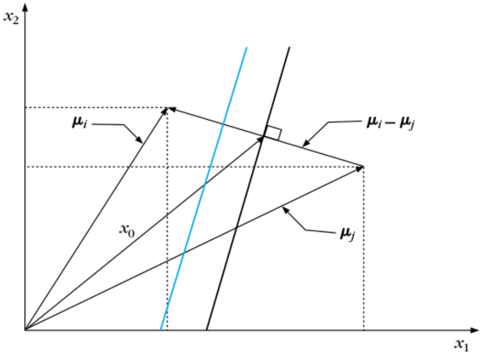


FIGURE 2.10

Decision lines for normally distributed vectors with $\Sigma = \sigma^2 I$. The black line corresponds to the case of $P(\omega_j) = P(\omega_i)$ and it passes through the middle point of the line segment joining the mean values of the two classes. The red line corresponds to the case of $P(\omega_j) > P(\omega_i)$ and it is closer to $\boldsymbol{\mu}_i$, leaving more "room" to the more probable of the two classes. If we had assumed $P(\omega_j) < P(\omega_i)$, the decision line would have moved closer to $\boldsymbol{\mu}_j$.

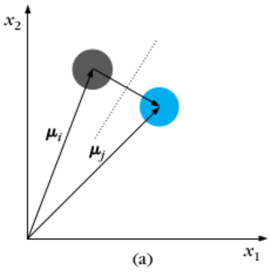


FIGURE 2.11

Decision line (a) for compact and (b) for noncompact classes. When classes are compact around their mean values, the location of the hyperplane is rather insensitive to the values of $P(\omega_1)$ and $P(\omega_2)$. This is not the case for noncompact classes, where a small movement of the hyperplane to the right or to the left may be more critical.

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \sigma^2 \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right) \frac{\mu_i - \mu_j}{||\mu_i - \mu_j||^2}$$

case 2

Nondiagonal covariance matrix: Following algebraic arguments similar to those used before, we end up with hyperplanes described by

$$g_{ij}(x) = g_i(x) - g_j(x) = w^T(x - x_0) = 0$$

where,

$$w = \Sigma^{-1}(\mu_i - \mu_j)$$

and

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right)\frac{\mu_i - \mu_j}{[(\mu_i - \mu_j)^T\Sigma^{-1}(\mu_i - \mu_j)]}$$

- The comments made before for the case of the diagonal covariance matrix are still valid,
- with one exception, The decision hyperplane is no longer orthogonal to the vector $(\mu_i - \mu_j)$ but orthogonal to its linear transformation $\Sigma^{-1}(\mu_i - \mu_j)$.

Perceptron

We now consider a two-class linear discriminant function whose output is binary : 0 for Class 0, and 1 for Class 1. This can be achieved by applying a unit step function $g(a)$ to the output of linear discriminant, so that the binary-output discriminant function is defined as

$$y(x) = g(w^T x + w_0)$$

where,

$$g(a) = \begin{cases} 1, & \text{if } a \geq 0 \\ 0, & \text{if } a < 0 \end{cases}$$

- This type of discriminant function is called 'perceptron', which was invented by Frank Rosenblatt in late 1950s.
- The Rosenblatt's original perceptron has very limited ability, but it has been extended in various ways, and it forms the basis of modern artificial neural networks.

Multi-layer Perceptron

Although the original perceptron is just a linear classifier, we can combine more than one perceptron to form complex decision boundaries and regions.

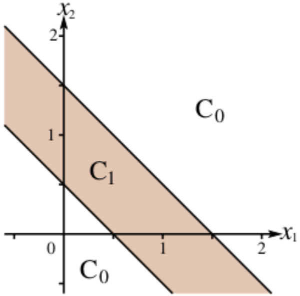
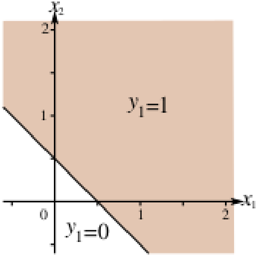


Figure 1: An example of a data set that is not linearly separable. There are two decision boundaries and three disjoint regions.

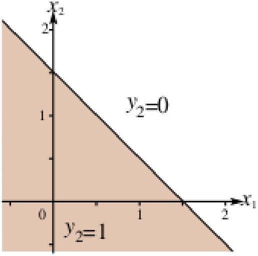
Multi-layer Perceptron

- Although each of the decision boundaries is linear, the data set is not linearly separable, and a single perceptron is unable to have more than one decision boundary.
- To tackle this problem, we start with considering two perceptrons, M_1 and M_2.
- each of them is responsible for one of the two decision boundaries.

Multi-layer Perceptron



(a) M_1

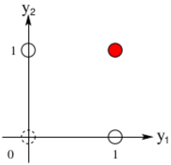


(b) M_2

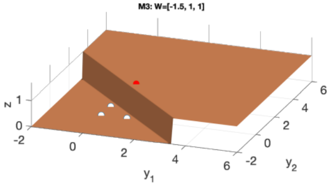
Figure 2: Decision boundaries and regions of M_1 and M_2. It can be confirmed that the intersection of the dark regions (where $y_1 = 1$ and $y_2 = 1$) corresponds to Class 1.

Multi-layer Perceptron

Since the output y_1 and y_2 take binary values, 0 or 1, there are only four possible combinations,
$\{(y1, y2)\} = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$, among which only the pair (1, 1) corresponds to Class 1, and $\{(0, 1), (1, 0)\}$ to Class 0.



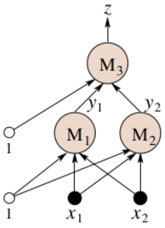
(a) Illustration of (y_1, y_2) plane.



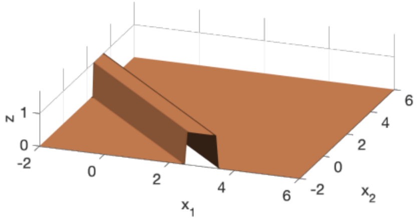
(b) Output of M_3.

It is easy to see that the point $(1, 1)$ in (y_1, y_2) plane can be separated from the other points by a single line, which can be done with another perceptron, say M_3 , taking (y_1, y_2) as input, and giving z as output: $z(y) = g(w_3^T y)$, where $y = (1, y_1, y_2)^T$.

Multi-layer Perceptron



(c) Structure of the multi-layer perceptron comprised of M_1 , M_2 , and M_3 .



(d) Output of the multi-layer perceptron.

This example indicates that multi-layer perceptrons can form complex decision boundaries and regions.

Quadratic Discriminant Analysis (QDA)

Without those assumptions, i.e., when the quadratic term exist because of the covariance matrix.
It is the

- Quadratic discriminant function:

$$
\begin{aligned}
g_k(x) &= \ln P(\omega_k|x) \equiv \ln p(x|\omega_k) + \ln P(\omega_k). \\
&= -\frac{1}{2}(x-\mu_k)^T \Sigma_k^{-1}(x-\mu_k) - \frac{1}{2}\ln|\Sigma_k| + \ln P(\omega_k)
\end{aligned}
$$

QDA

$$\mu_1 = [0,0]; \mu_2 = [4,0]; P(\omega_1) = P(\omega_2)$$

$$\Sigma_1 = \begin{bmatrix} 0.3 & 0.0 \\ 0.0 & 0.35 \end{bmatrix} \Sigma_2 = \begin{bmatrix} 1.2 & 0.0 \\ 0.0 & 1.85 \end{bmatrix}$$

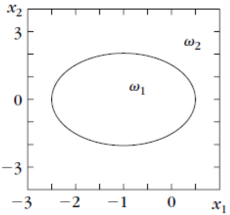


Figure 3: Ellipsoid decision boundary

$$\Sigma_1 = \begin{bmatrix} 0.1 & 0.0 \\ 0.0 & 0.75 \end{bmatrix} \Sigma_2 = \begin{bmatrix} 0.75 & 0.0 \\ 0.0 & 0.1 \end{bmatrix}$$

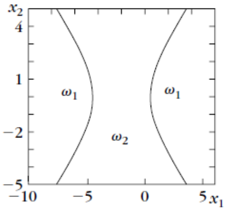


Figure 4: Hyperbolas decision boundary

YOU LOVE CODING ♡

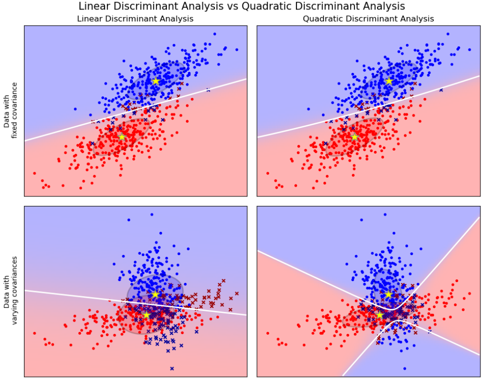


Figure 5: skitlearn(*click*)

Thank You !

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