Chapter 3: Classificatoin

Linear and Quadratic Discrimination

Ga young Moon

2025-02-23

Descendants of Lagrange
School of Mathematics, Statistics and Data Science
Sungshin Women's University

Outline

Linear / Quadratic Discrimination

- In discriminant analysis, the discriminant function is divided into two types according to its form.

- In the case of a discriminant function in the form:
  - a straight line $\rightarrow$ Linear Discrimination
  - a quadratic curve $\rightarrow$ Quadratic Discrimination

Distribution of $x \in \mathbb{R}^p$ given $y = \pm 1$

- The distribution of $x \in \mathbb{R}^p$ given $y = \pm 1$ is:
  $\rightarrow N(\mu_{\pm 1}, \sum_{\pm 1})$,
  that is, it is assumed to follow a multivariate normal distribution.

- Multivariate Normal Distribution:
  distribution that extends the normal distribution to multidimensional spaces.
  - $\mu_{\pm 1}$: means of each features, $\sum_{\pm 1}$: matrix of variance-covariance,

$$\sum = \begin{pmatrix} \sigma_{11}^2 & \cdots & \sigma_{1p}^2 \\ \vdots & \ddots & \vdots \\ \sigma_{p1}^2 & \cdots & \sigma_{pp}^2 \end{pmatrix}$$

Probability Density Function of $x \in \mathbb{R}^p$ given $y = \pm 1$

- Probability Density Function of $x \in \mathbb{R}^p$ given $y = \pm 1$:

$$f_{\pm 1}(x) = \frac{1}{\sqrt{(2\pi)^p \left|\sum\right|}} e^{\left\{-\frac{1}{2}(x-\mu_{\pm 1})^T \sum_{\pm 1}^{-1}(x-\mu_{\pm 1})\right\}}$$

Outline

Posterior probability

- We assume that the probabilities of $y = \pm 1$ are known before seeing the covariates x, and we call it 'Prior probability'.

- When the probability function of a random variable X is $f_{\pm 1}(x)$ and the prior probability at that time is $\pi_{\pm 1}$,
  $\rightarrow$ The posterior probability:

$$
\begin{aligned}
P(Y|X) &= \frac{P(X = x|Y = \pm 1)P(Y = \pm 1)}{P(X = x)} \\
&= \frac{x\ \pi_{\pm 1} f_{\pm 1}(x)}{\pi_1 f_1(x) + \pi_{-1} f_{-1}(x)}
\end{aligned}
$$

Minimizing the error probability

- Assuming that:
  1. $f_{\pm 1}$ follows a Gaussian distribution,
  2. The expectation $\mu_{\pm 1}$ and the covariance matrix $\sum_{\pm 1}$ are known
  3. $\pi_{\pm 1}$ is also known

$$\frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_{-1} f_{-1}(x)} \geq \frac{\pi_{-1} f_{-1}(x)}{\pi_1 f_1(x) + \pi_{-1} f_{-1}(x)}$$

- That is, we can minimize the error probability:
  - by estimating $y = 1$, when $\pi_1 f_1(x) \geq \pi_{-1} f_{-1}(x)$.
  - by estimating $y = -1$, when $\pi_1 f_1(x) \leq \pi_{-1} f_{-1}(x)$.

Maximizing the posterior probability

- When estimating $y = \hat{k}$, the probability that the estimate is correct:

$$1 - \sum_{k \neq \hat{k}} P(y = k|x) \;=\; 1 - P(y = \hat{k}|x)$$

- When the prior probability is known, choosing a k that maximizes the posterior probability $P(y = \hat{k}|x)$ as $\hat{k}$ minimizes the average error probability.

The border

- For simplicity, let's assume $K = 2$, and when maximizing the posterior probability,
  we see the properties at the border between $y = \pm 1$.

- Property:

$$-(x-\mu_1)^\top \sum_1^{-1}(x-\mu_1) + (x-\mu_{-1})^\top \sum_{-1}^{-1}(x-\mu_{-1}) \;=\; \log \frac{\left|\sum_1\right|}{\left|\sum_{-1}\right|} - 2\log\frac{\pi_1}{\pi_{-1}}$$

The border

- In general, the border is a function of the quadratic forms $x^\top \sum_1^{-1} x$ and $x^\top \sum_{-1}^{-1} x$ of x, and this case is 'Quadratic Discrimination'.

- In particular, when $\sum_1 = \sum_{-1}$ (if we write them as $\sum$), the border becomes a surface, that we call 'Linear Discrimination'.

The border

- When $x^\top \sum_1^{-1} x = x^\top \sum_{-1}^{-1} x$, the border becomes:

$$2(\mu_1 - \mu_{-1})^\top \sum^{-1} x - (\mu_1^\top \sum^{-1} \mu_1 - \mu_{-1} \sum^{-1} \mu_{-1}) \;=\; -2 \log \frac{\pi_1}{\pi_{-1}}$$

or more simply,

$$(\mu_1 - \mu_{-1})^\top \sum^{-1} (x - \frac{\mu_1 + \mu_{-1}}{2}) \;=\; -\log \frac{\pi_1}{\pi_{-1}}$$

The border

- Proof

$$2(\mu_1-\mu_{-1})^\top \sum^{-1} x - (\mu_1^\top \sum^{-1} \mu_1 - \mu_{-1} \sum^{-1} \mu_{-1}) \;=\; -2\log\frac{\pi_1}{\pi_{-1}}$$

$$\Rightarrow (\mu_1-\mu_{-1})^\top \sum^{-1} x - \frac{1}{2}(\mu_1^\top \sum^{-1} \mu_1 - \mu_{-1} \sum^{-1} \mu_{-1}) = -\log\frac{\pi_1}{\pi_{-1}}$$

$$\Rightarrow (\mu_1-\mu_{-1})^\top \sum^{-1} x - \frac{1}{2}(\mu_1^\top \sum^{-1} \mu_1 + \mu_1^\top \sum^{-1} \mu_{-1} - \mu_{-1}^\top \sum^{-1} \mu_1 - \mu_{-1}^\top \sum^{-1} \mu_{-1}) = -\log\frac{\pi_1}{\pi_{-1}}$$

($\because x^\top \sum_1^{-1} x = x^\top \sum_{-1}^{-1} x$ are canceled.)

$$\Rightarrow (\mu_1-\mu_{-1})^\top \sum^{-1} x - \frac{1}{2}(\mu_1^\top \sum^{-1} - \mu_{-1}^\top \sum^{-1})(\mu_1+\mu_{-1}) = -log\frac{\pi_1}{\pi_{-1}}$$

$$\Rightarrow (\mu_1-\mu_{-1})^\top \sum^{-1} x - \frac{1}{2}(\mu_1^\top - \mu_{-1}^\top) \sum^{-1}(\mu_1+\mu_{-1}) = -log\frac{\pi_1}{\pi_{-1}}$$

$$\Rightarrow (\mu_1-\mu_{-1})^\top \sum^{-1} x - \frac{1}{2}(\mu_1 - \mu_{-1})^\top \sum^{-1}(\mu_1+\mu_{-1}) = -log\frac{\pi_1}{\pi_{-1}}$$

$$\Rightarrow (\mu_1-\mu_{-1})^\top \sum^{-1} \left(x - \frac{\mu_1+\mu_{-1}}{2}\right) = -log\frac{\pi_1}{\pi_{-1}}$$

The border

- If $\pi_1 = \pi_{-1}$, then the border is $x = \frac{\mu_1+\mu_{-1}}{2}$.

- Proof

$$(\mu_1 - \mu_{-1})^\top \sum^{-1} \left(x - \frac{\mu_1 + \mu_{-1}}{2} \right) = 0$$

$$\therefore x = \frac{\mu_1 + \mu_{-1}}{2}$$

- If $\pi_{\pm 1}$ and $f_{\pm 1}$ are unknown, we need to estimate them from the training data.

Outline

[Example 35] Output the border of the Quadratic Discriminant using the R code

- The following code draws the border for estimating the mean and covariance of the covariates x for $y = \pm 1$.

```
mu.1=c(2,2); sigma.1=2; sigma.2=2; rho.1=0
mu.2=c(-3,-3); sigma.3=1; sigma.4=1; rho.2=-0.8

n=100

u=rnorm(n); v=rnorm(n); x.1=sigma.1*u+mu.1[1];
y.1=(rho.1*u+sqrt(1-rho.1^2)*v)*sigma.2+mu.1[2]

u=rnorm(n); v=rnorm(n); x.2=sigma.3*u+mu.2[1];
y.2=(rho.2*u+sqrt(1-rho.2^2)*v)*sigma.4+mu.2[2]
```

[Example 35] Output the border of the Quadratic Discriminant using the R code

```
f=function(x,mu,inv,de){
  drop(-0.5*t(x-mu)%*%inv%*%(x-mu)-0.5*log(de))
}

mu.1=mean(c(x.1,y.1)); mu.2=mean(c(x.2,y.2));

df=data.frame(x.1,y.1); mat=cov(df)
inv.1=solve(mat); de.1=det(mat)

df=data.frame(x.2,y.2); mat=cov(df)
inv.2=solve(mat); de.2=det(mat)
```

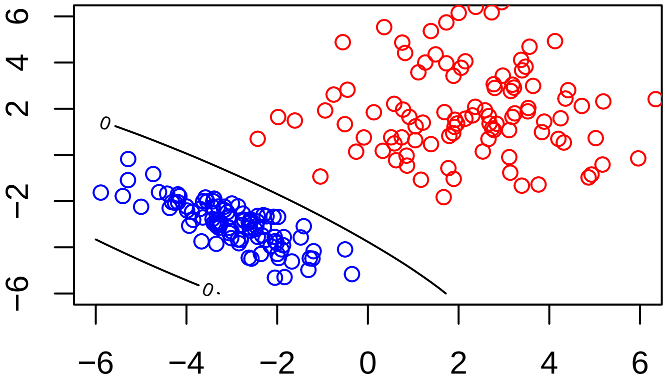
[Example 35] Output the border of the Quadratic Discriminant using the R code

```
f.1=function(u,v)f(c(u,v),mu.1,inv.1,de.1);
f.2=function(u,v)f(c(u,v),mu.2,inv.2,de.2)

pi.1=0.5; pi.2=0.5
u = v = seq(-6, 6, length=50); m=length(u); w=array(dim=c(m,m))
for(i in 1:m){
  for(j in 1:m){
    w[i,j]=log(pi.1)+f.1(u[i],v[j])-log(pi.2)-f.2(u[i],v[j])
  }
}

# plot
contour(u,v,w,level=0)
points(x.1,y.1,col="red"); points(x.2,y.2,col="blue")
```

[Example 35] Output the border of the Quadratic Discriminant using the R code



[Example 35] Output the border of the Linear Discriminant using the R code

- If the covariance matrices are equal, we can use the following code:

```
mu.1=c(2,2); sigma.1=2; sigma.2=2; rho.1=0
mu.2=c(-3,-3); sigma.3=1; sigma.4=1; rho.2=-0.8

n=100

u=rnorm(n); v=rnorm(n); x.1=sigma.1*u+mu.1[1];
y.1=(rho.1*u+sqrt(1-rho.1^2)*v)*sigma.2+mu.1[2]

u=rnorm(n); v=rnorm(n); x.2=sigma.3*u+mu.2[1];
y.2=(rho.2*u+sqrt(1-rho.2^2)*v)*sigma.4+mu.2[2]
```

[Example 35] Output the border of the Quadratic Discriminant using the R code

```
f=function(x,mu,inv,de){
  drop(-0.5*t(x-mu)%*%inv%*%(x-mu)-0.5*log(de))
}

mu.1=mean(c(x.1,y.1)); mu.2=mean(c(x.2,y.2));

df=data.frame(c(x.1,y.1)-mu.1, c(x.2,y.2)-mu.2)

inv.1=solve(mat)
de.1=det(mat)

inv.2=inv.1
de.2=de.1
```

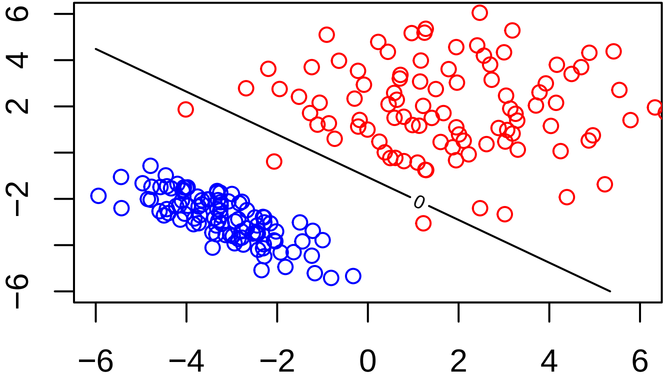
[Example 35] Output the border of the Quadratic Discriminant using the R code

```
f.1=function(u,v)f(c(u,v),mu.1,inv.1,de.1);
f.2=function(u,v)f(c(u,v),mu.2,inv.2,de.2)

pi.1=0.5; pi.2=0.5
u = v = seq(-6, 6, length=50); m=length(u); w=array(dim=c(m,m))
for(i in 1:m){
  for(j in 1:m){
    w[i,j]=log(pi.1)+f.1(u[i],v[j])-log(pi.2)-f.2(u[i],v[j])
  }
}

# plot
contour(u,v,w,level=0)
points(x.1,y.1,col="red"); points(x.2,y.2,col="blue")
```

[Example 35] Output the border of the Linear Discriminant using the R code



[Example 35] Output the Quadratic Discriminant graph using the R code

- Through two border pictures:
- We can see that if covariance matrices are equal, the border is a line, otherwise, it is a quadratic curve.
- In the linear discrimination, if the prior probabilities and the covariance matrices are equal, then the border is the vertical bisector of the line connecting the centers.

[Example 36] Iris data classification: using the classifier via Quadratic Discrimination

- When the response takes more than two values,
  we can choose the response with the maximum posterior probability.

- Fisher's Iris data to use in Ex.36:

- It contains four covariates, and the response variable which is the three species containing 50 samples. (N= 150, p= 4)

- We evaluate it using the test data set that is different from the training data set.

[Example 36] Iris data classification: using the classifier via Quadratic Discrimination

```
f=function(w,mu,inv,de)-0.5*(w-mu)%*%inv%*%t(w-mu)-0.5*log(de)
df=iris; df[[5]]=c(rep(1,50),rep(2,50),rep(3,50))
n=nrow(df); train=sample(1:n,n/2,replace=FALSE); test=setdiff(1:n,train)
mat=as.matrix(df[train,])
mu=list(); covv=list()
for(j in 1:3){
  x=mat[mat[,5]==j,1:4];
  mu[[j]]=c(mean(x[,1]),mean(x[,2]),mean(x[,3]),mean(x[,4]))
  covv[[j]]=cov(x)
}
g=function(v,j)f(v,mu[[j]],solve(covv[[j]]),det(covv[[j]]))
z=array(dim=n/2)
for(i in test){
  u=as.matrix(df[i,1:4]); a=g(u,1);b=g(u,2); c=g(u,3)
  if(a<b){if(b<c)z[i]=3 else z[i]=2}
  else {if(a<c)z[i]=3 else z[i]=1}
}
table(z[test],df[test,5])
```

[Example 36] Iris data classification: using the classifier via Quadratic Discrimination

```
##
##      1  2  3
##   1 29  0  0
##   2  0 19  0
##   3  0  3 24
```

- vertical axis: Values classified using the classifier
- horizontal axis: Values of test data set
- Setosa and Virginica can be seen as being well classified accroding to the test data set.
- On the other hand, in the case of Versicolor,

there are two values that were incorrectly classified as Virginica when analyzed by the QDA classifier.

Q & A

Thank you:)