Multivariate Analysis Lecture 14: More on Classification

Zhaoxia Yu
Professor, Department of Statistics

2023-05-25

Section 1

Outline

Outline

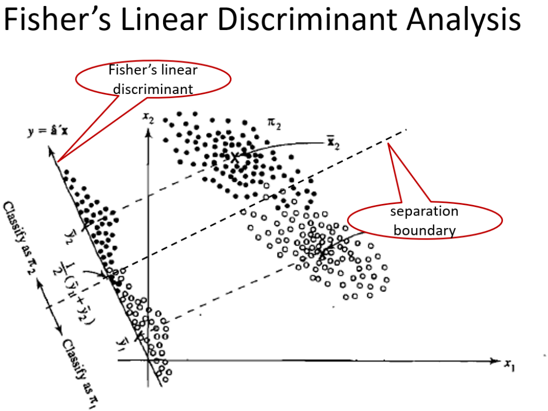
- Review of LDA
- QDA
- Decision theory
    - Equal costs
    - Unequal costs

Section 2

Review of LDA

Linear Discrminant Analysis

```r
knitr::include_graphics("img/FLDA.png")
```



Subsection 1

Two-Class Problems

FLDA: Maximum Separability

- The maximization problem is

$$\underset{a}{\operatorname{argmax}} \frac{a^T(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T a}{a^T \mathbf{\Sigma} a}$$

- Use an argument similar to PCA, such a is the first eigenvector of $\mathbf{\Sigma}^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T$.
- We can show that $a = \mathbf{S}_p^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$.
- The linear function

$$f(x) = a^T x \text{ where } a = \mathbf{S}_p^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$$

is called Fisher's linear discriminant function.

Allocate New Observations

- Consider an observation X_0. We compute

$$f(X_0) = a^T X_0$$

where $a = \mathbf{S}_p^{-1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$

- Let

$$m = a^T \frac{\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2}{2} = (\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)^T \mathbf{S}_p^{-1} \frac{\bar{\mathbf{X}}_1 + \bar{\mathbf{X}}_2}{2}$$

- Allocate X_0 to
  - class 1 if $f(X_0) > m$
  - class 2 if $f(x_0) < m$

Subsection 2

g-Class Problems

Quantify Separation in a g-Class Problem

- Measure separation using F statistic

$$
\begin{aligned}
F(a) &= \frac{MSB}{MSW} = \frac{SSB/(g-1)}{SSW/(n-g)} \\
&= \frac{\sum_{i=1}^{g} n_i(\bar{Y}_{i.}^{(1)} - \bar{Y}_{..}^{(1)})^2/(g-1)}{\sum_{i=1}^{g}(n_i - 1)S^2_{Y_i^{(1)}}/(n-g)} \\
&= \frac{a^T \sum n_i(\bar{X}_{i.} - \bar{X}_{..})(\bar{X}_{i.} - \bar{X}_{..})^T a}{a^T \sum_{i=1}^{g}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_{i.})(X_{ij} - \bar{X}_{i.})^T a} \frac{n-g}{g-1} \\
&= \frac{a^T \mathbf{B} a}{a^T \mathbf{W} a} \frac{n-g}{g-1}
\end{aligned}
$$

where $n = \sum_{i=1}^{g} n_i$, $\mathbf{B}$ is the between-group sample covariance matrix, and $\mathbf{W}$ is the within-group sample covariance matrix.

Linear Discriminants

- The first linear discriminant is the linear function that maximizes $F(a)$. It can also be shown that the first linear discriminant is given by the first eigenvector of $\mathbf{W}^{-1}\mathbf{B}$, i.e.,

$$Y_{ij}^{(1)} = \gamma_1^T X_{ij}$$

  where γ_1 is the first eigenvector of $\mathbf{W}^{-1}\mathbf{B}$.
- Similarly, for $k = 1, \cdots, rank(\mathbf{B})$, the kth linear discriminant is given by the kth eigenvector of $\mathbf{W}^{-1}\mathbf{B}$

$$Y_{ij}^{(k)} = \gamma_k^T X_{ij}$$

Use the Linear Discriminants

- Let X_0 be a new observation. We allocate it to the group with the minimum distance defined by the Euclidean distance in space spanned by the linear discriminants.
- Calculate $Y_0^{(k)} = \gamma_k^T X_0$, the projection of X_0 to the kth linear discriminant for $k = 1, \cdots, rank(B)$.
- Calculate the distance between $(Y_0^{(1)}, \cdots, Y_0^{(rank(B))})$ and $(\bar{Y}_{i.}^{(1)}, \cdots, \bar{Y}_{i.}^{(rank(B))})$

$$D^2(X_0, i) = \sum_{k=1}^{rank(B)} [Y_0^{(k)} - \bar{Y}_{i.}^{(k)}]^2$$

- Allocate X_0 to

$$\operatorname*{argmin}_i D^2(X_0, i)$$

Section 3

QDA

Subsection 1

Why Is QDA Useful?

Motivating Example

- Consider two populations $N(\mu_0 = 0, \sigma_0 = 1)$ and $N(\mu_1 = 0, \sigma_1 = 4)$. We simulated ten points from each distribution

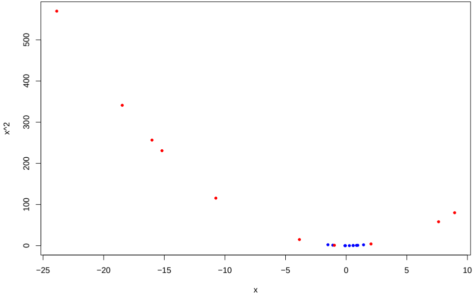
```
set.seed=(2);X0=rnorm(10, 0, 1); X1=rnorm(10, 0, 16)
plot(c(X0, X1), rep(0, 20), axes=FALSE, xlab="", ylab="", type="n", asp = 5, ylim=c(-1,0.1))
points(X0, rep(0,10), pch="0", col="blue")
points(X1, rep(0,10), pch="1", col="red")
```

$N(\mu_0 = 0, \sigma_0 = 1)$ vs $N(\mu_1 = 0, \sigma_1 = 4)$

- Does A Linear Classifier Work
- A linear classifier in 1D is a single value.
- If there was one, we simply classify based on the sign of $x - m_0$ for some m_0.
- Based on the figure in the previous one, there was no such linear classifier
- Let's try to expand the 1D data to 2D by including the squared of each observation

$N(\mu_0 = 0, \sigma_0 = 1)$ vs $N(\mu_1 = 0, \sigma_1 = 4)$

```r
X_org=c(X0, X1)
X_sq=X_org^2
plot(X_org, X_sq, type="n", xlab="x", ylab="x^2")
points(X_org[1:10], X_sq[1:10], pch=20, col="blue")
points(X_org[11:20], X_sq[11:20], pch=20, col="red")
```



Subsection 2

QDA for Two-Class Problems

QDA for Two-Class Problems

- The LDA can be derived using likelihood functions under the assumptions
    1. Multivariate normal
    2. Equal covariance matrix
- The assumption of equal covariance matrix is not always a good approximation to the true covariance matrices
- If we relax this assumption, we will have QDA

QDA for Two-Class Problems

- Let's consider a two-class classification problem with n_1 and n_2 observations in classes 1 and 2, respectively.
- Suppose we have two independent random samples
  - Sample 1: $X_{1j} \overset{iid}{\sim} N(\mu_1, \mathbf{\Sigma}_1)$, where $j = 1, \cdots, n_1$
  - Sample 2: $X_{2j} \overset{iid}{\sim} N(\mu_2, \mathbf{\Sigma}_2)$, where $j = 1, \cdots, n_2$
- Sample mean vectors:

$$\bar{\mathbf{X}}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} X_{1j}, \bar{\mathbf{X}}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} X_{2j}$$

- Remark: the sample mean vectors are the MLE of the corresponding mean vectors

QDA for Two-Class Problems

- MLE of covariance matrices

$$\hat{\boldsymbol{\Sigma}}_1 = \frac{n_1 - 1}{n_1} S_1, \hat{\boldsymbol{\Sigma}}_2 = \frac{n_2 - 1}{n_2} S_2$$

where S_i is the sample covariance matrix for sample i.

- Likelihood functions

$$L_1(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1) \propto |\boldsymbol{\Sigma}_1|^{-1/2} exp\{-\frac{1}{2}(x - \boldsymbol{\mu}_1)^T \boldsymbol{\Sigma}_1^{-1}(x - \boldsymbol{\mu}_1)\}$$
$$L_2(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2) \propto |\boldsymbol{\Sigma}_2|^{-1/2} exp\{-\frac{1}{2}(x - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}_2^{-1}(x - \boldsymbol{\mu}_2)\}$$

QDA for Two-Class Problems

- We can either check whether the ratio is greater than one or check whether the difference of log-likelihood is positive.

$$l_1 - l_2 = -\frac{1}{2} log(\frac{|\mathbf{\Sigma}_1|}{|\mathbf{\Sigma}_2|}) - \frac{1}{2}[(x-\boldsymbol{\mu}_1)^T \mathbf{\Sigma}_1^{-1}(x-\boldsymbol{\mu}_1) - (x-\boldsymbol{\mu}_2)^T \mathbf{\Sigma}_2^{-1}(x-\boldsymbol{\mu}_2)]$$

- The classification boundary is given by $l_1 - l_2 = 0$, i.e.,

$$(x-\boldsymbol{\mu}_1)^T \mathbf{\Sigma}_1^{-1}(x-\boldsymbol{\mu}_1) - (x-\boldsymbol{\mu}_2)^T \mathbf{\Sigma}_2^{-1}(x-\boldsymbol{\mu}_2) = log(\frac{|\mathbf{\Sigma}_2|}{|\mathbf{\Sigma}_1|})$$

- It is quadratic!

QDA for Two-Class Problems

- Replace unknown parameters with estimate, we have the classification rule: allocate x to class 1 if

$$(x - \bar{\mathbf{X}}_1)^T \mathbf{S}_1^{-1}(x - \bar{\mathbf{X}}_1) - (x - \bar{\mathbf{X}}_2)^T \mathbf{S}_2^{-1}(x - \bar{\mathbf{X}}_2) < log(\frac{|\mathbf{S}_2|}{|\mathbf{S}_1|})$$

QDA for g-Class Problems

- For the *i*th group, we compute a quadratic score, which is defined as

$$Q_i(x) = (x - \bar{\mathbf{X}}_i)^T \mathbf{S}_i^{-1}(x - \bar{\mathbf{X}}_i) + log(|\mathbf{S}_i|)$$

- Allocate x to the class with the minimum quadratic score

Motivating Example: $N(\mu_0 = 0, \sigma_0 = 1)$ vs $N(\mu_1 = 0, \sigma_1 = 4)$

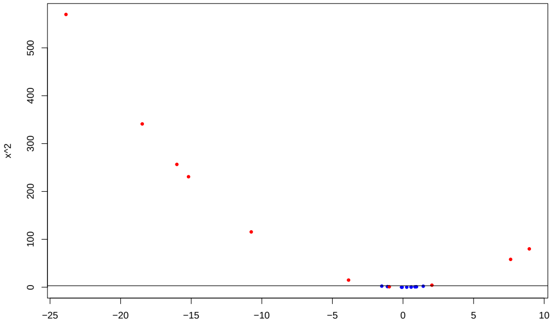
- The "theoretical" rule by QDA has the boundary

$$\frac{15}{16}x^2 = log(16/1) \Rightarrow x^2 = 16log(16)/15$$

- Note: this boundary is quadratic in 1D (the space of the original of measurement), but linear in 2D (the space of (x, x^2)).

Motivating Example: $N(\mu_0 = 0, \sigma_0 = 1)$ vs $N(\mu_1 = 0, \sigma_1 = 4)$

```
plot(X_org, X_sq, type="n", xlab="x", ylab="x^2")
points(X_org[1:10], X_sq[1:10], pch=20, col="blue")
points(X_org[11:20], X_sq[11:20], pch=20, col="red")
abline(h=16*log(16)/15)
```



Subsection 3

Example of QDA

Example of QDA

```
obj.lda=lda(Species~., data = iris)
obj.qda=qda(Species~., data = iris)

table(Pred=predict(obj.lda, iris)$class,
True=iris$Species)
table(Pred=predict(obj.qda, iris)$class,
True=iris$Species)
```

Example of QDA

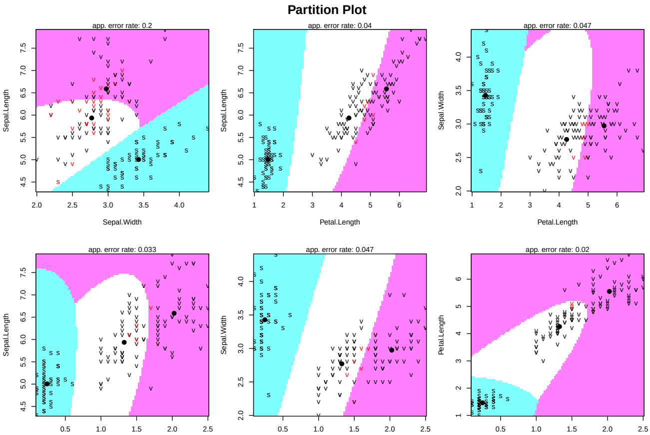
```
##              True
## Pred          setosa versicolor virginica
##   setosa          50          0         0
##   versicolor       0         48         1
##   virginica        0          2        49
```

```
##              True
## Pred          setosa versicolor virginica
##   setosa          50          0         0
##   versicolor       0         48         1
##   virginica        0          2        49
```

- The same result for this particular example

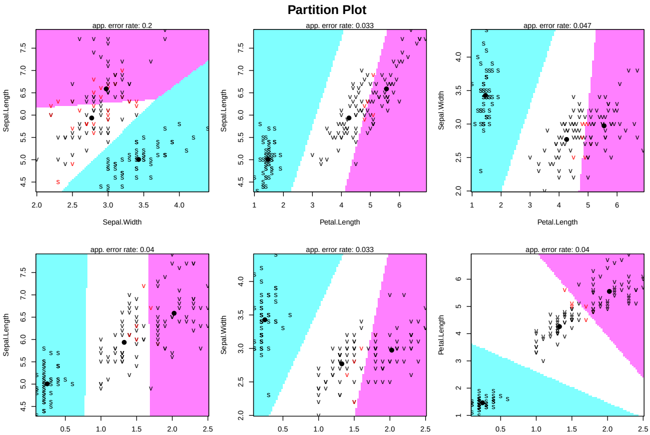
Visualize QDA Results

```
partimat(Species ~ ., data = iris, method = "qda")
```



Visualize LDA Results

```
partimat(Species ~ ., data = iris, method = "lda")
```



Section 4

Decision Theory

Cost and Prior Probabilities

- In practice, different types of errors have different costs
- Prior probabilities are often known but we haven't discussed how to use them
- Goals:
  - When different errors have the same cost, we look for a classification rule that minimizes the probability of misclassification
  - When different errors cost differently, we want to find a classification rule that minimizes the total cost

Section 5

Equal Costs

Subsection 1

Minimize Probability of Misclassification

Minimize Probability of Misclassification

- Notations:
- X: data
- Z: true class. It is binary, i.e., $Z = 1$ or $Z = 0$
- $P(Z = 1) = \pi$: prior probability, known
- $\delta(x)$: decision function / classifier
  - $\delta(x) = 1$: allocate x to group 1
  - $\delta(x) = 0$: allocate x to group 0

Risk and Posterior Risk

- Risk of a classifier δ

$$R(\delta, z) = \Pr[\delta(X) \neq Z | Z = z] = \mathbb{E}_{X|z}[\mathbb{I}_{\delta(X)\neq Z} | Z = z]$$
$$= \begin{cases} \Pr[\delta(X) = 0 | z = 1] & \text{if } z = 1 \\ \Pr[\delta(X) = 1 | z = 0] & \text{if } z = 0 \end{cases}$$

- The posterior risk of δ

$$PR(\delta(x)) = \Pr[\delta(x) \neq Z | x] = \mathbb{E}_{Z|x}[\mathbb{I}_{\delta(X)\neq Z} | X = x]$$
$$= \begin{cases} \Pr[Z = 0 | x] & \text{if } \delta(x) = 1 \\ \Pr[Z = 1 | x] & \text{if } \delta(x) = 0 \end{cases}$$

Bayes Risk

- Bayes risk

$$B(\delta) = \Pr[\delta(X) \neq Z]$$

- Note that

$$B(\delta) = \Pr[\delta(X) \neq Z] = \mathbb{E}_{XZ}[\mathbb{I}_{\delta(X)\neq Z}] = E_X[PR(\delta(X))] = E_Z[R(\delta, Z$$

- Rewrite the Bayes risk

$$\begin{aligned} B(\delta) &= \Pr[\delta(X) \neq Z] \\ &= \Pr[\delta(X) = 1, Z = 0] + \Pr[\delta(X) = 0, Z = 1] \\ &= \Pr[\delta(X) = 1|Z = 0]\Pr[Z = 0] + \Pr[\delta(X) = 0|Z = 1]\Pr[Z = \\ &= \pi \Pr[\delta(X) = 0|Z = 1] + (1 - \pi)\Pr[\delta(X) = 1|Z = 0] \end{aligned}$$

- The above expression is baesd on the fact

$$B(\delta) = \mathbb{E}_{XZ}[\mathbb{I}_{\delta(X)\neq Z}]$$

Bayes Classification Rule

- Want to find δ^* that minimizes $B(\delta)$
- Claim 1: the δ^* that minimizes $PR(\delta(x))$ also minimizes $B(\delta)$
  - This is because $B(\delta) = \mathbb{E}[PR(\delta(X)] \geq \mathbb{E}[PR(\delta^*(X)] = B(\delta^*)]$
- Need to find δ^* that minimizes $PR(\delta(x))$. It can be shown that

$$\delta^*(x) = \begin{cases} 1 & \text{if } \frac{\Pr(Z=1|x)}{\Pr(Z=0|x)} > 1 \\ 0 & \text{if } \frac{\Pr(Z=1|x)}{\Pr(Z=0|x)} < 1 \end{cases}$$

- Skip next slide if you are not interested in the proof

The Classifier that Minimizes Posterior Risk

- Recall that

$$PR(\delta(x) = 0) = \Pr(Z = 1|x), PR(\delta(x) = 1) = \Pr(Z = 0|x)$$

- Therefore, we $\delta^*(x)$ should be 1 if

$$\begin{aligned} PR(\delta(x) = 0) > PR(\delta(x) = 1) &\Leftrightarrow \Pr(Z = 1|x) > \Pr(Z = 0|x) \\ &\Leftrightarrow \frac{\Pr(Z = 1|x)}{\Pr(Z = 0|x)} > 1 \end{aligned}$$

The Bayes Classificatin Rule

- We say $\delta^*(x)$ is the Bayes classification rule

$$\delta^*(x) = \begin{cases} 1 & \text{if } \frac{\Pr(Z=1|x)}{\Pr(Z=0|x)} > 1 \\ 0 & \text{if } \frac{\Pr(Z=1|x)}{\Pr(Z=0|x)} < 1 \end{cases}$$

- Computation

$$\frac{\Pr(Z=1|x)}{\Pr(Z=0|x)} \overset{\text{Bayes' theorem}}{=} \frac{\frac{f(x|z=1)\Pr(Z=1)}{f(x)}}{\frac{f(x|z=0)\Pr(Z=0)}{f(x)}}$$

$$= \frac{f(x|z=1)}{f(x|z=0)} \frac{\pi}{1-\pi}$$

- A short review of Bayes' theorem is on next slide. Feel free to skip if you are very familiar with it already

Bayes' Theorem

- Read this slide if you would like to review Bayes' theorem
- Let A and B be two events.
- Bayes' theorem says

$$\Pr(B|A) = \frac{\Pr(A, B)}{\Pr(A)} = \frac{\Pr(A|B)\Pr(B)}{\Pr(A)}$$

where $\Pr(A, B)$ means the joint probability that both A and B occur. We can use alternative expressions such as $\Pr(A \text{ and } B)$ and $\Pr(A \cap B)$.

Subsection 2

Example 1: Univariate

Example 1: Univariate

- Let's consider a univariate example. Suppose that the population consists for two underlying populations
  - Population 1 with π probability and $N(\mu_1 = 1, \sigma^2 = 0.25)$
  - population 0 with $1 - \pi$ probability and $N(\mu_0 = 0, \sigma^2 = 0.25)$
- Would like to allocate $x = 0.8$
- According to Bayes classification rule, we need to compute

$$
\begin{aligned}
\frac{f(x|z=1)\pi}{f(x|z=0)(1-\pi)} &= \frac{f(x|\mu_1 = 1, \sigma^2)\pi}{f(x|\mu_0 = 0, \sigma^2)(1-\pi)} \\
&= \frac{\frac{1}{\sigma\sqrt{2\pi}} exp\{-\frac{1}{2\sigma^2}(x-1)^2\}}{\frac{1}{\sigma\sqrt{2\pi}} exp\{-\frac{1}{2\sigma^2}(x-0)^2\}} \frac{\pi}{1-\pi} \\
&= exp\{\frac{1}{2\sigma^2}(2x-1)\} \frac{\pi}{1-\pi}
\end{aligned}
$$

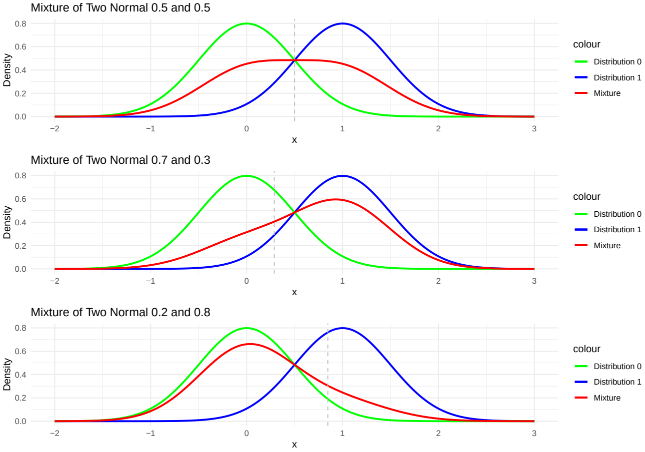
Example 1: Univariate

- The classification boundary is

$$exp\{\frac{1}{2\sigma^2}(2x-1)\} = (1-\pi)/\pi \Leftrightarrow \frac{1}{2\sigma^2}(2x-1) = log((1-\pi)/\pi)$$
$$\Leftrightarrow x = \sigma^2 log((1-\pi)/\pi) + 0.5$$

- The boundary is linear!
    - If $\pi = 0.5$, the boundary is $x = 0.5$, we classify $x = 0.8$ to class 1.
    - If $\pi = 0.7$, the boundary is $x = 0.288$, we classify $x = 0.8$ to class 1.
    - If $\pi = 0.2$, the bondary is $x = 0.846$, we classify $x = 0.8$ to class 0.

Example 1: Density and Classification Boundary



Subsection 3

Example 2: Multivariate

Bayes' Classification under Equal Covariance

- For a two-class problem, the classification boundary by Bayes' classification rule is
$$\frac{f(x|z=1)\pi}{f(x|z=0)(1-\pi)} = 1 \Leftrightarrow log(\frac{f(x|z=1)}{f(x|z=0)}) = log(\frac{1-\pi}{\pi})$$
- Suppose the two underlying distributions are $N(\boldsymbol{\mu}_1, \boldsymbol{\Sigma})$ and $N(\boldsymbol{\mu}_2, \boldsymbol{\Sigma})$.
- The boundary is
$$-\frac{1}{2}(x-\boldsymbol{\mu}_1)^T\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}_1)+\frac{1}{2}(x-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}(x-\boldsymbol{\mu}_0) = log(\frac{1-\pi}{\pi})$$
which is equivalent to

$$(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}x = (\boldsymbol{\mu}_1-\boldsymbol{\mu}_0)^T\boldsymbol{\Sigma}^{-1}\frac{\boldsymbol{\mu}_1+\boldsymbol{\mu}_0}{2} + log(\frac{1-\pi}{\pi})$$

Bayes' Classification under Equal Covariance

- In practice, we substitute the unknown parameters by their estimates

$$(\bar{\mathbf{X}}_{1.} - \bar{\mathbf{X}}_{0.})^T \mathbf{\Sigma}^{-1} x = (\bar{\mathbf{X}}_{1.} - \bar{\mathbf{X}}_{0.})^T \mathbf{\Sigma}^{-1} \frac{\bar{\mathbf{X}}_{1.} + \bar{\mathbf{X}}_{0.}}{2} + log(\frac{1-\pi}{\pi})$$

- Recall that in LDA the linear boundary is

$$a^T x = a^T \frac{\bar{\mathbf{X}}_{1.} + \bar{\mathbf{X}}_{0.}}{2}$$

  Therefore, Bayes' classification is the same as the LDA when $\pi = 1/2$.
- Similarly, in a g-class problem, LDA is the same as Bayes classification under the assumptions (1) multivariate normality, (2) equal covariance, and (3) uniform prior probabilities.

Connection with Logistic Regression

- A logistic regression can be used for a two-class problem. Note: both can be extended to model multi-class problem.
- It models the log-odds, which is defined as

$$\frac{\Pr(Z = 1|x)}{\Pr(Z = 0|x)}$$

  This is the ratio of posterior risks.
- More specifically, it models the log-odds as a linear function of the covariates.
- The LDA under the Bayes rule computes the ratio of the posterior risk. The decision function is also based on a linear function of the covariates.
- Therefore we see a connection between them.

Differencees between LDA and Logistic Regression

- The two approaches were derived from different models with different assumptions.
- Logistic regression models models the randomness of a binary variable
- Although we started LDA without assuming multivariate normality, LDA can be derived from the likelihood principle by modeling a random vector using a multivariate normal
- When the assumptions are met, LDA might have better performance, according to "The efficiency of logistic regression compared to normal discriminant analysis, J Am Stat Assoc, 70, 892-898 (1975)."
- The practical differences is often small
- Logistic regression provides an estimated probability; LDA only provides predicted labels.

Subsection 4

Example 3: Univariate, Unequal Variance

Example 3: Univariate, Unequal Variance

- Again, consider a univariate example. This time we relax the assumption of equal variance
- Suppose that the population consists for two underlying populations
  - Population 1 with π probability and $N(\mu_1 = 0, \sigma_1^2 = 4^2)$
  - population 0 with $1 - \pi$ probability and $N(\mu_0 = 0, \sigma_0^2 = 1^1)$
- Using Bayes' classification, we allocate x to population 1 if $\frac{f(x|z=1)\pi}{f(x|z=0)(1-\pi)} > 1$, i.e., the boundary is

Example 3:

- The Bayes' classification boundary is

$$
\begin{aligned}
1 &= \frac{f(x|\mu_1 = 0, \sigma_1^2)\pi}{f(x|\mu_0 = 0, \sigma_0^2)(1-\pi)} \\
&= \frac{\frac{1}{\sigma_1\sqrt{2\pi}} exp\{-\frac{1}{2\sigma_1^2}(x-0)^2\}}{\frac{1}{\sigma_0\sqrt{2\pi}} exp\{-\frac{1}{2\sigma_0^2}(x-0)^2\}} \frac{\pi}{1-\pi} \\
&= exp\{(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2})x^2\}\frac{\pi}{1-\pi}\frac{\sigma_0}{\sigma_1} \\
&= exp\{\frac{1}{2}\frac{15}{16}x^2\}\frac{\pi}{1-\pi}\frac{1}{4}
\end{aligned}
$$

which is equivalent to

$$
x^2 = \frac{16}{15} log(16\frac{1-\pi}{\pi})
$$

Subsection 5

Example 4: Univariate, Unequal Variance

Example 4: Univariate, Unequal Variance

- Suppose that the population consists for two underlying populations
    - Population 1 with π probability and $N(\mu_1 = 1, \sigma_1^2 = 0.25)$
    - population 0 with $1 - \pi$ probability and $N(\mu_0 = 0, \sigma_2^2 = 1)$
- Would like to allocate $x = 0.8$

Example 4: Univariate, Unequal Variance

- According to Bayes classification rule, we need to compute

$$\frac{f(x|z=1)\pi}{f(x|z=0)(1-\pi)} = \frac{f(x|\mu_1=1,\sigma_1^2)\pi}{f(x|\mu_0=0,\sigma_0^2)(1-\pi)}$$

$$= \frac{\frac{1}{\sigma_1\sqrt{2\pi}} exp\{-\frac{1}{2\sigma_1^2}(x-1)^2\}}{\frac{1}{\sigma_0\sqrt{2\pi}} exp\{-\frac{1}{2\sigma_0^2}(x-0)^2\}} \frac{\pi}{1-\pi}$$

$$= exp\{(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2})x^2 + \frac{x}{\sigma_1^2} - \frac{1}{2\sigma_1^2}\} \frac{\pi}{1-\pi} \frac{\sigma_0}{\sigma_1}$$

- The classification boundary is

$$(\frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_1^2})x^2 + \frac{x}{\sigma_1^2} - \frac{1}{2\sigma_1^2} = log[\frac{1-\pi}{\pi}\frac{\sigma_1}{\sigma_0}]$$

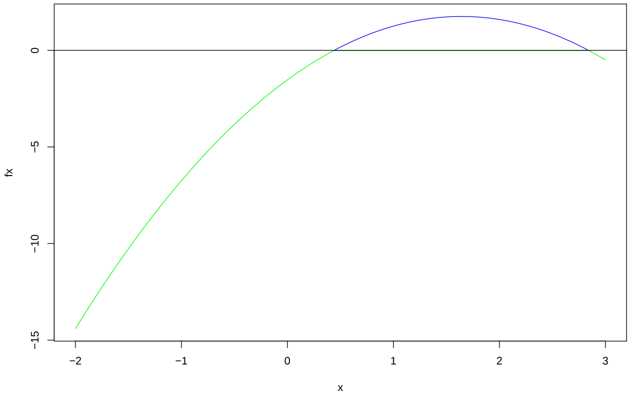
- It is quadratic!

Example 4: Univariate, Unequal Variance

```
mean1 <- 1
var1 <- 0.25
mean0 <- 0
var0 <- 0.64
weight1 <- 0.5
weight0 <- 1 - weight1

# doesn't seem to be correct
x <- seq(-2, 3, length.out = 1000)
fx=(1/var0 - 1/var1)/2*x^2 + x/var1 - 1/2/var1 - log(weight0*sqrt(var1)/weight1/sqrt(var0))
plot(x, fx, type="n")
lines(x[fx>0], fx[fx>0], col="blue")
lines(x[fx<0], fx[fx<0], col="green")
abline(h=0)
```

Example 4: Univariate, Unequal Variance



Section 6

Unequal Costs

Subsection 1

Risk and Cost

Risk and Cost

- Different types of misclassifications might cost differently
- Let $L(\delta(x), z)$ denote the cost function
- Let $C(1|0) = L(1, 0)$, the cost of misclassifying 0 to 1
- Let $C(0|1) = L(0, 1)$, the cost of misclassifying 1 to 0
- The Risk and Bayes risk need to be revised accordingly

Risk and Posterior Risk

- Risk of a classifier δ

$$R(\delta, z) = \mathbb{E}_{X|Z=z}[L(\delta(X), z)]$$
$$= \begin{cases} C(0|1)\Pr[\delta(X) = 0|z = 1] & \text{if } z = 1 \\ C(1|0)\Pr[\delta(X) = 1|z = 0] & \text{if } z = 0 \end{cases}$$

- The posterior risk of δ

$$PR(\delta(x)) = \mathbb{E}_{Z|x}[L(\delta(x), Z)]$$
$$= \begin{cases} C(1|0)\Pr[Z = 0|x] & \text{if } \delta(x) = 1 \\ C(0|1)\Pr[Z = 1|x] & \text{if } \delta(x) = 0 \end{cases}$$

Bayes Risk

- Bayes risk

$$B(\delta) = \mathbb{E}_{XZ}[L(\delta(X), Z)]$$

- Rewrite the Bayes risk

$$\begin{aligned}
B(\delta) &= \mathbb{E}_{XZ}[L(\delta(X), Z)] \\
&= L(\delta(X) = 1, Z = 0) \Pr[\delta(X) = 1, Z = 0] + L(\delta(X) = 0, Z = 1)[\delta(X) = 0, Z = 1] \\
&= C(1|0) \Pr[\delta(X) = 1, Z = 0] + C(0|1) \Pr[\delta(X) = 0, Z = 1] \\
&= C(1|0) \Pr[\delta(X) = 1|Z = 0] \Pr[Z = 0] + C(0|1) \Pr[\delta(X) = 0|Z = 1] \Pr[Z = 1] \\
&= C(0|1)\pi \Pr[\delta(X) = 0|Z = 1] + (1 - \pi)C(1|0) \Pr[\delta(X) = 1|Z = 0]
\end{aligned}$$

Bayes Classification Rule with Unequal Costs

- Use a derivation similar to the equal cost situation, we can show that the Bayes classification rule is

$$
\begin{aligned}
& PR(\delta(x) = 0) > PR(\delta(x) = 1) \\
\Leftrightarrow\ & C(0|1)\Pr(Z = 1|x) > C(1|0)\Pr(Z = 0|x) \\
\Leftrightarrow\ & \frac{\Pr(Z = 1|x)}{\Pr(Z = 0|x)} > \frac{C(1|0)}{C(0|1)} \\
\Leftrightarrow\ & \frac{f(x|z = 1)}{f(x|z = 0)} > \frac{C(1|0)}{C(0|1)} \frac{1 - \pi}{\pi}
\end{aligned}
$$

Other Related Topics

- There are numerous issues/methods / models
- Training error vs testing error
- Model / variable selection / shrinkage
- Classification tree. Random forest
- Support vector machine
- Neural network and deep neural network