Classification

Observe $(g_1, \underset{\sim}{x}_1), \cdots, (g_n, \underset{\sim}{x}_n)$

g_i takes values in $\{1, \cdots, k\}$

<u>Model</u>: Dist'n of $(G, \underset{\sim}{X})$: $P(G=j) = \lambda_j$

Conditional density of $\underset{\sim}{X}$ given $G=j$. $f_j(\underset{\sim}{x})$

Marginal density of $\underset{\sim}{X}$ is $f(\underset{\sim}{x}) = \lambda_1 f_1(\underset{\sim}{x}) + \lambda_2 f_2(\underset{\sim}{x}) + \cdots + \lambda_k f_k(\underset{\sim}{x})$

Optimal classification: 2 approaches

① Minimize $P(\text{error})$

Given $\underset{\sim}{x}$, clarify as $\hat{G}(\underset{\sim}{x}) = j$ if $\lambda_j f_j(\underset{\sim}{x}) > \lambda_i f_i(\underset{\sim}{x})$ for all $i \neq j$

i.e. $R_j = \{\underset{\sim}{x} : \lambda_j f_j(\underset{\sim}{x}) > \lambda_i f_i(\underset{\sim}{x})$ for all $i \neq j$

② Look at conditional dist'n of G given $\underset{\sim}{X} = \underset{\sim}{x}$

Bayes Thm: $P(G=j \mid \underset{\sim}{X} = \underset{\sim}{x}) \text{ "=" } \dfrac{P(G=j, \underset{\sim}{X} = \underset{\sim}{x})}{P(\underset{\sim}{X} = \underset{\sim}{x})}$

$$= \frac{P(G=j) P(\underset{\sim}{X} = \underset{\sim}{x} \mid G=j)}{P(\underset{\sim}{X} = \underset{\sim}{x})}$$

$$= \frac{\lambda_j f_j(\underset{\sim}{x})}{\lambda_1 f_1(\underset{\sim}{x}) + \cdots + \lambda_k f_k(\underset{\sim}{x})}$$

$\Rightarrow$ if we choose $\hat{G}(\underset{\sim}{x})$ to maximize $P(G=j \mid \underset{\sim}{X} = \underset{\sim}{x})$

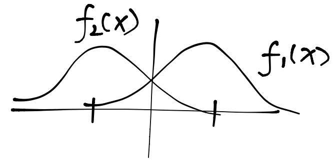
we get the classification rule in ①

But $P(G=j \mid \underset{\sim}{X} = \underset{\sim}{x})$ provides more information.

<u>Example</u> $k=2$, $p=1$, $\lambda_1 = \lambda_2 = \frac{1}{2}$

$f_1(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x-1)^2)$ $(N(1,1))$

$f_2(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}(x+1)^2)$ $(N(-1,1))$



① gives $\hat{G}(x) = 1$ if $x > 0$

$\hat{G}(x) = 2$ if $x < 0$

② $P(G=1 \mid X=x) = \dfrac{f_1(x)}{f_1(x) + f_2(x)}$

x	$P(G=1 \mid X=x)$
-1	0.119
-0.5	0.269
-0.25	0.378
0	0.500
0.25	0.622
0.5	0.731
1	0.881

Linear discriminant analysis (LDA)

$f_1(\underset{\sim}{x}), \cdots, f_k(\underset{\sim}{x})$ multivariate normal with distinct means $\underset{\sim}{\mu}_1, \cdots, \underset{\sim}{\mu}_k$ and common covariance matrix C.

If everything is known then $\hat{G}(\underset{\sim}{x}) = j$ if $d_j(\underset{\sim}{x}) > d_i(\underset{\sim}{x})$ for all $i \neq j$ where

$d_j(\underset{\sim}{x}) = \underset{\sim}{x}^T C^{-1} \underset{\sim}{\mu}_j - \frac{1}{2} \underset{\sim}{\mu}_j^T C^{-1} \underset{\sim}{\mu}_j + \ln(\lambda_j)$

Given data $(g_1, \underset{\sim}{x}_1), \dots, (g_n, \underset{\sim}{x}_n)$ we have $\hat{\underset{\sim}{\mu}}_j = \dfrac{\sum_{i=1}^n \underset{\sim}{x}_i I(g_i = j)}{\sum_{i=1}^n I(g_i = j)}$ $(j = 1, \dots, k)$

$$\hat{C} = \frac{1}{n-k} \sum_{i=1}^n (\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_j)(\underset{\sim}{x}_i - \hat{\underset{\sim}{\mu}}_j)^T$$

<u>Example</u>: Fisher's iris data (see Blackboard)

3 species of irises $\left\{ \begin{array}{l} \text{virginica} \\ \text{setora} \\ \text{versciolor} \end{array} \right.$

4 variables $\left\{ \begin{array}{l} \text{sepal length} \\ \text{sepal width} \\ \text{petal length} \\ \text{petal width} \end{array} \right.$

- LDA works very well here.
- But, could prob do as well from pairwise scatterplots!

How to estimate misclassification rate?
(as well as posterior dist'n)
- resubstitution estimate is typically biased downwards.
- bias increases often severely, as model complexity increases.

<u>Solution</u>: Do some sort of cross-validation
- divide data into 2 sets: training set ($n-m$ obs) and test set (m obs)
- estimate classification rate based on training data and use test data to estimate misclassification rate.

For example, 10-fold cross-validation
- divide data into 10 sets
- successively leave out one set $\rightarrow$ test data
  use other 9 as training data

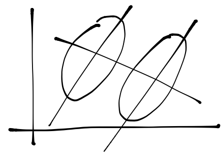
- leave-one-out CV: for each "observation" $\underset{\sim}{x}_i$,
  we compare $\hat{G}_{-i}(\underset{\sim}{x}_i)$ where $\hat{G}_{-i}$ is the classification rate using all data except $\underset{\sim}{x}_i$.
  $\Rightarrow$ used in R function: lda, qda

For the iris data: est'd misclassification rate $= \frac{3}{150}$ (assume $\lambda_1 = \lambda_2 = \lambda_3 = \frac{1}{3}$)
$= 2\%$

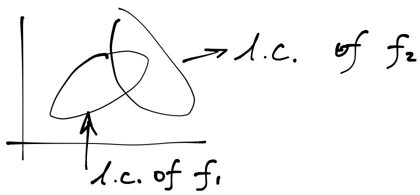
$$\text{Est'd misclassification rate} = \frac{\sum_{i=1}^n \lambda_{g_i} I(\hat{G}_{-i}(\underset{\sim}{x}_i) \neq g_i)}{\sum_{i=1}^n \lambda_{g_i}}$$

Can also estimate $P(G=j \mid \underset{\sim}{x}_i)$ use LOO CV $\nearrow$ 1.-0.-0.
- assume model is approx correct.

- ## Quadratic Discriminant Analysis (QDA)



$k=2,\ p=2,\ C_1=C_2=C$
level curves of $f_1(\underset{\sim}{x})$ & $f_2(\underset{\sim}{x})$ have same orientation



$\Rightarrow$ no linear boundaries for classification rule.

- general form of classification rule is the same but regions more complicated