Classification II

Linear Discriminant Analysis (LDA)

Classify an observation into one of K classes

Data: $\mathbf{X}$ = continuous inputs
Y = categorical output

$\pi_k =$ $P(Y = k)$ prior probability that a randomly chosen observation comes from class k

$f_k(x) =$ probability density function for $(X|Y = k)$

Comparison with logistic regression

- When classes are well separated, parameter estimates in logistic regression are unstable. Not a problem in LDA.
- Small n and approximately multinormally distributed $\mathbf{X}$, LDA more stable than logistic regression.
- A natural approach when Y has many classes (albeit an extension to logistic regression is available)

Bayes theorem

Starting with the definition of conditional probability

$$P(Y = k|X = x) \equiv \frac{P(X = k, X = x)}{P(X = x)}$$

For discrete $X-$variables:

$$P(Y = k|X = x) = \frac{P(X = x|Y = k)P(Y = k)}{P(X = x)}$$

$$\begin{aligned} P(X = x) &= \sum_{k=1}^{K} P(X = x, Y = k) \\ &= \sum_{k=1}^{K} P(X = x|Y = k)P(Y = k) \end{aligned}$$

For discrete $X-$variables:

$$P(Y = k|X = x) = \frac{P(X = x|Y = k)P(Y = k)}{\sum_{k=1}^{K} P(X = x|Y = k)P(Y = k)}$$

Possible to show that for continuous X:

$$p_k(x) \equiv P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{k=1}^{K} \pi_k f_k(x)}$$

$p_k(x_i)$ is the <u>posterior probability</u> that observation i belongs to class k, given that $X_i = x_i$.

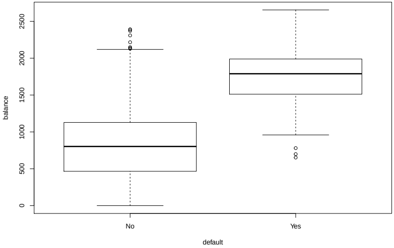
<u>Bayes classifier:</u> Predict, "allocate", observation nr i to the class k with highest $p_k(x_i)$.

Default data

```r
require(ISLR)
```

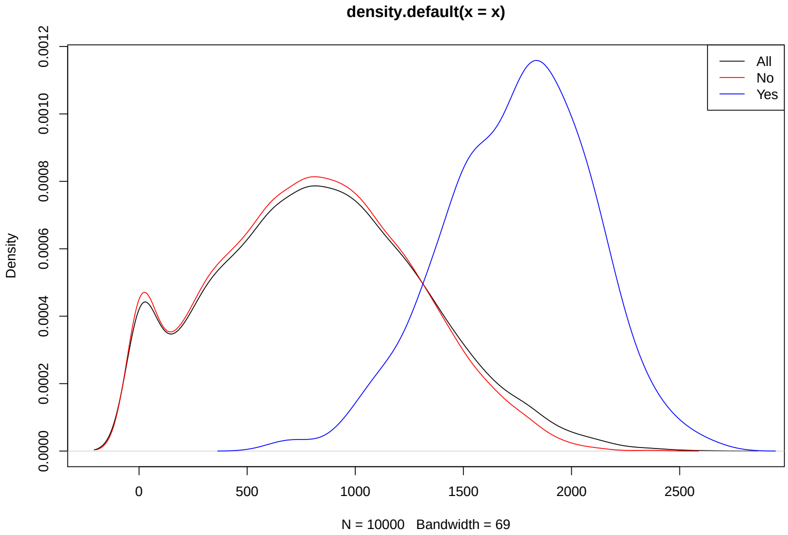
```
## Loading required package: ISLR
```

```r
plot(balance~default,data=Default)
```



Default observations have a higher average account balance

Default data



Default data

```r
by(Default$balance,Default$default,mean)
```

```
## Default$default: No
## [1] 803.9438
## ------------------------------------------------------------
## Default$default: Yes
## [1] 1747.822
```

A very simple example

$p = 1$ (one X-variable) and $K = 2$ (two classes)

We assume $(X|Y = k) \in N(\mu_k, \sigma^2)$, $k = 1, 2$

Prior probabilities $P(Y = k) = \pi_k$, $k = 1, 2$

Densities: $f_k(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}$, $k = 1, 2$

Reminder:

$$p_k(x) \equiv P(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{k=1}^{K} \pi_k f_k(x)}$$

$$p_k(x) \frac{e^{-\frac{(x-\mu_k)^2}{2\sigma^2}}}{e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} + e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}}$$

Allocate i to class 1 if $p_1(x_i) > p_2(x_i)$

Solving the simple example

$$p_1(x) > p_2(x)$$
$$\iff$$
$$e^{-\frac{(x-\mu_1)^2}{2\sigma^2}} > e^{-\frac{(x-\mu_2)^2}{2\sigma^2}}$$
$$\iff$$
$$(x-\mu_1)^2 < (x-\mu_2)^2$$

Allocate i to class 1 if distance to μ_1 shorter than distance to μ_2.
Make sense?

Find Bayes classifier

$$p_k(x) = \frac{\pi_k f_k(x)}{\underbrace{\sum_{k=1}^{K} \pi_k f_k(x)}_{\text{same for all } k}}$$

$$\text{Largest } p_k(x) \iff \text{Largest } \pi_k f_k(x) \iff \text{Largest } (\ln \pi_k + \ln f_k(x))$$

$$\begin{aligned}
\ln \pi_k + \ln f_k(x) &= \ln \pi_k + \ln \left[\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu_k)^2}{2\sigma^2}} \right] \\
&= \ln \pi_k + \frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(x^2 - 2\mu_k x + \mu_k^2) \\
&= \ln \pi_k + \underbrace{\frac{1}{2} \ln(2\pi\sigma^2) - \frac{x^2}{2\sigma^2}}_{\text{constant over } k \text{ for given } x} + \frac{\mu_k x}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2}
\end{aligned}$$

We use the discriminant function

$$\delta_k(x) = \ln \pi_k + \frac{\mu_k x}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2}$$

i.e. allocate *i* to the class with largest $\delta_k(x_i)$

Becomes *quadratic* function if the variance is not equal for the two classes. Quadratic Discriminant Analysis (QDA)

Empirical implementation - unknown parameters

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i: y_i = k} x_i$$

$$\hat{\sigma}_k^2 = \frac{1}{n - K} \sum_{k=1}^{K} \sum_{i: y_i = k} (x_i - \hat{\mu}_k)^2$$

$$\hat{\pi}_k = \frac{n_k}{n} = \frac{\text{number of obs in class k}}{\text{number of obs}}$$

Implementing on Default data

```
nk=table(cl)
n=length(cl)
pi=nk/n
mu=as.matrix(by(x,cl,mean))
s2=(1/(n-2))*(sum((x[cl==1]-mu[1])^2)+sum((x[cl==2]-mu[2])^2))
mu
```

```
##          [,1]
## 1  803.9438
## 2 1747.8217
```

```
s2
```

```
## [1] 205318.6
```

```
pi
```

```
## cl
##      1      2
## 0.9667 0.0333
```

```r
delta=function(x,k,mu,s2,pi) mu[k]*x/s2-mu[k]^2/(2*s2)+log(pi[k])
delta1=delta(x,1,mu,s2,pi)
delta2=delta(x,2,mu,s2,pi)
pred=matrix(1,n,1)
pred[delta2>delta1]=2
head(cbind(cl,pred))
```

```
##      cl
## [1,]  1 1
## [2,]  1 1
## [3,]  1 1
## [4,]  1 1
## [5,]  1 1
## [6,]  1 1
```

```r
table(pred,cl)
```

```
##     cl
## pred    1    2
##    1 9643  257
##    2   24   76
```

p>1, more than one X-variable

Multvariate normally distributed $\boldsymbol{X} \in N_p(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|} e^{\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

Similar to the univariate case it is possible to show that

$$\delta_k(x) = \ln \pi_k + \boldsymbol{x}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k - \frac{1}{2}\boldsymbol{\mu}_k'\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_k$$

is the discriminant function in the p-variate case if $\boldsymbol{X}$ have the same variance-covariance matrix for all groups.

Confusion matrix

	Predicted class	
True class	True Neg (TN)	False Pos (FP)
	False Neg (FN)	True Pos (TP)

```
##     pred
## cl      1    2
##    1 9643   24
##    2  257   76
```

True positive rate $= \frac{\text{TP}}{\text{FN+TP}} = \frac{76}{257+76} = 0.23$

False positive rate $= \frac{\text{FP}}{\text{TN+FP}} = \frac{24}{9643+24} = 0.0025$

Threshold - value

- We choose the class maximizing the posterior probability of belonging to a class. If there are only two classes, a natural threshold is 0.5; but we could use others.

- Default example: Classify an observation i as $\mathrm{default}$ if $p_2(x_i) > 0.5$; or if it is > 0.2. Try several alternatives to evaluate a model, e.g. LDA or logistic regression.

ROC-curve

Plot True Positive Rate against False Positive Rate for many different thresholds.

```r
require(MASS)
```

```
## Loading required package: MASS
```

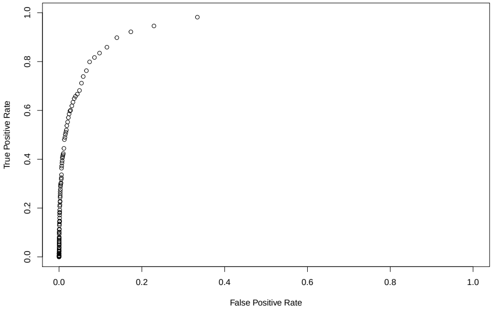
```r
lda1=lda(cl~x)
pr=predict(lda1)$posterior
head(pr)
```

```
##           1           2
## 1 0.9972130 0.002786981
## 2 0.9958358 0.004164240
## 3 0.9865931 0.013406929
## 4 0.9988882 0.001111757
## 5 0.9963955 0.003604464
## 6 0.9933487 0.006651334
```

ROC-curve

```
thrange=seq(0.1,0.9,0.1)
nth=length(thrange)
roc=data.frame(FPrate=0,TPrate=0)
ii=1
for(th in thrange)
{
  pred=pr[,2]>th
  cm=table(cl,pred)
  fprate=cm[1,2]/sum(cl==1)
  tprate=cm[2,2]/sum(cl==2)
  roc[ii,]=cbind(fprate,tprate)
  ii=ii+1
}
plot(roc,xlim=c(0,1),ylim=c(0,1))
```

ROC-curve



$\mathrm{AUC} =$Area Under Curve gives a summary measure of the model's performance.

K-nearest neighbors classifier

x_0 is a *prediction point* (a test observation)

1. Find the K training-points that are closest to x_0, call this set $\mathbb{N}_0$.
2. $P(Y = k|X = x_0) = \frac{1}{K}\sum_{x_i \in \mathbb{N}_0} I(y_i = k)$ where I is the indicator function.