Classification

Jing Qin

April-May/2022

Bull-data (again)

Breed	SalePr	YrHgt	FtFrBody	PrctFFB	Frame	BkFat	SaleHt	SaleWt
5	1300	48.7	1056	72.9	5	0.15	52.6	1525
1	1525	49.4	959	68.4	6	0.15	52.6	1565
1	1525	49.6	1083	75.8	6	0.30	54.6	1640
8	1850	53.1	964	70.8	8	0.10	55.5	1535
1	1500	49.5	963	69.4	6	0.35	53.1	1670
8	1825	53.0	1055	76.8	8	0.10	56.7	1526
5	1375	51.0	1002	72.1	7	0.25	51.9	1410
1	1400	47.6	974	69.7	5	0.15	51.9	1570
1	2250	51.9	1108	72.1	7	0.25	55.3	1575
8	2000	53.5	1175	74.5	8	0.10	57.4	1686
8	1725	51.4	1034	71.2	7	0.10	56.0	1655

Discrimination and classfication:
same **rules** but *different* purposes

Discrimination (separation): to describe, graphically or algebraically, the differential features of objects from several known collections (populations). We try to find 'discriminants' whose numerical values are such that the collections are separated as much as possible.

Breed	SalePr	YrHgt	FtFrBody	PrctFFB	Frame	BkFat	SaleHt	SaleWt
5	1300	48.7	1056	72.9	5	0.15	52.6	1525
1	1525	49.4	959	68.4	6	0.15	52.6	1565
1	1525	49.6	1083	75.8	6	0.30	54.6	1640
8	1850	53.1	964	70.8	8	0.10	55.5	1535
1	1500	49.5	963	69.4	6	0.35	53.1	1670
8	1825	53.0	1055	76.8	8	0.10	56.7	1526
5	1375	51.0	1002	72.1	7	0.25	51.9	1410
1	1400	47.6	974	69.7	5	0.15	51.9	1570
1	2250	51.9	1108	72.1	7	0.25	55.3	1575
8	2000	53.5	1175	74.5	8	0.10	57.4	1686
8	1725	51.4	1034	71.2	7	0.10	56.0	1655

Discrimination and classfication:

same rules but *different* purposes

Classification (allocation): to sort objects into two or more labeled classes. The emphasis is on deriving a rule that can be used to optimally assign new objects to the labeled classes.

?Breed

SalePr	YrHgt	FtFrBody	PrctFFB	Frame	BkFat	SaleHt	SaleWt
1300	48.7	1056	72.9	5	0.15	52.6	1525
1525	49.4	959	68.4	6	0.15	52.6	1565
1525	49.6	1083	75.8	6	0.30	54.6	1640
1850	53.1	964	70.8	8	0.10	55.5	1535
1500	49.5	963	69.4	6	0.35	53.1	1670
1825	53.0	1055	76.8	8	0.10	56.7	1526
1375	51.0	1002	72.1	7	0.25	51.9	1410
1400	47.6	974	69.7	5	0.15	51.9	1570
2250	51.9	1108	72.1	7	0.25	55.3	1575
2000	53.5	1175	74.5	8	0.10	57.4	1686
1725	51.4	1034	71.2	7	0.10	56.0	1655

We start with a bit easier setting: only two populations

Hemophilia A data set (Example 11.3)

Example: detection of hemophilia A carriers

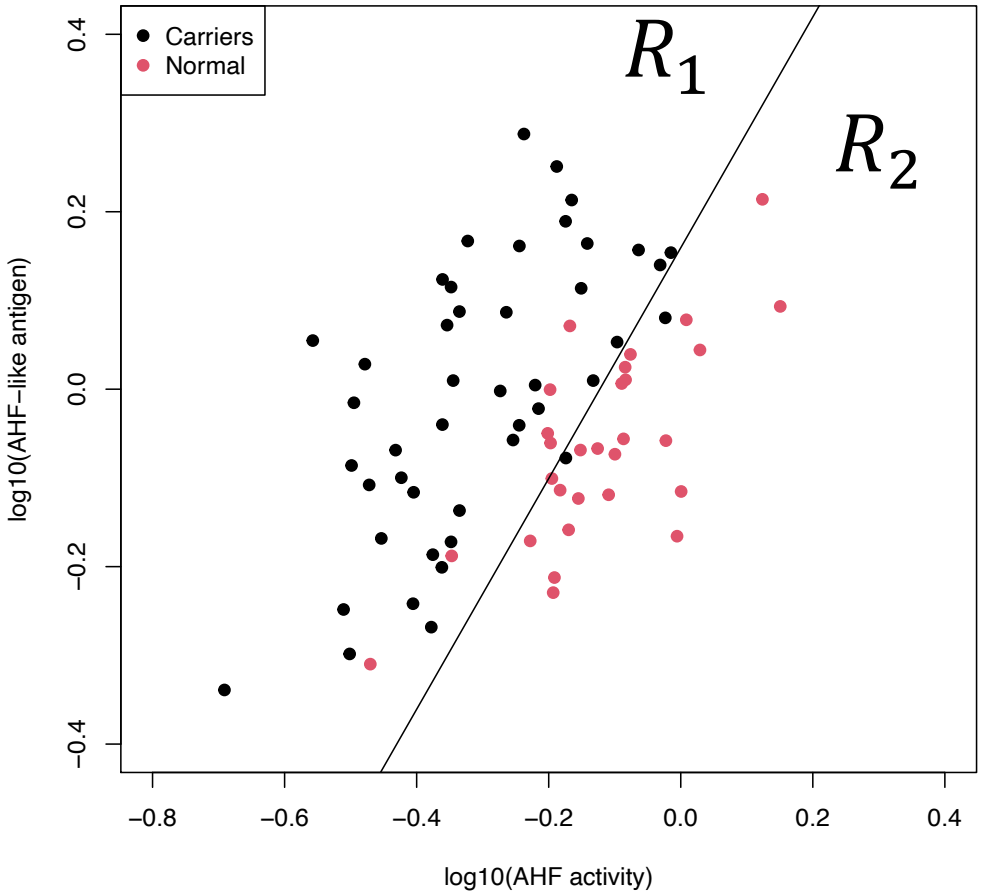
Classify people as normal, i.e. not carrying the hemophilia gene, or as obligatory carrier on the basis of the following blood sample measurements

$$X_1 = \log(\text{AHF activity}),$$

$$X_2 = \log(\text{AHF-like antigen}).$$

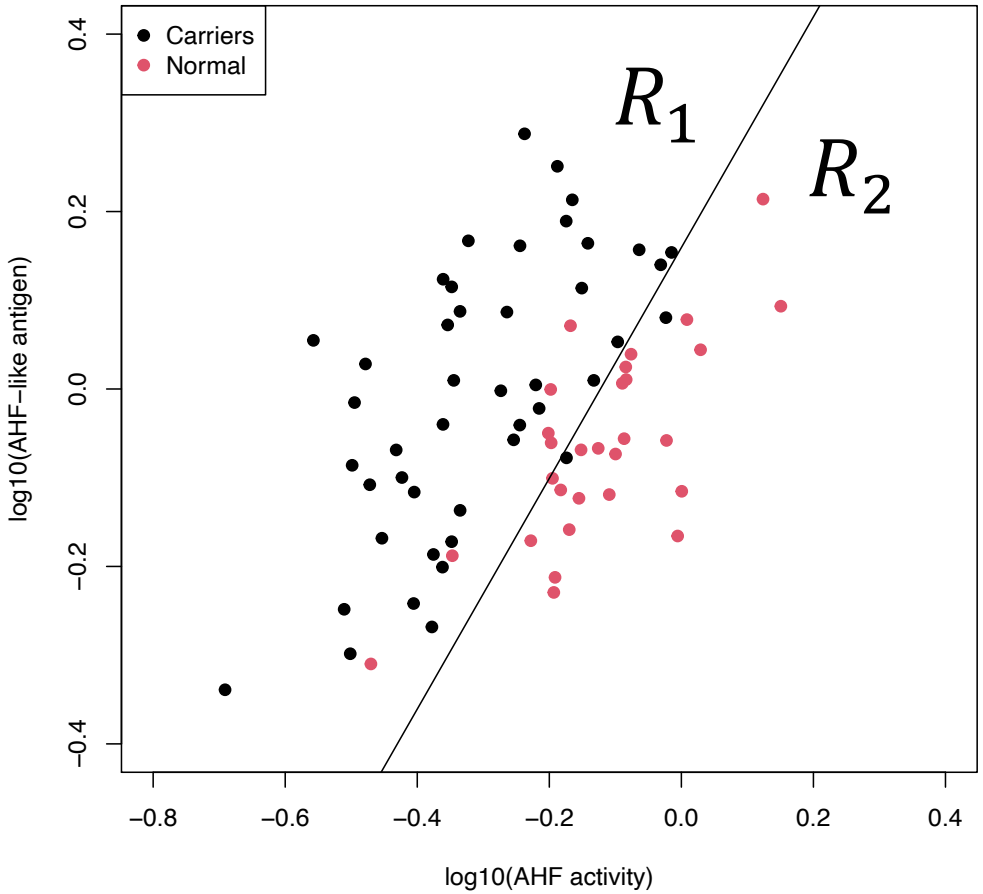
π_1 : <u>truely</u> in group 1 (carriers)
π_2 : <u>truely</u> in group 2 (normal)

R_1 : <u>allocated</u> in group 1 (carriers)
R_2 : <u>allocated</u> in group 2 (normal)



Some rules are designed to minimize the *risks*

Risk of Misclassification #1



$$P(\text{ a normal observation is classified as carrier})$$

$$\|$$

$$P(\text{ observation is normal and classified as carrier})$$

$$\|$$

$$P(A \cap B) = P(B|A) \cdot P(A)$$

$$\|$$

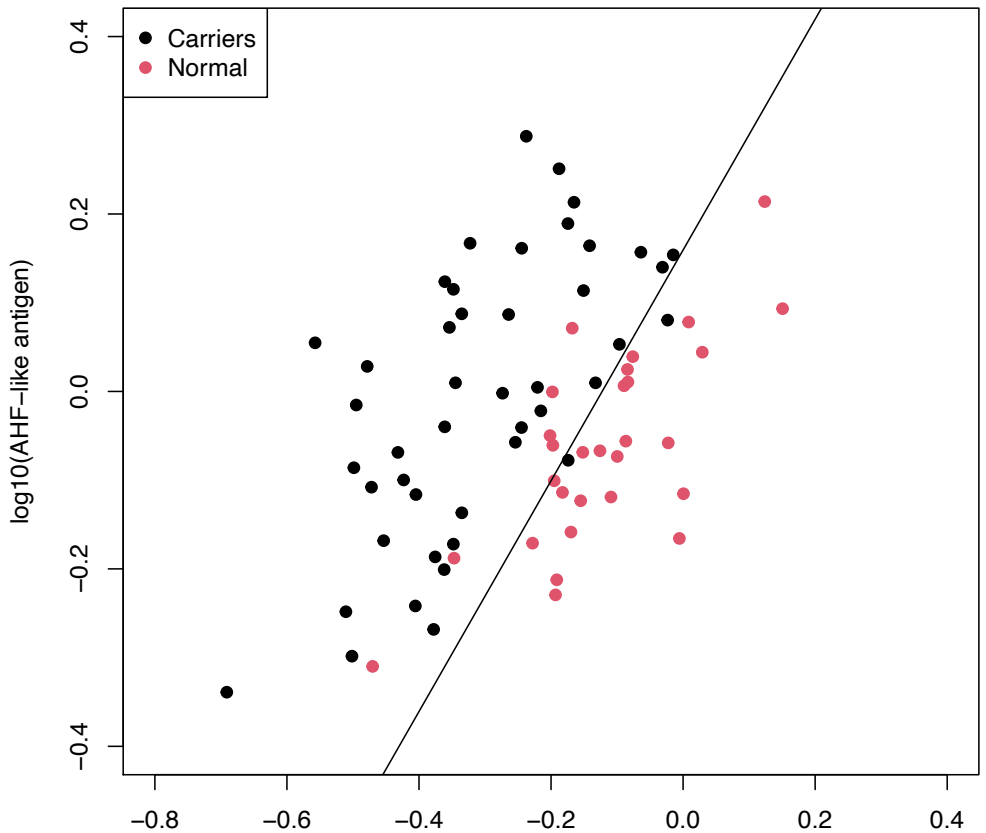
$$P(\text{classified as carrier } | \text{observation is normal})$$

$$\times$$

$$P(\text{ observation is normal})$$

Some rules are designed to minimize the *risks*

Risk of Misclassification #2



$$P(\text{ a carrier observation is classified as normal})$$

$$\|$$

$$P(\text{ observation is carrier and classified as normal})$$

$$\|$$

$$P(A \cap B) = P(B|A) \cdot P(A)$$

$$P(\text{classified as normal } | \text{observation is carrier})$$

$$\times$$

$$P(\text{ observation is carrier})$$

Expected Cost of Misclassification (ECM)

$$\boxed{\begin{gathered} Cost\ \#1 \times P(classified\ as\ carrier \mid observation\ is\ normal) \\ \times \\ P(\ observation\ is\ normal) \end{gathered}}$$

$$+$$

$$\boxed{\begin{gathered} Cost\ \#2 \times P(classified\ as\ normal \mid observation\ is\ carrier) \\ \times \\ P(\ observation\ is\ carrier) \end{gathered}}$$

$$\text{ECM} = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \qquad (11\text{-}5)$$

$$P(classified\ as\ Group\ 2 \mid observation\ is\ supposed\ to\ be\ in\ Group\ 1)$$

Expected Cost of Misclassification (ECM)

$$\boxed{\begin{gathered} Cost\ \#1 \times P(classified\ as\ carrier \mid observation\ is\ normal) \\ \times \\ P(\ observation\ is\ normal) \end{gathered}}$$

$$+$$

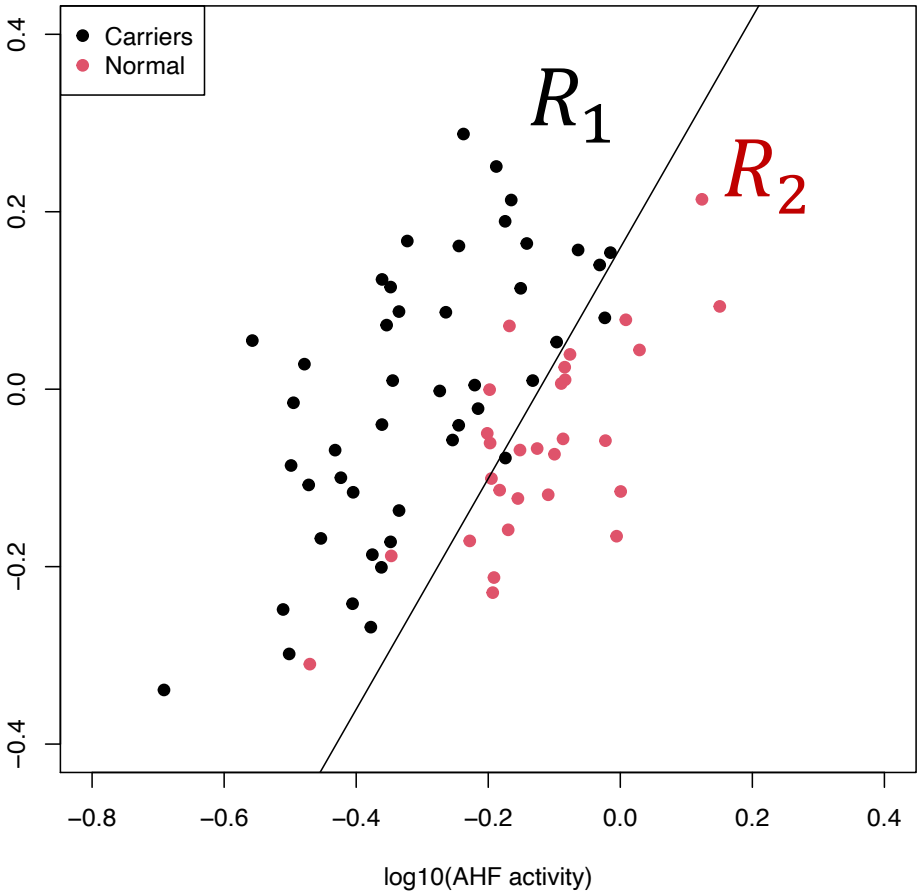
$$\boxed{\begin{gathered} Cost\ \#2 \times P(classified\ as\ normal \mid observation\ is\ carrier) \\ \times \\ P(\ observation\ is\ carrier) \end{gathered}}$$

$$\text{ECM} = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \qquad (11\text{-}5)$$

$$P(classified\ as\ Group\ 2 \mid observation\ is\ supposed\ to\ be\ in\ Group\ 1)$$

$$\text{ECM} = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \qquad (11\text{-}5)$$

Result 11.1. The regions R_1 and R_2 that minimize the ECM are defined by the values $\mathbf{x}$ for which the following inequalities hold:



$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} \geq \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix} \qquad (11\text{-}6)$$

$$R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)$$

$$\begin{pmatrix}\text{density}\\ \text{ratio}\end{pmatrix} < \begin{pmatrix}\text{cost}\\ \text{ratio}\end{pmatrix}\begin{pmatrix}\text{prior}\\ \text{probability}\\ \text{ratio}\end{pmatrix}$$

Special Cases of Minimum Expected Cost Regions

(a) $p_2/p_1 = 1$ (equal prior probabilities)

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{c(1|2)}{c(2|1)} \qquad R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{c(1|2)}{c(2|1)}$$

(b) $c(1|2)/c(2|1) = 1$ (equal misclassification costs)

TPM

Total probability of misclassification rule

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \frac{p_2}{p_1} \qquad R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \frac{p_2}{p_1} \tag{11-7}$$

(c) $p_2/p_1 = c(1|2)/c(2|1) = 1$ or $p_2/p_1 = 1/(c(1|2)/c(2|1))$
(equal prior probabilities and equal misclassification costs)

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq 1 \qquad R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < 1$$

Example 11.2

Given the prior probabilities and costs of misclassification, we can use (11-6) to derive the classification regions R_1 and R_2. Specifically, we have

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

$$R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

Suppose the density functions evaluated at a new observation $\mathbf{x}_0$ give $f_1(\mathbf{x}_0) = .3$ and $f_2(\mathbf{x}_0) = .4$. Do we classify the new observation as π_1 or π_2? To answer the question, we form the ratio

$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} = \frac{.3}{.4} = .75$$

Example 11.2

Given the prior probabilities and costs of misclassification, we can use (11-6) to derive the classification regions R_1 and R_2. Specifically, we have

$$R_1: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} \geq \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

$$R_2: \quad \frac{f_1(\mathbf{x})}{f_2(\mathbf{x})} < \left(\frac{10}{5}\right)\left(\frac{.2}{.8}\right) = .5$$

Suppose the density functions evaluated at a new observation $\mathbf{x}_0$ give $f_1(\mathbf{x}_0) = .3$ and $f_2(\mathbf{x}_0) = .4$. Do we classify the new observation as π_1 or π_2? To answer the question, we form the ratio

$$\frac{f_1(\mathbf{x}_0)}{f_2(\mathbf{x}_0)} = \frac{.3}{.4} = .75$$

$\boldsymbol{x}_0$ should be allocated to Group 1, R_1

If, normally distributed (Test it yourself!)

$$N(\boldsymbol{\mu}_1, \Sigma_1) \rightarrow f_1(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_1|} \exp\{-(\boldsymbol{x}-\boldsymbol{\mu}_1)'\Sigma_1^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_1)/2\}$$

The (R_1, R_2) minimize ECM is

$$R_1 = \left\{ \boldsymbol{x} \middle| \frac{\boxed{f_1(\boldsymbol{x})}}{\boxed{f_2(\boldsymbol{x})}} \geq \frac{c(1|2)}{c(2|1)} \frac{p_2}{p_1} \right\}$$

$$N(\boldsymbol{\mu}_2, \Sigma_2) \rightarrow f_2(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\Sigma_2|} \exp\{-(\boldsymbol{x}-\boldsymbol{\mu}_2)'\Sigma_2^{-1}(\boldsymbol{x}-\boldsymbol{\mu}_2)/2\}$$

If, normally distributed, further with MASS in R

$\boxed{\Sigma_1 = \Sigma_2 = \Sigma}$

$\boxed{\Sigma_1 = \Sigma_2\ ?}$

Homogeneous?

$\boxed{\Sigma_1 \neq \Sigma_2}$

Allocate $\mathbf{x}_0$ to π_1 if

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (11\text{-}18)$$

Allocate $\mathbf{x}_0$ to π_2 otherwise.

$$\mathbf{S}_{\text{pooled}} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_2 \quad (11\text{-}17)$$

linear discriminant analysis (R cmd `lda()` and `predict()`)

Allocate $\mathbf{x}_0$ to π_1 if

$$-\frac{1}{2}\mathbf{x}_0'(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{x}_0 + (\bar{\mathbf{x}}_1'\mathbf{S}_1^{-1} - \bar{\mathbf{x}}_2'\mathbf{S}_2^{-1})\mathbf{x}_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (11\text{-}29)$$

Allocate $\mathbf{x}_0$ to π_2 otherwise.

quadratic discriminant analysis (R cmd `qda()` and `predict()`)

If **not** normally distributed, use logistic regression model §11.7

- The method depends on a logistic regression model based on the log odds ratio $\ln\left(\frac{p}{1-p}\right)$

$$value = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

  in which coefficients

$$\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$$

  are determined based on maximum likelihood theory.

- In practice, the model including $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$ can be constructed with R cmd

```
glm('group' ~ 'predictors', data, family=binomial(link="logit"))
```

- Classification criterion: Allocate $\boldsymbol{x}$ to group 1 if the estimated posterior probability

$$\hat{P}(1|\boldsymbol{X} = \boldsymbol{x}) = \frac{e^{value}}{1 + e^{value}} \geq 1/2$$

Well, we have LDA, QDA and logistic…so compare

Comparison?

APER. apparent error rate

more criterion coming later

$$\frac{\#\text{ of data misclassified}}{\text{size of data.}}$$

Confusion matrix table()

		predicted π_1	predicted π_2	
Actual	π_1	n_{1C}	n_{1M}	n_1
	π_2	n_{2M}	n_{2C}	n_2

$$APER = \frac{n_{1M} + n_{2M}}{n_1 + n_2}$$

LDA

27	3
8	37

QDA

27	3
8	37

Logistic

25	5
4	41

✓

$$\frac{9}{75} = 12\%$$

$\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2$?
Homogeneous in general: Box's M-test §6.6

Assume g different groups with distributions $N_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1), N_p(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_2), \ldots, N_p(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$ and there is *independence* between the observations belonging to different groups. We are interested in testing

$$\begin{aligned} H_0 &: \boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \ldots = \boldsymbol{\Sigma}_g, \\ H_1 &: \text{at least two } \boldsymbol{\Sigma}_i \text{ not equal} , \end{aligned}$$

at the significance level α.

Box's M-test (maximum likelihood test)

- Test statistic: $M = -2 \ln \Lambda$

$$\boxed{M = (n-g)\ln|\boldsymbol{S}_{\text{pooled}}| - \sum_{\ell=1}^{g}(n_\ell - 1)\ln|\boldsymbol{S}_\ell|,}$$

In which, likelihood ratio $\Lambda = \prod_{\ell=1}^{g}\left(\frac{|\boldsymbol{S}_\ell|}{|\boldsymbol{S}_{\text{pooled}}|}\right)^{(n_\ell-1)/2}$,

with

$$\begin{aligned}
\boldsymbol{S}_\ell &= \frac{1}{n_\ell - 1}\sum_{j=1}^{n_\ell}(\boldsymbol{X}_{\ell j} - \bar{\boldsymbol{X}}_\ell)(\boldsymbol{X}_{\ell j} - \bar{\boldsymbol{X}}_\ell)', \\
\boldsymbol{S}_{\text{pooled}} &= \frac{1}{n-g}\sum_{\ell=1}^{g}\sum_{j=1}^{n_\ell}(\boldsymbol{X}_{\ell j} - \bar{\boldsymbol{X}}_\ell)(\boldsymbol{X}_{\ell j} - \bar{\boldsymbol{X}}_\ell)',
\end{aligned}$$

where $n = \sum_{\ell=1}^{g} n_\ell$.

Box's M-test

Then, under H_0

$$(1-u)M \dot{\sim} \chi^2_\nu$$

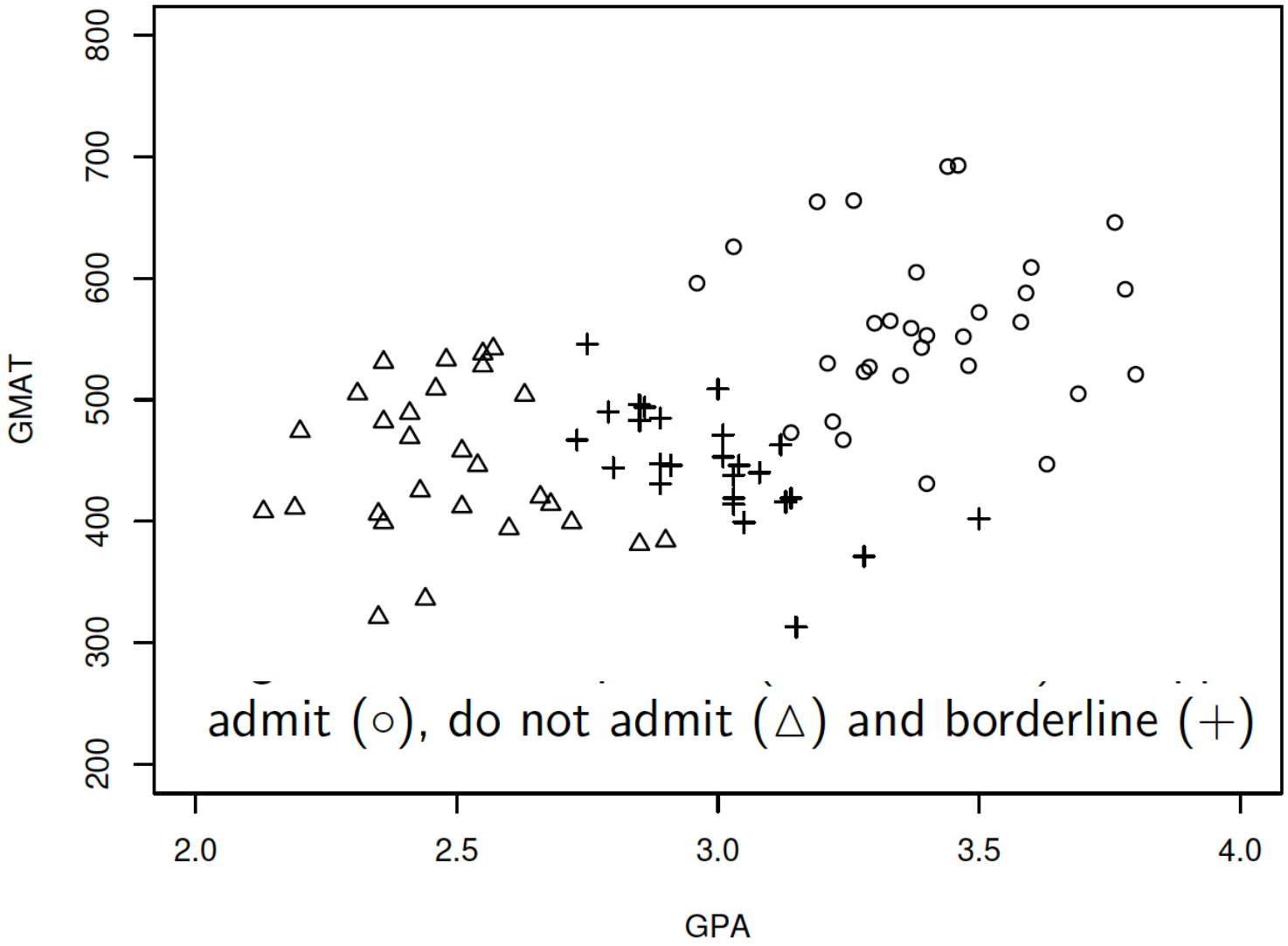
$$u = \left[\sum_{\ell=1}^{g} \frac{1}{n_\ell - 1} - \frac{1}{n-g}\right] \frac{2p^2+3p-1}{6(p+1)(g-1)}. \qquad \nu = \frac{1}{2}\ p(p+1)(g-1).$$

Reject H_0, if $(1-u)M > \chi^2_v(\alpha)$

Try it out with the Example 11.3, i.e. hemophilia data set

What if there are $\geq$ **3** groups…

Re-visit Example 11.11 (Business-school-data)



ECM again

The conditional expected cost of misclassifying an $\mathbf{x}$ from π_1 into π_2, or $\pi_3, \ldots,$ or π_g is

$$\begin{aligned}\mathrm{ECM}(1) &= P(2|1)c(2|1) + P(3|1)c(3|1) + \cdots + P(g|1)c(g|1) \\ &= \sum_{k=2}^{g} P(k|1)c(k|1)\end{aligned}$$

$$\begin{aligned}\mathrm{ECM} &= p_1\mathrm{ECM}(1) + p_2\mathrm{ECM}(2) + \cdots + p_g\mathrm{ECM}(g) \\ &= p_1\left(\sum_{k=2}^{g} P(k|1)c(k|1)\right) + p_2\left(\sum_{\substack{k=1\\k\neq 2}}^{g} P(k|2)c(k|2)\right) \\ &\quad + \cdots + p_g\left(\sum_{k=1}^{g-1} P(k|g)c(k|g)\right) \\ &= \boxed{\sum_{i=1}^{g} p_i\left(\sum_{\substack{k=1\\k\neq i}}^{g} P(k|i)c(k|i)\right)}\end{aligned}$$

Minimize this!

Result 11.5 (choose the one with the least risk)

Result 11.5. The classification regions that minimize the ECM (11-37) are defined by allocating $\mathbf{x}$ to that population π_k, $k = 1, 2, \ldots, g$, for which

Extra weight, If identical, ECM->TPM

$$\sum_{\substack{i=1 \\ i \neq k}}^{g} \boxed{p_i} \boxed{f_i(\mathbf{x})} \boxed{c(k \mid i)} \tag{11-38}$$

Density of individual group!

is smallest. If a tie occurs, $\mathbf{x}$ can be assigned to any of the tied populations.

Proportion of the group I

TPM+MVN+Unequal Σ_i ➡ QDA

Minimum Total Probability of Misclassification (TPM) Rule for Normal Populations—Unequal Σ_i

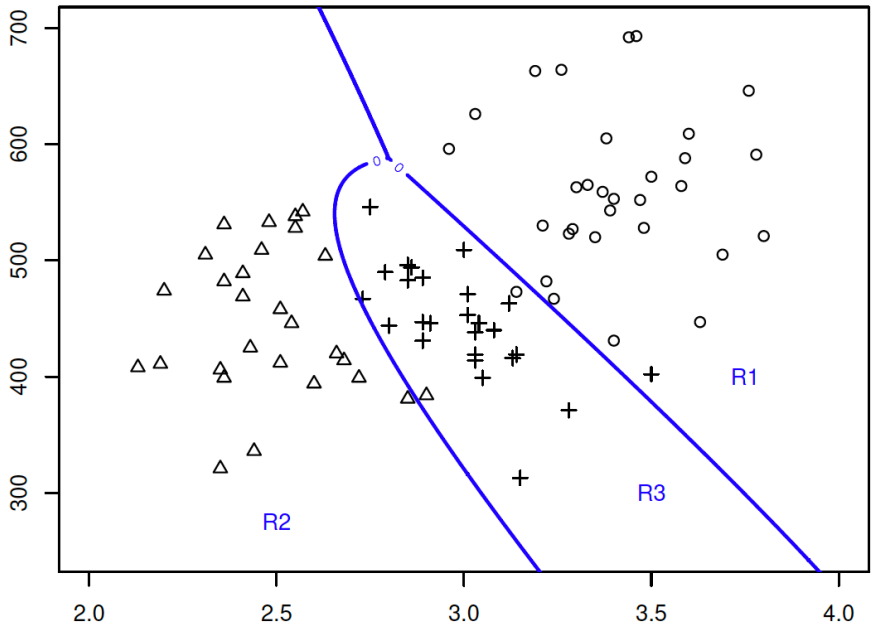
Allocate $\mathbf{x}$ to π_k if

$$\text{the quadratic score } d_k^Q(\mathbf{x}) = \text{largest of } d_1^Q(\mathbf{x}), d_2^Q(\mathbf{x}), \ldots, d_g^Q(\mathbf{x}) \qquad (11\text{-}46)$$

where $d_i^Q(\mathbf{x})$ is given by (11-45).

$$d_i^Q(\mathbf{x}) = -\tfrac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)'\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln p_i \qquad i = 1, 2, \ldots, g \qquad (11\text{-}45)$$

R cmd: qda() + predict ()



MVN+equal Σ ➡ LDA

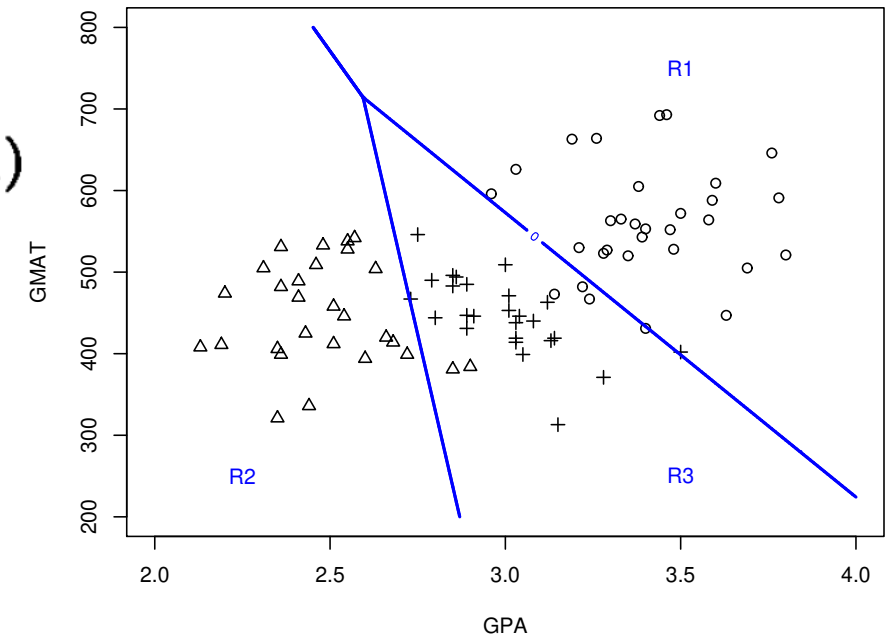
Estimated Minimum TPM Rule for Equal-Covariance Normal Populations

Allocate $\mathbf{x}$ to π_k if

the linear discriminant score $\hat{d}_k(\mathbf{x})$ = the largest of $\hat{d}_1(\mathbf{x}), \hat{d}_2(\mathbf{x}), \ldots, \hat{d}_g(\mathbf{x})$ (11-52)

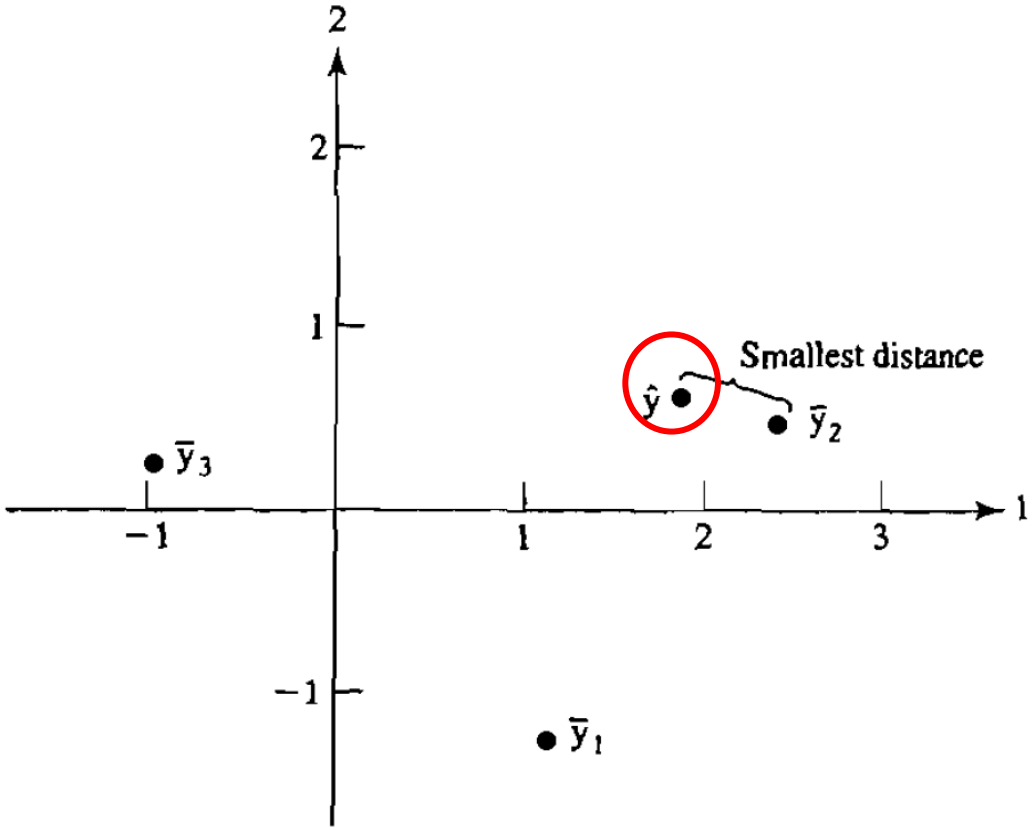
with $\hat{d}_i(\mathbf{x})$ given by (11-51).

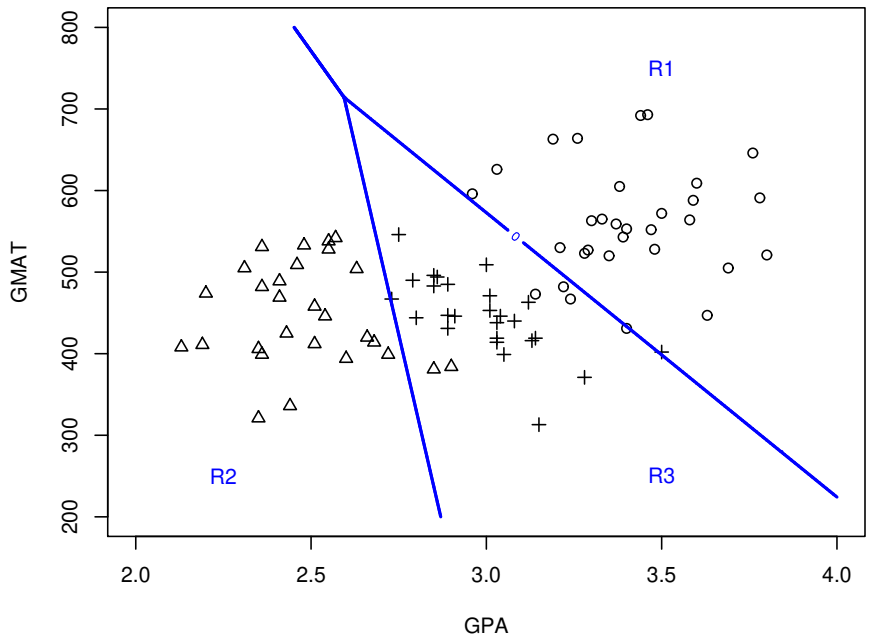
$$\hat{d}_i(\mathbf{x}) = \bar{\mathbf{x}}_i' \mathbf{S}_{\text{pooled}}^{-1} \mathbf{x} - \tfrac{1}{2} \bar{\mathbf{x}}_i' \mathbf{S}_{\text{pooled}}^{-1} \bar{\mathbf{x}}_i + \ln p_i \quad \text{for } i = 1, 2, \ldots, g \qquad (11\text{-}51)$$



R cmd: lda() + predict ()

~~MVN~~+equal Σ ➡ Fisher's Method (*nothing new*)





R cmd: lda() + predict ()

If **not** normally distributed, use logistic regression model §11.7

- The method depends on a logistic regression model based on the log odds ratio

$$\ln\left(\frac{p}{1-p}\right) \quad value = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

in which coefficients

$$\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$$

are determined based on maximum likelihood theory.

- In practice, the model including $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$ can be constructed with R cmd

```
glm('group' ~ 'predictors', data, family=binomial(link="logit"))
```

- Classification criterion: Allocate $\boldsymbol{x}$ to group 1 if the estimated posterior probability

$$\hat{P}(1|\boldsymbol{X} = \boldsymbol{x}) = \frac{e^{value}}{1 + e^{value}} \geq 1/2$$

~~MVN+equal Σ~~ → Multi-nomial Logistic Discrimination

Assume there are in total K categories

1. The log odds ratios are modelled as the following.

   Coefficients are selected using maximum likelihood method.

$$\log\left(\frac{\mathbb{P}(Y=1|X)}{\mathbb{P}(Y=K|X)}\right) = \beta_1^T X$$

$$\log\left(\frac{\mathbb{P}(Y=2|X)}{\mathbb{P}(Y=K|X)}\right) = \beta_2^T X$$

$$\log\left(\frac{\mathbb{P}(Y=K-1|X)}{\mathbb{P}(Y=K|X)}\right) = \beta_{K-1}^T X$$

2. This is then equivallent to →

$$\mathbb{P}(Y=1|X) = \frac{\exp(\beta_1^T X)}{1+\sum_{j=1}^{K-1}\exp(\beta_j^T X)}$$

$$\mathbb{P}(Y=2|X) = \frac{\exp(\beta_2^T X)}{1+\sum_{j=1}^{K-1}\exp(\beta_j^T X)}$$

$$\vdots$$

$$\mathbb{P}(Y=K-1|X) = \frac{\exp(\beta_{K-1}^T X)}{1+\sum_{j=1}^{K-1}\exp(\beta_j^T X)}$$

$$\mathbb{P}(Y=K|X) = \frac{1}{1+\sum_{j=1}^{K-1}\exp(\beta_j^T X)}$$

3. For a particular element, we select the category with the highest chance.

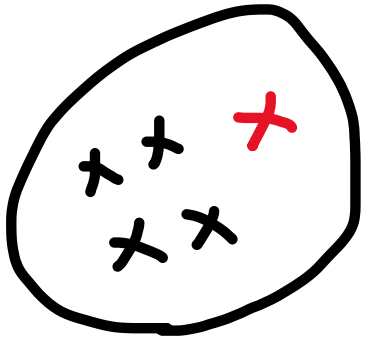
Cross validation→ $\hat{E}$(APER)

It is also called "hold-out" procedure in some literature (including textbook)

Basically, each data point takes its turn to be classified based on the rest data, then the proportion of "misclassified" becomes $\hat{E}$(APER) [expected actural error rate]

Because they employ estimates of population parameters, the sample classification rules (11-48) and (11-52) may no longer be optimal. Their performance, however, can be evaluated using Lachenbruch's holdout procedure. If $n_{iM}^{(H)}$ is the number of misclassified holdout observations in the ith group, $i = 1, 2, \ldots, g$, then an estimate of the expected actual error rate, $E(\text{AER})$, is provided by

$$\hat{E}(\text{AER}) = \frac{\sum_{i=1}^{g} n_{iM}^{(H)}}{\sum_{i=1}^{g} n_i} \tag{11-57}$$



$$\hat{E}(\text{AER}) = \frac{n_{1M}^{(H)} + n_{2M}^{(H)}}{n_1 + n_2} \tag{11-36}$$

Cross validation→ $\hat{E}$(APER)

R cmd regarding this part:

- For this course, we do **not** use '70%-30%' or sth similar ☺
- Easy ones: lda() and qda() both have an option 'CV=TRUE'
- Medium: A for-loop in combination with glm() [see hemophiliaEVA.R]
- Challenge: A for-loop in combination with multinom()

  [show me what you got]

If **not** normally distributed, use logistic regression model §11.7

- The method depends on a logistic regression model based on the log odds ratio

$$\ln\left(\frac{p}{1-p}\right) \quad value = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

in which coefficients

$$\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$$

are determined based on maximum likelihood theory.

- In practice, the model including $\boldsymbol{\beta}' = (\beta_0, \beta_1, \ldots, \beta_p)$ can be constructed with R cmd

```
glm('group' ~ 'predictors', data, family=binomial(link="logit"))
```

- Classification criterion: Allocate $\boldsymbol{x}$ to group 1 if the estimated posterior probability

$$\hat{P}(1|\boldsymbol{X} = \boldsymbol{x}) = \frac{e^{value}}{1 + e^{value}} \geq 1/2$$

If, normally distributed, further with MASS in R

$\boxed{\Sigma_1 = \Sigma_2 = \Sigma}$

$\boxed{\Sigma_1 = \Sigma_2\ ?}$

Homogeneous?

$\boxed{\Sigma_1 \neq \Sigma_2}$

Allocate $\mathbf{x}_0$ to π_1 if

$$(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}\mathbf{x}_0 - \frac{1}{2}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)'\mathbf{S}_{\text{pooled}}^{-1}(\bar{\mathbf{x}}_1 + \bar{\mathbf{x}}_2) \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (11\text{-}18)$$

Allocate $\mathbf{x}_0$ to π_2 otherwise.

$$\mathbf{S}_{\text{pooled}} = \left[\frac{n_1 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_1 + \left[\frac{n_2 - 1}{(n_1 - 1) + (n_2 - 1)}\right]\mathbf{S}_2 \quad (11\text{-}17)$$

linear discriminant analysis (R cmd `lda()` and `predict()`)

Allocate $\mathbf{x}_0$ to π_1 if

$$-\frac{1}{2}\mathbf{x}_0'(\mathbf{S}_1^{-1} - \mathbf{S}_2^{-1})\mathbf{x}_0 + (\bar{\mathbf{x}}_1'\mathbf{S}_1^{-1} - \bar{\mathbf{x}}_2'\mathbf{S}_2^{-1})\mathbf{x}_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (11\text{-}29)$$

Allocate $\mathbf{x}_0$ to π_2 otherwise.

quadratic discriminant analysis (R cmd `qda()` and `predict()`)

Receiver Operating Characteristic Curve

$$P(2|x) > cut\_off$$

ROC by default only works for classification into 2 groups.

One of these two is viewed to be 'positive' (say the 2nd group)

- somehow more important to be correctly predicted
- In ROCR, simply the group with Class=1

- Confusion matrix (for a fixed rule)

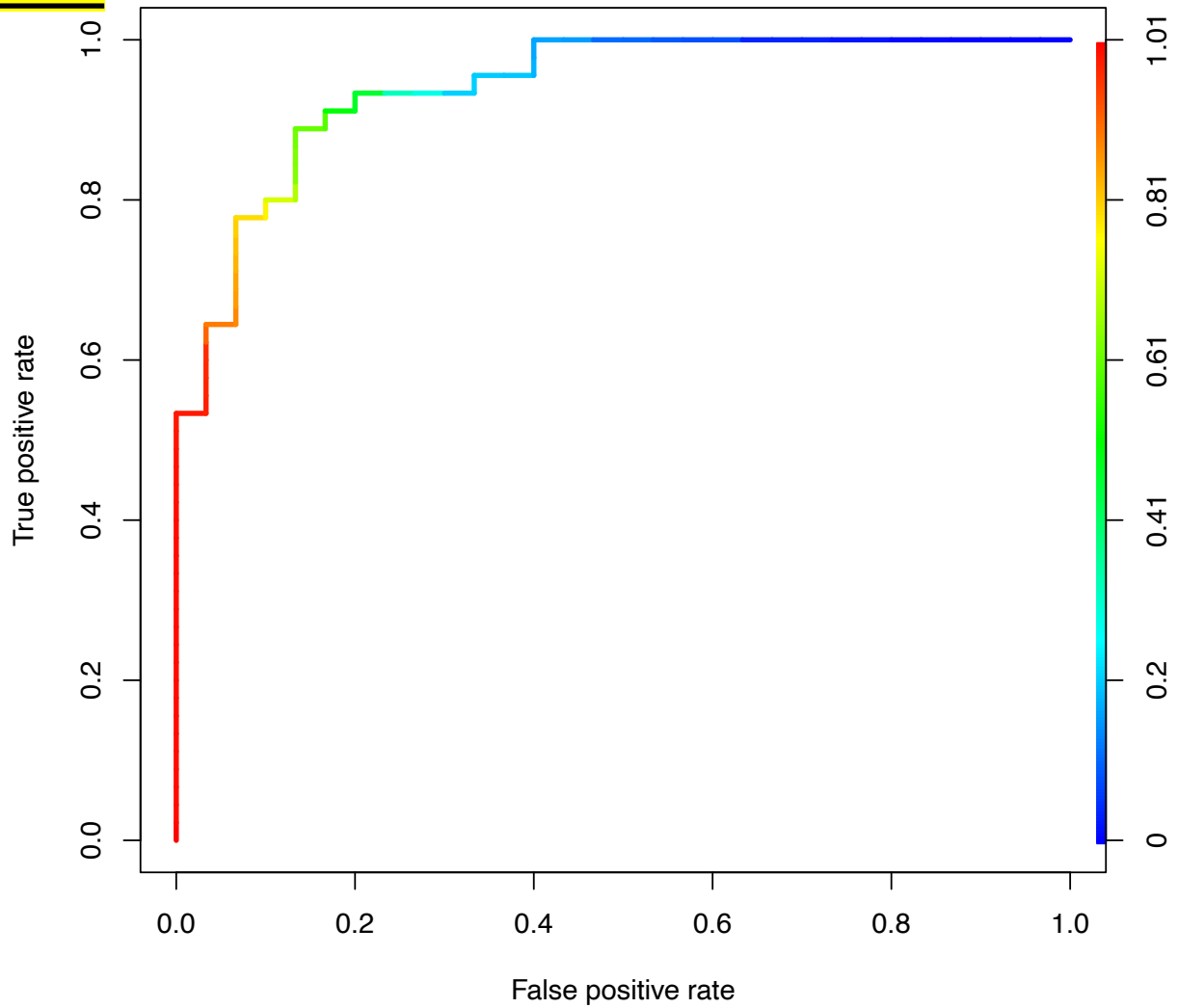
		predicted negative π_1	predicted positive π_2	
Actual	π_1	n_{1C}	n_{1M}	n_1
	π_2	n_{2M}	n_{2C}	n_2

of π_1-element misclassified as π_2

of π_2-element misclassified as π_1

- False positive rate (FPR) $= \dfrac{n_{1M}}{n_1}$, $n_1 = n_{1C} + n_{1M}$

True positive rate (TPR) $= \dfrac{n_{2C}}{n_2}$



Receiver Operating Characteristic Curve

- Confusion matrix (for a fixed rule)

		predicted negative	predicted positive	
		π_1	π_2	
Actual	π_1	n_{1C}	n_{1M}	n_1
	π_2	n_{2M}	n_{2C}	n_2

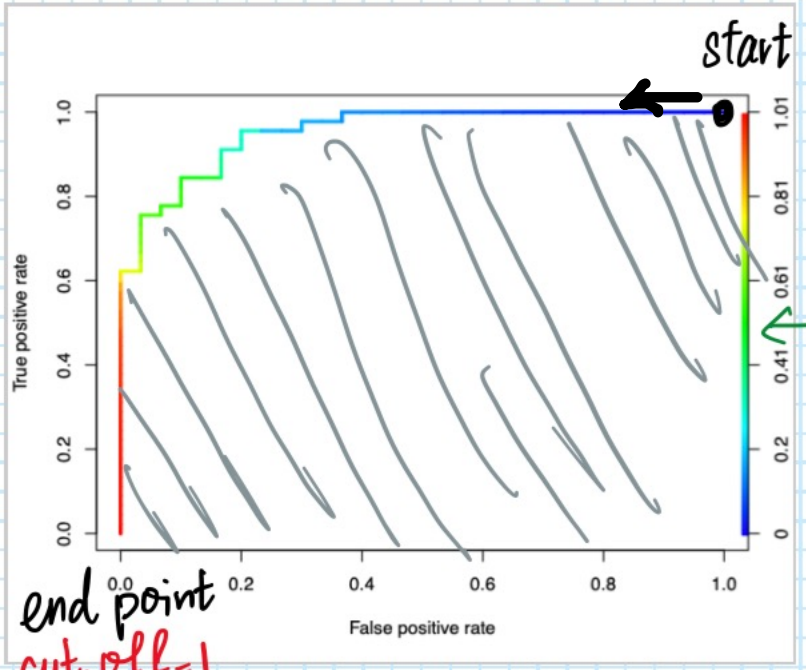
n_{1M}: # of π_1-element misclassified as π_2

n_{2M}: # of π_2-element misclassified as π_1

- False positive rate (FPR) $= \dfrac{n_{1M}}{n_1}$, $n_1 = n_{1C} + n_{1M}$

True positive rate (TPR) $= \dfrac{n_{2C}}{n_2}$

ROC curve cut-off c from 0 to 1



start point cutoff = 0

end point cut-off = 1

color indicate value of cutoff.

start point

cut-off ⇒ "all to π_2"

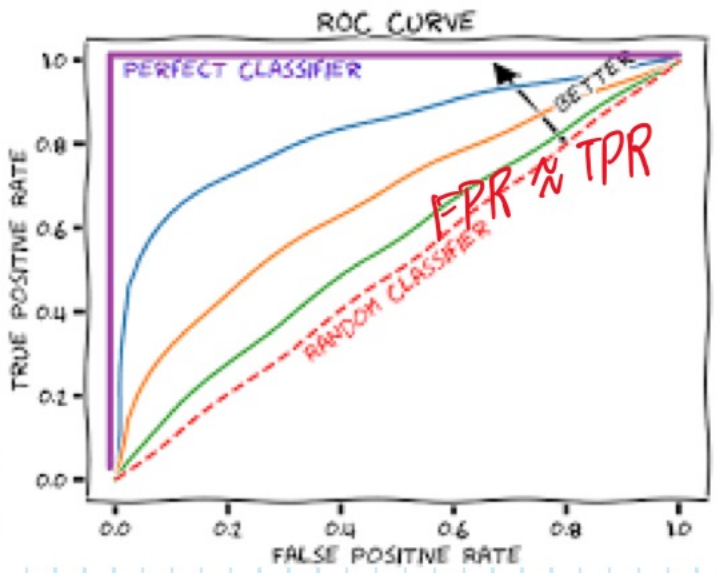
0 ⇓ FPR = 1, TPR = 1

↓

1 ⇒ "all to π_1"

FPR = 0, TPR = 0

Receiver Operating Characteristic Curve



AUC "area under the curve"

① AUC ≤ 1

② better classifier. bigger AUC

"perfect" at certain cutoff

FPR = 0 TPR = 1

• Confusion matrix (for a fixed rule)

		predicted negative	predicted positive	
Actual	π_1	n_{1C}	n_{1M}	n_1
	π_2	n_{2M}	n_{2C}	n_2

of π_1-element misclassified as π_2

of π_2-element misclassified as π_1

· False positive rate (FPR) $= \dfrac{n_{1M}}{n_1}$, $n_1 = n_{1C} + n_{1M}$

True positive rate (TPR) $= \dfrac{n_{2C}}{n_2}$

Variable selection

- Sometimes a big pool of variables is available and one would like to find a good subset of these, in order to have a model that is as parsimonious as possible, but still provides a good classification.

- In other cases a good classifier might already be available, but one would like to know if all variables in the model contribute substantially to the classification.

- One possibility to obtain information that is relevant for investigating the above questions is to perform a complete enumeration of models, i.e. fit all models with one variable, all models with two variables,...

- The models then have to be evaluated/compared on the basis of some criterion function, e.g. $APER$

Variable selection (Exercise 11.27 IRIS data)

Three species of iris flowers (Rscript `irisaper.R`)

π_1: iris setosa,

π_2: iris versicolor,

π_3: iris virginica.

Assign an iris flower to one of the classes on the basis of

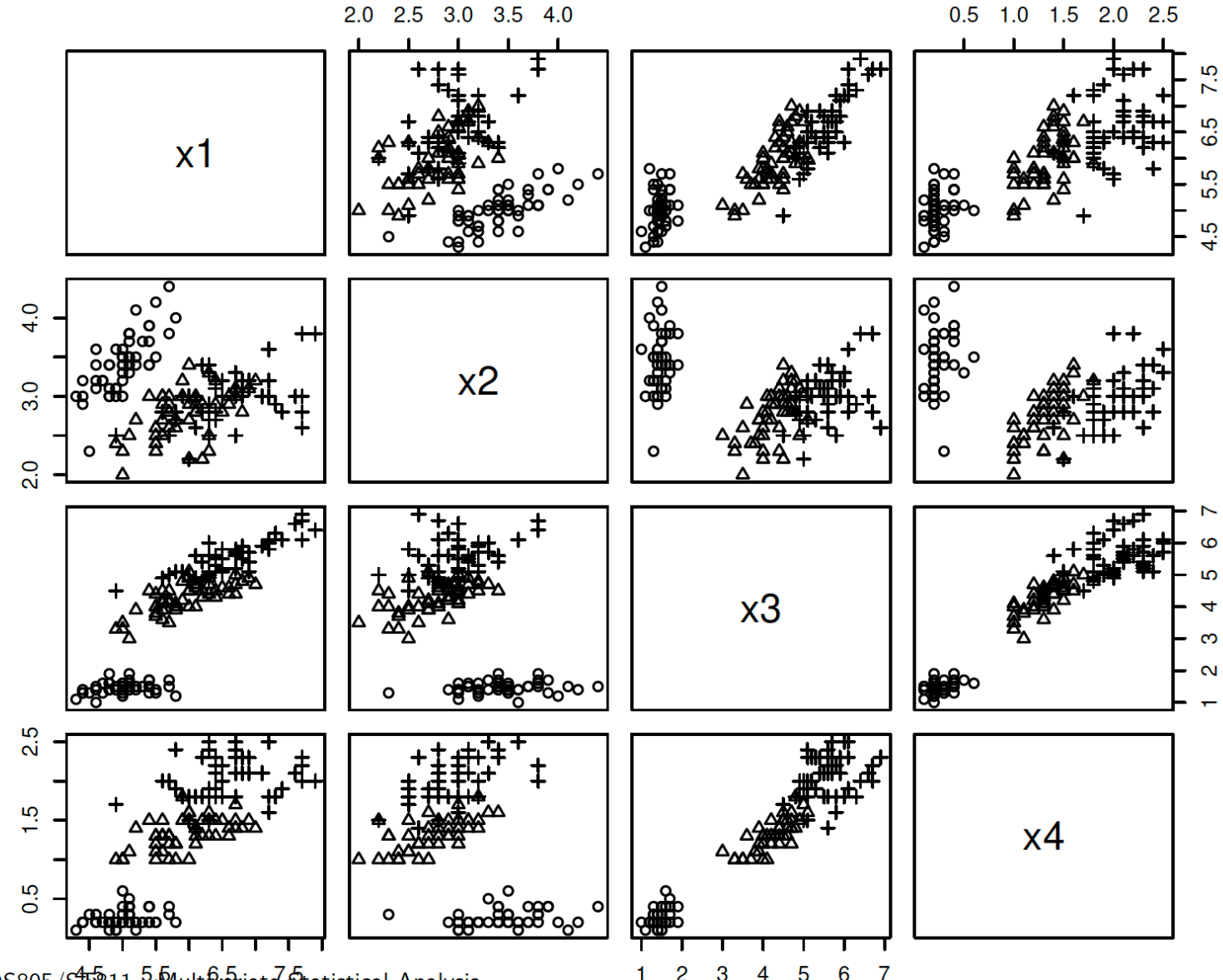
X_1: sepal length,

X_2: sepal width,

X_3: petal length,

X_4: petal width.

Variable selection (Exercise 11.27 IRIS data)



Variable selection (Exercise 11.27 IRIS data)

Table 1: Classification results for linear classifier, equal costs, equal priors

Subset size	Variables	$APER$ (exercise)
1	X_1	
	X_2	
	X_3	
	X_4	
2	X_1, X_2	
	X_1, X_3	
	X_1, X_4	
	X_2, X_3	
	X_2, X_4	
	X_3, X_4	
3	X_1, X_2, X_3	
	X_1, X_2, X_4	
	X_1, X_3, X_4	
	X_2, X_3, X_4	
4	X_1, X_2, X_3, X_4	0.020