12] leave one out cross validation (CV):

For $i = 1, \ldots, n$, leave out case i, compute the discriminant rule, and see if case i is correctly classified.

Let m_c be the number of cases correctly classified. The

CV error rate $= 1 - \frac{m_c}{n}$.

13] Leave out a validation set (data splitting!) with enough cases n_v, 10% to 50%, so that a good error estimate can be obtained. Compute the discriminant rule from the $n - n_v$ cases not in the validation set. Then classify the cases in the validation set. Let m_L be the number of cases correctly classified. The

"validation set" error rate $= 1 - \frac{m_L}{n_v}$.

14) ✓ K-fold CV: make k folds, leave fold out fit method classify cases in fold, repeat

Let m_K be the number of cases correctly classified. Then the K-fold CV error rate is

$$1 - \frac{m_K}{n}.$$

Note: a) 12] is n-fold CV. For 14], often $K = 5$ or 10.

b) 14] attempts to choose a model good for classification (good for predicting the group).

More on classification:

---

ex] Patient with heart attack symptoms:

3 tests are done $\underline{W} = (W_1, W_2, W_3)^T$

and the patient is classified as

1 = had heart attack or 0 = did not have a heart attack.

ex) person applies for credit card based on $(\text{salary}, \text{credit rating})^T$ and is classified as 1 = acceptable, 0 = not acceptable.

15] ✓ Suppose training data consists of a random sample of n_j cases

$\underline{x}_{1j}, \ldots, \underline{x}_{n_j j}$ for each group j.

Let $(\bar{\underline{x}}_j, S_j)$ be the sample mean and sample covariance matrix for each group. Let $\underline{w}_i$ be a new test random vector from 1 of the $G = 2$ groups, but the group is unknown.

Usually have $\underline{w}_1, \ldots, \underline{w}_m$ = test data, and discriminant analysis or classification attempts to allocate the $\underline{w}_i$ to correct groups.

16} Several discriminant rules can be modified to incorporate $\pi_j = P(Y=j)$ (↑ group label) and costs of correct and incorrect allocation. We will assume that the costs are unknown or equal and that the π_j are unknown or $\pi_j = \frac{1}{G}$, $j=1,\dots,G$.

17} The pooled covariance matrix

$$S_{pool} = \frac{1}{n-G} \sum_{j=1}^{G} (n_j - 1) S_j.$$ This estimator is good if the G groups have the same pop cov matrix $\Sigma_{\underline{x}}$, and can be useful if the n_j are not large enough for S_j to be a good estimator of $\Sigma_{\underline{x}_j} = \Sigma_j$, $j=1,\dots,G$.

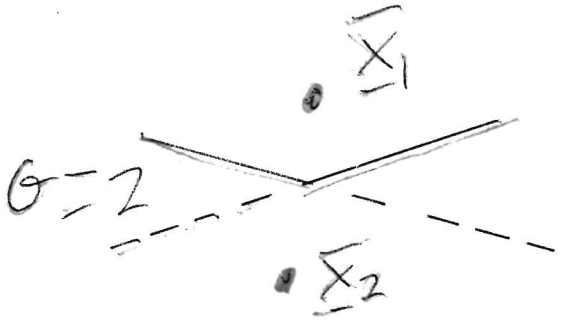
The G Σ_i have $G \frac{p(p+1)}{2}$ unknown parameters.

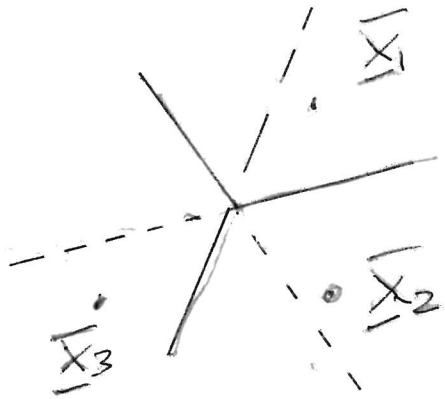
Roughly want $n_j \geq 10p$.

18] The linear discriminant analysis

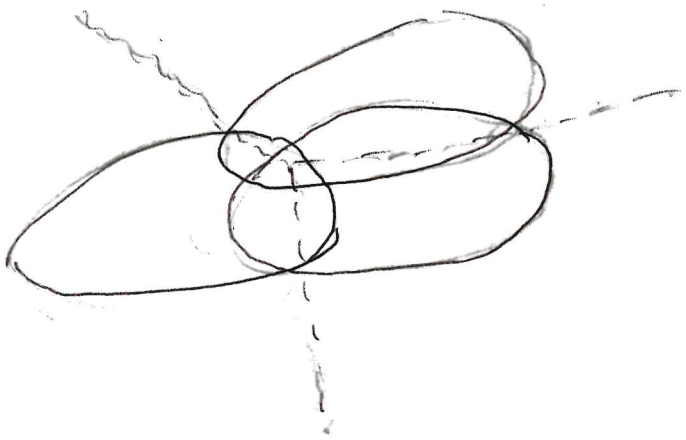
LDA rule allocates $\underline{w}$ to the group g with the largest value of $d_j(\underline{w}) = \bar{x}_j^T S_{pool}^{-1} \underline{w} - \frac{1}{2} \bar{x}_j^T S_{pool}^{-1} \bar{x}_j$

$= \alpha_j + \beta_j^T \underline{w}$. LDA is widely used and basically separates the G groups with G hyperplanes.





19] LDA roughly makes G hyperellipsoids $\{ D_j^2(\bar{x}_j, S_p) \leq h^2 \}$ of the same volume and shape that cover most of the training data. The hyperplanes minimize the overlap of the hyperellipsoids.



3 ellipses are supposed to be the same up to location

20) In high dimensions, replace S_{pool} by $diag(S_{pool})$ or I_p, etc.

ex) LR output

label	coef estimate	stderr	Est/SE	pval
constant	$\hat{\alpha}$			
x_1	$\hat{\beta}_1$			
⋮				
x_p	$\hat{\beta}_p$			

ex) 0 = F 1 = M

label	estimate
constant	-19.7762
head measurements: circum	0.0144698
head measurements: length	0.037147

Let circum $= x_1 = 550$, length $= x_2 = 200$.

a) Find ESP for $\underline{x}$. (variant: $\underline{w}$)

soln] $ESP = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 =$

$-19.7762 + 0.0244688(550) + 0.0371472(200)$

$= \boxed{1.11108}$

b) Is $\underline{x}$ $(= \underline{w})$ classified in group $\overset{F}{0}$ or $\overset{M}{1}$?

soln] group 1 since $ESP > 0$

c) Find $\hat{g}(x) = \hat{p}(ESP)$

$$\text{soln} = \frac{e^{ESP}}{1+e^{ESP}} = \frac{3.0376}{4.0376} = \boxed{0.7523}$$

21] crude forward selection:

step 1) Choose $w_1 = x_{j_1}$ that minimizes AER.

2) keep w_1 in model and add $w_2 = x_{j_2}$ that minimizes AER. keep w_1 and w_2 in the model.

⋮

K) $w_1, \ldots, w_{K-1}$ are in the model.
Add $w_K = x_{j_K}$ that minimizes the AER.

⋮

p) $\{w_1, \ldots, w_p\} = \{x_1, \ldots, x_p\}$.

Final model might be the one with smallest AER given $n \geq jk^*$.

1 and 2 sample tests

See paper back on website and Ch. 9

1) 1 sample HD test

$H_0: \underline{\mu} = \underline{0}$ vs $H_1, H_A: \underline{\mu} \neq \underline{0}$

data $\underline{x}_1, \ldots, \underline{x}_n$ $p \times 1$, often $p >> n$.

To test $H_0: \underline{\mu} = \underline{\mu}_0$, use

$\underline{w}_i = \underline{x}_i - \underline{\mu}_0 \quad i = 1, \ldots, n$

Then $E(\underline{w}_i) = \underline{0}$ if H_0 is true.

2] $\underline{\mu} = \underline{0}$ iff $\|\underline{\mu}\|_2 = 0$ iff

$\underline{\mu}^T \underline{\mu} = 0$.

3] Some good tests try to estimate $\underline{\mu}^T \underline{\mu}$. In low dimensions, $\bar{\underline{X}}^T \bar{\underline{X}}$ is a good estimator.

4] If $\underline{x}_i \perp\!\!\!\perp \underline{x}_j$, then

$$E(\underline{x}_i^T \underline{x}_j) = E\left[\sum_k x_{ik} x_{jk}\right]$$

$$\overset{ind}{=} \sum_{k=1}^{p} E[x_{ik}] E[x_{jk}] \overset{\mu_k = E(x_{ik})}{=} \sum_k \mu_k^2 = \underline{\mu}^T \underline{\mu}$$

5] $E\left[\underline{x}_i^T \underline{x}_i\right] = E\left[\sum_k x_{ik} x_{ik}\right]$

$$= \sum_k E[x_{ik}^2] = \sum_k [\sigma_k^2 + \mu_k^2] = \underline{\mu}^T \underline{\mu} + \underline{\sigma}^T \underline{\sigma}$$

$\sigma_k^2 = V(x_{ik})$ $\qquad$ $\underline{\sigma} = (\sigma_1, \ldots, \sigma_p)^T$

6] $E(\bar{x}^T\bar{x}) = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i^T \frac{1}{n}\sum_{j=1}^{n} x_j\right]$

$i=j$ terms

$= \frac{1}{n^2}\sum_i\sum_j E(x_i^T x_j) = \frac{1}{n^2}\left(\sum_{i\neq j}^{n}\sum E[x_i^T x_j] + \sum_{i=1}^{n} E[x_i^T x_i]\right)$

$= \frac{1}{n^2}\left[(n^2-n)\,\mu^T\mu + n\,\mu^T\mu + n\,\sigma^T\sigma\right]$

$= \mu^T\mu + \frac{\sigma^T\sigma}{n} = \mu^T\mu + \frac{\sum_{k=1}^{p}\sigma_k^2}{n}$,

and the second term could be huge

if $p >> n$.

7] Some good HD tests use

$T_n = \frac{1}{n(n-1)}\sum\sum_{i\neq j} x_i^T x_j$ as an

estimator of $\mu^T\mu$.

8] When Ho $\underset{\sim}{\mu} = \underset{\sim}{0}$ is true, it has

been shown that $\frac{T_n}{\sqrt{V(T_n)}} \xrightarrow{D} N(0,1)$,

$\frac{T_n}{\sqrt{V(T_n)}}$ = z score if $H_0: \underline{\mu} - \underline{\mu} = 0$ is true

and thus $\frac{T_n}{S_n} \xrightarrow{D} N(0,1)$

if $S_n^2 \xrightarrow{P} V(T_n)$.

9] The low dimensional analog is the 1 sample Hotelling's T^2 statistic

$$T^2 = n(\bar{\underline{x}} - \underline{\mu}_0)^T \hat{\Sigma}_{\underline{x}}^{-1} (\bar{\underline{x}} - \underline{\mu}_0).$$

Set $\underline{\mu}_0 = \underline{0}$ and replace $\hat{\Sigma}_{\underline{x}}^{-1}$ by I to get an early HD test.

10] The nonparametric bootstrap (several other names) takes a sample of size n selected with replacement from $\underline{x}_1, \ldots, \underline{x}_n$ to get a bootstrap data set $\underline{x}_1^*, \ldots, \underline{x}_n^*$. Compute the statistic $T_n^* = T_n(\underline{x}_1^*, \ldots, \underline{x}_n^*)$.

Repeat B times to get the bootstrap sample for the statistic

$T_1^*, \ldots, T_B^*$. The bagging estimator $\bar{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^* =$ sample mean of the T_i^*.

ex] (T_n = MED(n) or $\bar{x}$ common)

data 1, 2, 5, 10, 50 MED(n) = 5 = T_n

bootstrap dataset	ordered	bootstrap sample B = 3
2, 10, 1, 2, 2	1, 2, 2, 2, 10	$T_1^* = 2$
50, 10, 50, 2, 2	2, 2, 10, 50, 50	$T_2^* = 10$
10, 50, 2, 1, 1	1, 1, 2, 10, 50	$T_3^* = 2$

$\bar{T}^* = \frac{1}{3}(2+10+2) = \frac{14}{3}$

[1] The nonparametric bootstrap uses the empirical dist

	x_1	...	x_n
$P(W = x_i)$	$\frac{1}{n}$		$\frac{1}{n}$

$P(\underline{x}_i \notin$ bootstrap data set$)$

$= \frac{1-\frac{1}{n}}{1} \cdots \frac{1-\frac{1}{n}}{n} = (1-\frac{1}{n})^n \to e^{-1} \approx .3679$

$\approx \frac{1}{3}$. So about $\frac{1}{3}$ of the bootstrap data set cases are replicates: $\underline{x}_i^* = \underline{x}_j^*$ for some $i \neq j$, and then $\underline{x}_i^{*T}\underline{x}_i^* = \underline{x}_i^{*T}\underline{x}_j^*$ behaves badly. So the nonparametric bootstrap (for $\frac{n}{p}$ small) is poor for $T_n = \frac{1}{n(n-1)} \sum_{i \neq j} \sum \underline{x}_i^T \underline{x}_j$

$= \frac{1}{n(n-1)} (\underline{a}^T\underline{a} - \underline{c}^T\underline{c}) = \bar{\underline{x}}^T\bar{\underline{x}} - \frac{1}{n} + r(S)$ where $\hat{\Sigma}_{\underline{x}}$

$\underline{a} = \sum_{i=1}^n \underline{x}_i = n\bar{\underline{x}}$ and $\underline{c} = (\underline{x}_1^T, \ldots, \underline{x}_n^T)^T = \begin{pmatrix} \underline{x}_1 \\ \vdots \\ \underline{x}_n \end{pmatrix}$.

So $\underline{c}^T\underline{c} = \sum_i \sum_j x_{ij}^2 = \sum_{i=j} \underline{x}_i^T\underline{x}_i$.

12] The m out of n bootstrap = deleted jackknife (with $n-d=m$) =

Subsampling with m cases drawn without replacement

draws cases $\underline{x}_1^*, \ldots, \underline{x}_m^*$ without replacement and computes $T^* = T_m(\underline{x}_1^*, \ldots, \underline{x}_m^*)$ where m near $\frac{n}{2}$ is common and there is theory for $\frac{m}{n} \to \tau \in (0,1)$ and for $\frac{m}{n} \to 0$ but $m \to \infty$.

using $m = \lfloor \frac{2n}{3} \rfloor$ simulated well.

[3] Let $m = \lfloor \frac{n}{2} \rfloor$.

Let $w_1, \ldots, w_m = \underline{x}_1^T \underline{x}_2, \underline{x}_3^T \underline{x}_4, \ldots, \underline{x}_{2m-1}^T \underline{x}_{2m}$.

Then the w_i are iid with

$E(w_i) = \underline{\mu}^T \underline{\mu}$ and $V(w_i) = V(\underline{x}_1^T \underline{x}_2) = \sigma_w^2$.

By the CLT, $\sqrt{m}(\overline{w} - \underline{\mu}^T \underline{\mu}) \xrightarrow{D} N(0, \sigma_w^2)$.

see HW2E).

So usual t-test and t CI applied to $w_1, \ldots, w_m$ works when the $\underline{x}_i$ are high dimensional.