of $n = \sum_{j=1}^{K} n_j$ cases were correctly classified.

Then the apparent error rate $AER = 1 - \frac{m_A}{n}$.
= prop. of training data misclassified

17) Expect the apparent error rate to be too low; the method works better on the training data than on new test data to be classified.

18) Cross validation (CV): for $i = 1, \ldots, n$ leave case i out, compute discriminant rule, see if case i is correctly classified. Let m_C be the number of cases correctly classified. Then the CV error rate is $1 - \frac{m_C}{n}$. Some techniques have fast formulas for computing the CV error rate.

19) Leave out a subset with enough cases n_V, 10% to 50%, so that a good error estimate can be obtained. compute the discriminant rule on the cases not leftout. Let m_L be the number of leftout cases correctly classified. The "leave a subset out" error rate is $1 - \frac{m_L}{n_V}$.

20) variable selection: find a subset of variables that

does a good job of classification.

forward selection $x_1, \ldots, x_p$ are variables

Step 1) choose $w_1 = x_j$ that minimizes the AER.

2) keep w_1 in the model and add $w_2 = x_j$ that minimizes the AER. So w_1, w_2 are in the model.

k) Have $w_1, \ldots, w_{k-1}$ in model. Add the $w_k = x_j$ that minimizes the AER.

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

w is a permutation of x

backwards elimination = backward elimination

Step 1) $x_1, \ldots, x_p = w_1, \ldots, w_p$ are in the model

Step 2) Delete $w_p = x_j$ such that the model with $p-1$ variables minimizes the AER $w_1, \ldots, w_{p-1}$ are in the model.

3) Delete $w_{p-1} = x_j$ such that the model with $p-2$ variables minimizes the AER. $w_1, \ldots, w_{p-2}$ are in the model.

k) Delete w_{p-k+2} such that the model with $p-k+1$ variables minimizes the AER. $w_1, \ldots, w_{p-k+2}$ are in the model. $w_1, \ldots, w_{p-k+1}$ are in the model.

p) Delete w_2 such that the model with 1 variable minimizes the AER. w_1, w_2 are in the model

21) Other criterion can be used. proc stepdisc in SAS does variable selection

22) LDA seems to be good for variable selection. See text, p 257-7.

23) p121 suppose the distribution of the jth pop is $EC(\underset{\sim}{\mu}_j, \Sigma_j, g)$ where $g\downarrow$ is free of j so the G groups are from the same family of EC distributions (eg all are MVN). Then $f_j(\underset{\sim}{w}) = k_p |\Sigma_j|^{-\frac{1}{2}} g\left[(\underset{\sim}{w}-\underset{\sim}{\mu}_j)^T \Sigma_j^{-1} (\underset{\sim}{w}-\underset{\sim}{\mu}_j)\right]$ and maximizing $f_j(\underset{\sim}{w})$ is equivalent to maximizing

$$\log f_j(\underset{\sim}{w}) = \log(k_p) - \tfrac{1}{2}\log|\Sigma_j| + \log\left(g\left[(\underset{\sim}{w}-\underset{\sim}{\mu}_j)^T \Sigma_j^{-1} (\underset{\sim}{w}-\underset{\sim}{\mu}_j)\right]\right)$$

$$= \log(k_p) - \tfrac{1}{2}\log(|\Sigma_j|) + \log\left(g\left[D^2_{\underset{\sim}{w}}(\underset{\sim}{\mu}_j, \Sigma_j)\right]\right).$$

a) If pop $j \sim N_p(\underset{\sim}{\mu}_j, \Sigma_j)$, $j = 1, \ldots, k$, then $g(u) = \exp(-\frac{1}{2}u)$ and maximizing $f_j(\underset{\sim}{w})$ is equivalent to maximizing $\tilde{Q}_j(\underset{\sim}{w}) = -\frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(\underset{\sim}{w}-\underset{\sim}{\mu}_j)^T \Sigma_j^{-1}(\underset{\sim}{w}-\underset{\sim}{\mu}_j)$. Putting in estimates leads to the quadratic rule with $Q_j(\underset{\sim}{w})$.

b) If $\det(\Sigma_j) = |\Sigma_j| \equiv |\Sigma|$ for $j = 1, \ldots, k$ (eg rotate one pop with an orthogonal matrix), then maximizing $f_j(\underset{\sim}{w})$ is equivalent to minimizing $D^2_{\underset{\sim}{w}}(\underset{\sim}{\mu}_j, \Sigma_j)$. Plugging in estimates leads to the distance rule.

c) $D^2_{\underset{\sim}{w}}(\underset{\sim}{\mu}_j, \Sigma_j) = \underset{\sim}{w}^T \Sigma_j^{-1} \underset{\sim}{w} + \underset{\sim}{\mu}_j^T \Sigma_j^{-1}(-2\underset{\sim}{w} + \underset{\sim}{\mu}_j)$. Hence if $\Sigma_j \equiv \Sigma$ for $j = 1, \ldots, k$, maximizing $f_j(\underset{\sim}{w})$ is equivalent to minimizing $D^2_{\underset{\sim}{w}}(\underset{\sim}{\mu}_j, \Sigma)$ which is equivalent to maximizing $\underset{\sim}{\mu}_j^T \Sigma^{-1}(2\underset{\sim}{w} - \underset{\sim}{\mu}_j)$. Plugging in estimates leads to the linear discriminant rule.

24) p153 Let $k(\underline{z})$ be a pdf. Then

$\frac{1}{h^p} k\left[\frac{1}{h}(\underline{z}-\underline{x}_i)\right]$ is a pdf centered at $\underline{x}_i$ instead of $\underline{0}$ and with $h>0$ decreasing or increasing the spread of the pdf.

ex) $k(\underline{x}) = \frac{1}{(2\pi)^{\frac{p}{2}}} \exp\left(-\frac{1}{2}\underline{x}^T\underline{x}\right)$

is the $N_p(\underline{0}, I)$ pdf.

$\frac{1}{h^p} \frac{1}{(2\pi)^{p/2}} \exp\left[-\frac{1}{2}\frac{1}{h^2}(\underline{z}-\underline{x}_i)^T(\underline{z}-\underline{x}_i)\right]$

is the $N_p(\underline{x}_i, h^2 I)$ pdf

($\Sigma = h^2 I$ so $\Sigma^{-1} = \frac{1}{h^2} I$ and $\sqrt{\det(h^2 I)} = \sqrt{h^{2p}} = h^p$.)

25] So $\frac{1}{n}\frac{1}{h^p}\sum_{i=1}^{n} k\left(\frac{1}{h}(\underline{z}-\underline{x}_i)\right) = \hat{f}(\underline{z})$ is a kernel density estimator of $f(\underline{z})$ found by placing

$\underbrace{\frac{1}{n}\frac{1}{h^p} k\left(\frac{1}{h}(\underline{z}-\underline{x}_i)\right)}_{\text{area} = \frac{1}{n}}$ at each $\underline{x}_i$ and summing

See handout for $p=1$ case.

26] If p is small, $\hat{f}_j(\underline{w})$ can be estimated with a kernel density estimator for $j=1,\dots,k$. Then use the maximum likelihood rule.
Problem: Kernel density estimators are poor for $p>2$.

27] ddiscr and ddiscr2 try to modify distance and ML kernel density rules, but the modifications are not good. Try LDA and QDA before these 2 methods.

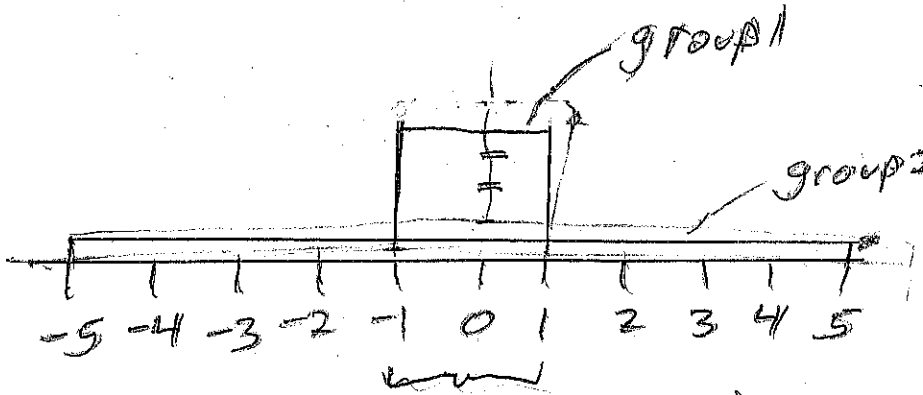
$\mu = \frac{a+b}{2}$ $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

$\mu_1 = \mu_2 = 0$

$\sigma_1 = \sqrt{\frac{4}{12}} = .5774$

$\sigma_2 = \sqrt{\frac{100}{12}} = 2.8868$

28]



distances $Z_1 = \left|\frac{x-\mu_1}{\sigma_1}\right|$ for group 1 are larger than distances $Z_2 = \left|\frac{x-\mu_2}{\sigma_2}\right|$ for group 2.

So distance rule allocates w to group 2 even if $-1<w<1$. Remedy: let $v = Z_1$ and apply distance rule to v. Now rule works well since

U is small for $-1 \le U < 1$ and large for $|U| > 1$.

Ex 1 in § 8.3 is reproduced in HW 7.

end exam 2 material

Ch 9 § 9.1 One sample Hotelling's T^2 test

1) p160 know for final

i) $H_0: \underline{\mu} = \underline{\mu}_0$ $H_A: \underline{\mu} \neq \underline{\mu}_0$

ii) $T_c^2 = n(\bar{\underline{x}} - \underline{\mu}_0)^T \hat{D}^{-1} (\bar{\underline{x}} - \underline{\mu}_0) = n D^2_{\underline{\mu}_0}(\bar{\underline{x}}, \hat{D})$ from output

iii) $pval = P\left(\frac{n-p}{(n-1)p} T_c^2 < F_{p,n-p}\right)$ table or output

$pval = P\left(F_{p,n-p} > \frac{n-p}{(n-1)p} T_c^2\right)$ ← better

iv) reject H_0 if $pval < \alpha$ and conclude $\underline{\mu} \neq \underline{\mu}_0$

fail to reject H_0 if $pval \ge \alpha$, and conclude either $\underline{\mu} = \underline{\mu}_0$ or that there is not enough evidence to conclude that $\underline{\mu} \neq \underline{\mu}_0$.

* Use $\alpha = 0.05$ if α is not given.

2) $\sqrt{n}(\bar{\underline{x}} - \underline{\mu}) \xrightarrow{D} N_p(\underline{0}, \Sigma_{\underline{x}})$ by MCLT.

$\sqrt{n}\,\Sigma_{\underline{x}}^{-\frac{1}{2}}(\bar{\underline{x}} - \underline{\mu}) \xrightarrow{D} N_p(\underline{0}, I_p)$,

so $n(\bar{\underline{x}} - \underline{\mu})^T \Sigma_{\underline{x}}^{-1} (\bar{\underline{x}} - \underline{\mu}) \xrightarrow{D} \chi^2_p$.