olvenotes & James et al notes

01, 1

1) Statistical learning could be defined as the statistical analysis of multivariate data. Machine learning, data mining, big data and predictive analytics are synonymous.

2) 01, 30 A <u>case</u> or <u>observation</u> consists of k random variables measured on one person or thing. The ith case $\underset{\sim}{z}_i = (z_{i1}, \ldots, z_{ik})^T$. The <u>training data</u> consists of $\underset{\sim}{z}_1, \ldots, \underset{\sim}{z}_n$. The statistical model is fit (trained) on the training data. The <u>test data</u> is $\underset{\sim}{z}_{n+1}, \ldots, \underset{\sim}{z}_{n+m}$ and is used to evaluate the quality of the ~~fitted~~ model.

3) 02 The focus of <u>supervised learning</u> is predicting a future value of a response variable Y_f given $\underset{p\times 1}{x_f}$ and training data $(Y_1, \underset{\sim}{x}_1), \ldots, (Y_n, \underset{\sim}{x}_n)$. The focus of

unsupervised learning is to group

$\underline{x}_1, \ldots, \underline{x}_n$ into clusters. Data mining is looking for relationships in large data sets.

ex] Y = college GPA x_1 = high school GPA

x_2 = ACT or SAT score x_3 = gender

marry data

ex] ○ ex 3.2 p96 Y = # of women married to civilians in $n = 26$ districts in Prussia in 1843.

x_1 = constant = trivial predictor

x_2 = pop of district in 1843

x_3 = mmen = # married civilian men in the district

x_4 = mmilmen = # married men in the military in the district

x_5 = milwmn = # women married to husband in the military

$n = 26 > 5p$, $p = 5$. Sometimes the surveyor would not count the spouse if the spouse was not at home but $Y = x_3 + e$ is a good model.

ex] p4 $n = 64$ cancer cell lines from 14 cancer types $p = 6830$, $x_1, \ldots, x_p$ are 6830 gene expression measurements

4)

classical statistics	statistical learning
i) $n \geq 10p$	$n \geq p$ or $n < p$, often $n \geq 10d$ where d = model degrees of freedom
ii) all p variables $x_1, \ldots, x_p$ are used in the model	there is variable selection $x_{i_1}, \ldots, x_{i_d}$ are used
iii) 1 model see ii) eg $Y = \beta_1 + \beta_2 x_3 + e$ for many data	model selection choose 1 model from several eg $Y = \beta_1 + \beta_2 x_3 + e$ $Y = \beta_1 + \beta_2 x_3 + \beta_3 x_2 + e$ $Y = \beta_1 + \beta_2 x_3 + \beta_3 x_2 + \beta_4 x_4 + e$ $Y = \beta_1 + \beta_2 x_3 + \beta_3 x_2 + \beta_4 x_4 + \beta_5 x_5 +$
iv) inference focus in more on hypothesis testing than prediction	inference focus is more on prediction

5) 4 statistical learning principles:

i) there is more interest in prediction or classification eg producing $\hat{Y}_F$

than other types of inference such as parameter estimation, hypothesis testing, confidence intervals or which model fits best.

ii) often the focus is on extracting information

when n/p is not large eg $p > n$.
If d is a crude estimator of fitted model degrees of freedom (df), we want n/d large. A sparse model has few nonzero coefficients and often d = # nonzero coefficients
($\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$; p is large but most $\hat{\beta}_i = 0$.)

iii) Interest is in performance on test data.
Performance on training data, is overly optimistic for estimating performance on training data.

iv) Some methods are flexible while others are unflexible. ($y = \beta_1 + \beta_2 x_2$ is a line
$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3$ is a plane.)
$sp = x'\beta$ is a hyperplane and is unflexible

Flexibility tends to increase with d. Unflexible regression methods are often used when the mean function $E(Y|x) = M(x^T\beta)$ is known up to the p unknown parameters β. Flexible methods tend to be

6} The response variable is the variable of interest; the variable you want to predict. SL p3

The predictors or features $x_1, \ldots, x_p$ are used to predict Y.

7} 04 Regression investigates how Y changes with the $p \times 1$ vector $\underline{x}$ of predictors. Often this conditional distribution $Y|\underline{x}$ is described by a 1D regression model where Y is conditionally independent of $\underline{x}$ given the sufficient predictor $SP = h(\underline{x})$, written

$Y \perp\!\!\!\perp \underline{x} \mid SP$ or $Y \perp\!\!\!\perp \underline{x} \mid h(\underline{x})$ where $h: \mathbb{R}^p \to \mathbb{R}$. The estimated sufficient predictor $ESP = \hat{h}(\underline{x})$.

ex] $Y = m(\underline{x}) + e \quad \to h(\underline{x}) = m(\underline{x}), \quad \hat{h}(\underline{x}) = \hat{m}(\underline{x})$

$Y \sim \text{Poisson}\left[\exp(\underline{\beta}'\underline{x})\right] \to h(\underline{x}) = \underbrace{\underline{\beta}'\underline{x}}_{SP \text{ for a generalized linear model GLM}}, \quad \hat{h}(\underline{x}) = \hat{\underline{\beta}}'\underline{x}$

$Y \sim \text{binomial}\left(1, \dfrac{e^{\alpha + \sum_{j=1}^{p} S_j(x_j)}}{1 + e^{\alpha + \sum_{j=1}^{p} S_j(x_j)}}\right), \quad h(\underline{x}) = \underbrace{\alpha + \sum_{j=1}^{p} S_j(x_j)}_{SP \text{ for a generalized additive model GAM}}$

$\hat{h}(\underline{x}) = \hat{\alpha} + \sum_{j=1}^{p} \hat{S}_j(x_j)$

8) 05 know A response plot is a plot of the ESP $\hat{h}(\underline{x})$ vs. Y. A residual plot is a plot of the ESP vs. the residuals r_i.

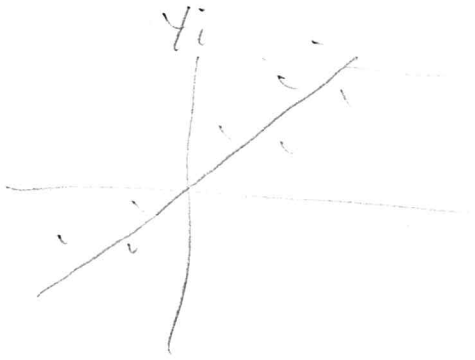
ex] If $Y_i = m(\underline{x}_i) + e_i$ then the ith residual

$$r_i = Y_i - \hat{m}(\underline{x}_i) \qquad i = 1, \ldots, n.$$

(estimated) model mean function and a scatterplot smoother to the response plot.

ex) For $Y = m(\underline{x}) + e$ the identity line (unit slope 0 intercept) is the est. mean function $Y = m(\underline{x})$

$Y = \hat{m}(\underline{x})$ is the identity line in a plot of $\hat{m}(\underline{x})$ vs Y



9) 05 6.2.2 p8 response transformations

Let Z be the variable of interest and let $Y = t(Z) = \underline{x}'\beta + e$ follow the multiple linear regression (MLR) model.

10) 09,10 Assume all values of the "response" Z are positive. The ladder of powers $\Lambda_L = \{-1, -\frac{1}{2}, -\frac{1}{3}, 0, \frac{1}{3}, \frac{1}{2}, 1\}$.

A power transformation has the form

$$Y = t_\lambda(Z) = Z^\lambda \quad \text{for } \lambda \neq 0$$

$$Y = t_0(Z) = \log(Y) \quad \text{for } \lambda = 0.$$

A modified power transformation has the form

$$t_\lambda(Z) = \frac{Z^\lambda - 1}{\lambda} \qquad \lambda \neq 0$$

$$t_0(Z) = \log(Z) \qquad \lambda = 0.$$

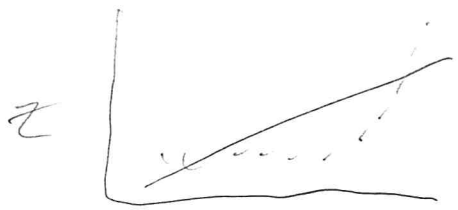
11) For the additive error regression model $Y = m(x) + e$ for MLR $Y = t(Z) = x'\beta + e$, a graphical method for response transformations computes the "fitted values" $\hat{W}_i$ using $W_i = t_\lambda(Z_i)$ as the response. A transformation plot is a plot of $\hat{W}_i$ vs W_i and is made for each of the 7 values of $\lambda \in \Lambda_L$ with the identity line added as a visual aid. Use the transformation

12) $Y = t_{\lambda^*}(Z)$ if the plotted points in the transformation plot for λ^* follow the identity line in a roughly evenly populated band.

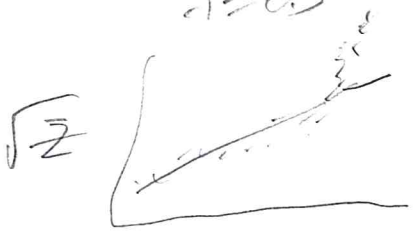
See Oex 1.3 and HW1 C.

13) I know for exam 1 how to pick the transformation from a few plots

$\lambda = 1$

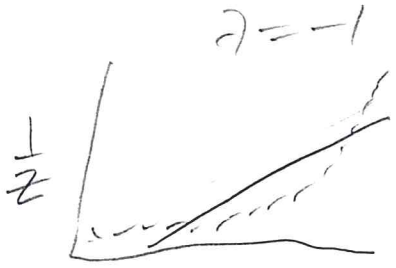


$\lambda = 0.5$



$\lambda = 0$

$\lambda = -1$



pick $Y = \log(Z)$ since this transformation has a linear transformation plot.

13) If more than 1 transformation plot is linear take the simplest or most reasonable transformation or the transformation that

that makes the most sense to subject matter experts. Also check which of the competing transformations has the best "residual plot" of $\hat{w}$ vs $w - \hat{w}$.

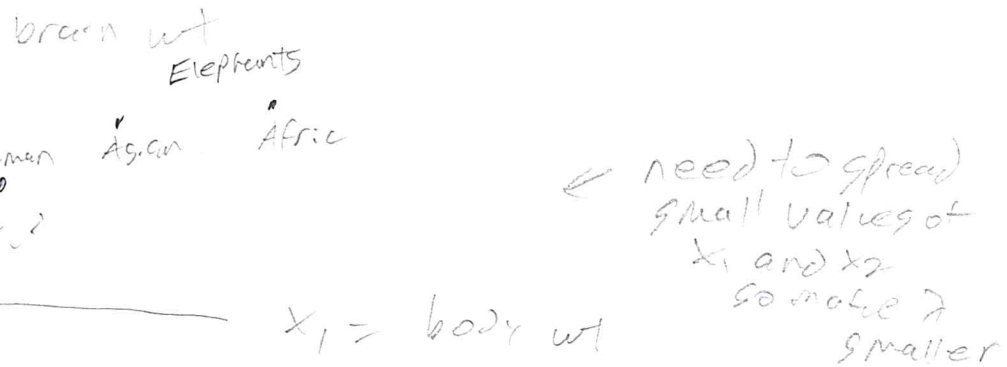
The values of λ in decreasing order of importance are 1, 0, $\frac{1}{2}$, -1 and $\frac{1}{3}$ so a log transformation is preferred over a cube root transformation if both transformation plots look equally good.

143 know for exam 1. The log rule says that a positive variable $W > 0$ with

$\frac{\max(W)}{\min(W)} > 10$ suggests using $\log(W)$ instead of W. This rule removes skew from predictors and can greatly improve the model.

153 know for exam 1 The following rule is used for plots of 2 variables, usually ESP vs Y in this class for response transformations

Ladder rule In a plot of x_1 vs x_2, to spread small values of a variable, make λ smaller. To spread large values of a variable make λ larger.



Both variables are right skewed so there are lots of small values and a few large values

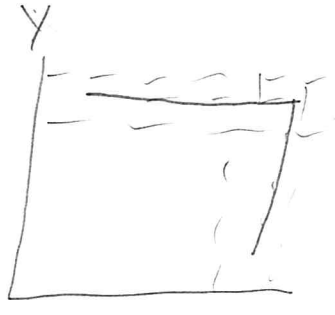
$$\frac{\max x_2}{\min x_2} \approx \frac{6654.2}{0.005} > 10$$

$$\frac{\max x_1}{\min x_1} \approx \frac{5711.9}{0.14} > 10$$

} log rule

ex Fig 1.2 09

small values of w and large values of y need spreading



large values of both variables need spreading

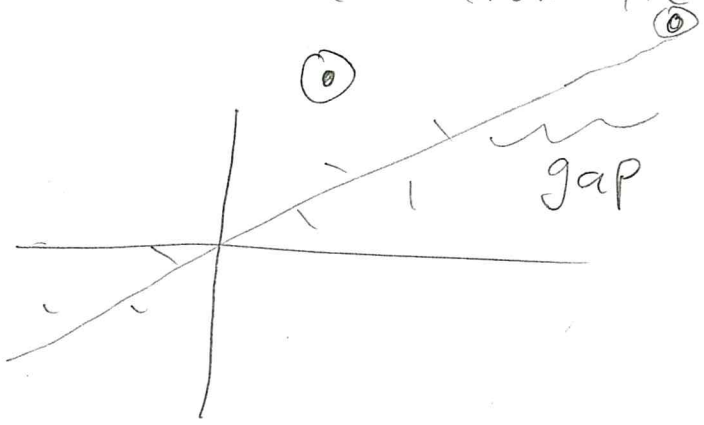
small values of both variables

small values of y and large values of w

0 & 1.3

16] Outliers are cases that lie far away from the bulk of the data and can ruin the statistical analysis.

17] If $Y = X\beta + e$ look for gaps in the response plot and cases far from the identity line.



18] The sample mean $\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n}$.

19] know
Arrange the data set $W_1, \ldots, W_n$ in ascending order. $\underbrace{Y_{(1)}}_{\min} \leq Y_{(2)} \leq \cdots \leq Y_{(n-1)} \leq \underbrace{Y_{(n)}}_{\max}$ are the order statistics.

20] The sample median $MED(Y_1, \ldots, Y_n) =$ $MED(n) = Y_{(\frac{n+1}{2})}$ n odd

$\frac{Y_{(\frac{n}{2})} + Y_{(\frac{n}{2}+1)}}{2}$ n even

21] The sample median absolute deviation

$MAD(Y_1, \ldots, Y_n) = MAD(n) = MED(|Y_i - MED(n)|, i = 1, \ldots, n)$

ex} the median is the middle ordered value or average of the 2 middle values.

1, 2, 3, 4, 5, 6, 7, 8, 9
↑
MED(n) = 5

$|Y_i - MED(n)|$: +4, +3, +2, +1, 0, 1, 2, 3, 4

ordered → 0, 1, 1, 2, 2, 3, 3, 4, 4
↑
MAD(n) = 2

22) Collect the data: $W = \underset{n \times p}{X} = \begin{pmatrix} \underline{x}_1^T \\ \vdots \\ \underline{x}_n^T \end{pmatrix} = (\underline{v}_1 \cdots \underline{v}_p)$

$\underline{x}_i$ ← ith case

$\underline{v}_j$ ← jth variable

$= \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ \vdots & & & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}$

↑ 2nd variable

23) 0.14 The coordinatewise median $MED(W) =$ $(MED(x_1), \ldots, MED(x_n))^T$ where $MED(x_i)$ is the sample median of the ith column

24) Let $\bar{x}$ and

median for a small data set.

see Dixon 1.5 problems (1.1) and (1.2) HW2 see notes 6.5

25) 015 The sample covariance S_{ij} estimates the pop covariance $\sigma_{ij} = E[(X_i - E(X_i))(X_j - EX_j)]$ and

$$S_{ij} = \frac{1}{n-1} \sum_{k=1}^{n} (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j).$$

The sample variance $S_{ii} = S_i^2$, $\sigma_{ii} = \sigma_i^2 =$ pop var.

The sample correlation r_{ij} estimates the pop corr $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$ and $r_{ij} = \frac{S_{ij}}{S_i S_j} = \frac{S_{ij}}{\sqrt{S_{ii} S_{jj}}}$.

26) 015-16 Let $\underset{p\times 1}{\underline{x}_1}, \ldots, \underline{x}_n$ be the data.

The sample covariance matrix $\underset{p\times p}{\underline{S}} = (S_{ij})$.

The classical estimator of multivariate location and dispersion (MLD) is $(T, C) = (\bar{x}, S)$. The sample correlation matrix $\underline{R} = (r_{ij})$.

27) 017 know the ith Mahalanobis distance $D_i = \sqrt{D_i^2}$ where $D_i^2 = D_i^2(T(W), C(W)) =$

$$(\underline{x}_i - T(W))^T C^{-1}(W)(\underline{x}_i - T(W))$$

$$= (\underline{x}_i - T)^T C^{-1} (\underline{x}_i - T) = D^2_{\underline{x}_i}(T, C).$$

$$D^2_{\underline{x}}(T, C) = (\underline{x} - T)^T C^{-1} (\underline{x} - T)$$

↑ center of hyperellipsoid

← eigenvectors determine axes of hyperellipsoids

28] Note that $D^2_T(\underline{x}, C) = D^2_{\underline{x}}(T, C)$.

29] The pop (squared) Mahalanobis distance

$$D^2_{\underline{x}}(\mu, \Sigma) = (\underline{x} - \mu)^T \Sigma^{-1} (\underline{x} - \mu)$$ where

μ = pop location vector and Σ = pop dispersion matrix, often the pop cov. matrix.

30] The squared Euclidean distance of $\underline{x}$ from T is $(\underline{x} - T)^T (\underline{x} - T) = D^2_{\underline{x}}(T, I_p)$ where $I_p = \text{diag}(1, \ldots, 1)$ is the $p \times p$ identity matrix.

31] § 13.3 Outlier detection if $p > n$; more than $\frac{n}{2}$ cases are in the bulk of the data.

a) Use Euclidean distances from the coordinatewise median $D_i(MED(W), I_p)$.

ex 4 measurements on 5 trees

SL 6.5

N	E	S	W
72	66	76	77
60	53	66	63
56	57	64	58
41	29	36	38
32	32	35	36
Σ 261	237	277	272

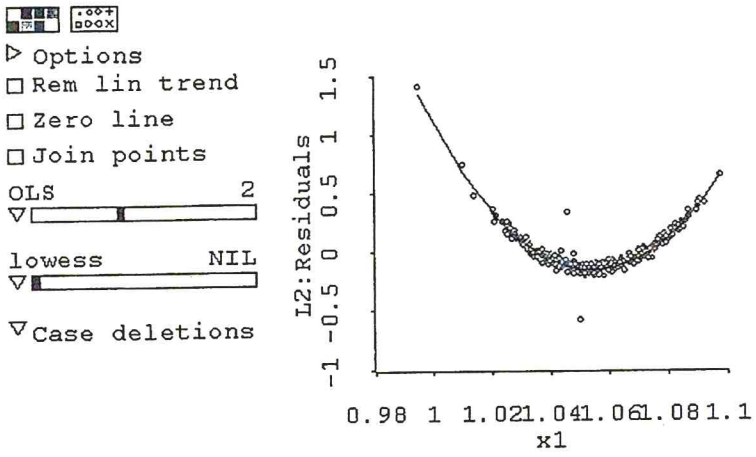
ordered

56	60	72	N
53	57	66	E
64	66	76	S
58	63	77	W
	MED (10)		

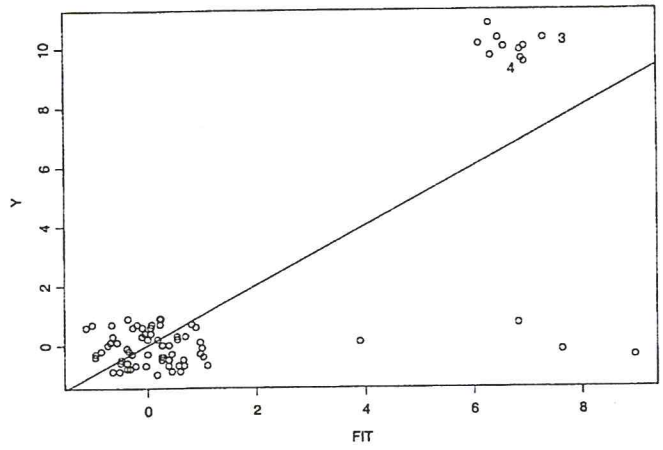
32	41
29	32
35	36
36	38

show work →

$$\frac{1}{5}\begin{pmatrix}261\\237\\277\\272\end{pmatrix} = \begin{pmatrix}52.5\\47.4\\55.4\\54.4\end{pmatrix} = \bar{x}$$


Options
Rem lin trend
Zero line
Join points
OLS
2
lowess
NIL
Case deletions
I2:Residuals
-1 -0.5 0 0.5 1 1.5
0.98 1 1.02 1.04 1.06 1.08 1.1
x1

a)


Y
0 2 4 6 8 10
0 2 4 6 8
FIT
3
4

b)

b) Let MED_j be the coordinatewise median computed from cases $\underline{x}_i$ with $D_i^2 \leq MED\left(D_i^2(MED_{j-1}, I_p)\right)$ where $MED_0 = MED(W)$. Often use $j=0$ or $j=9$. Let $D_i = D_i(MED_j, I_p)$. Let $W_i = \begin{cases} 1 & \text{if } D_i < MED(D_1, \ldots, D_n) + 5\,MAD(D_1, \ldots, D_n) \\ 0 & \text{else.} \end{cases}$

The coumbz set B consists of the $m \geq \frac{n}{2}$ cases with weight $W_i = 1$. The coumbz estimator (T, C) has $T = \dfrac{\sum_{i=1}^{n} W_i \underline{x}_i}{\sum_{i=1}^{n} W_i}$, $C = \dfrac{\sum_{i=1}^{n} W_i (\underline{x}_i - T)(\underline{x}_i - T)^T}{\sum_{i=1}^{n} W_i - 1}$

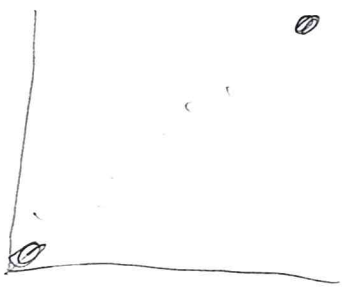
which is the sample mean and sample covariance matrix computed from the cases in set B.

c) Let $\underline{w}_i = (Y_i, \underline{x}_i)$ and let the continuous predictors from $\underline{x}_i$ be $\underline{v}_i$ (the predictors that take on many values, so not gender). Apply the regression method to the m cases $\underline{w}_i$ corresponding to the coumbz set $B = \{i_1, \ldots, i_m\}$ applied to $\underline{v}_1, \ldots, \underline{v}_n$, MLR, GLMs, GAMs, LDA, QDA, kNN etc

33) 014 The function ddplot5 plots

$D_i(MED(W), I_p)$ vs $D_i(T_{COUMBZ}, I_p)$.

The plotted points tend to cluster about the identity line with outliers in the upper right corner of the plot with a gap between the bulk of the data and the outliers.



Buxton data

34) To detect outliers in one group and the bulk of the data in another group, the distance of the outliers from the bulk of the data increases roughly with $\sqrt{p}$.

0.6 1.4 Large Sample Theory Stem

35) Know 024 Central Limit Theorem (CLT):
Let $Y_1, \ldots, Y_n$ be iid with $E(Y_1) = \mu$ and $V(Y_1) = \sigma^2$.
Then $\sqrt{n}(\overline{Y}_n - \mu) \xrightarrow{D} N(0, \sigma^2)$.

36) If $W_n \xrightarrow{D} X$, then X is the asymptotic distribution or limiting distribution of W_n. X does not depend on n. The approximate

$\bar{Y} \approx N(\mu, \frac{\sigma^2}{n})$, which does depend on n.

37] 027 Let $\{Z_n \; n=1,2,\ldots\}$ be a sequence of RVs with CDFs F_n and let X be a RV with CDF F. Then Z_n converges in distribution to X, written $Z_n \xrightarrow{D} X$, if $\lim_{n\to\infty} F_n(t) = F(t)$ at each continuity point of F.

38] If $X_n \xrightarrow{D} X$, then $P(a < X_n \leq b) = F_n(b) - F_n(a)$

$\rightarrow F(b) - F(a) \doteq P(a < X \leq b)$ if F is continuous at a and b. So F can be used to approximate probabilities and percentiles.

See ex's 1.13 and 1.14.

39] 029 X_n converges in probability to X,

$X_n \xrightarrow{P} X$, if $\lim_{n\to\infty} P(|X_n - X| \geq \varepsilon) = 0 \quad \forall \varepsilon > 0$

$\iff \lim_{n\to\infty} P(|X_n - X| < \varepsilon) = 1 \quad \forall \varepsilon > 0.$

40] A sequence of estimators T_n is consistent for $\tau(\theta)$ if $T_n \xrightarrow{P} \tau(\theta) \; \forall \theta \in \Theta$. Then T_n is a consistent estimator of $\tau(\theta)$,

41) i) If $\lim_{n\to\infty} V_\theta(T_n) = 0$ and $\lim_{n\to\infty} E_\theta(T_n) = \tau(\theta)$ both $\forall \theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

ii) Let $0 < \delta \leq 1$. If $n^\delta (T_n - \tau(\theta)) \xrightarrow{D} N(0, \sigma(\theta))$, $\forall \theta \in \Theta$, then T_n is a consistent estimator of $\tau(\theta)$.

42) $X_n \xrightarrow{P} \tau(\theta)$ iff $X_n \xrightarrow{0} \tau(\theta)$.

43) If $n^\delta (W_n - \tau) \xrightarrow{D} X$ for some nondegenerate RV X, then W_n has rate n^δ, and

$n^{1/2}$ rate = square root n consistency is good.

Multivariate Limit Th's

44) 038 Let $\underset{\sim}{X}_n$ have joint CDF $F_n(\underset{\sim}{x})$ and $\underset{\sim}{X}$ have joint CDF $F(\underset{\sim}{x})$.

a) $\underset{\sim}{X}_n \xrightarrow{D} \underset{\sim}{X}$ if $F_n(\underset{\sim}{x}) \to F(\underset{\sim}{x})$ at all continuity points $\underset{\sim}{x}$ of F.

b) $\underset{\sim}{X}_n \xrightarrow{P} \underset{\sim}{X}$ if $\forall \varepsilon > 0$, $P(\|\underset{\sim}{X}_n - \underset{\sim}{X}\| > \varepsilon) \to 0$

45) know | Multivariate CLT (MCLT)

If $\underset{\sim}{X}_1, \ldots, \underset{\sim}{X}_n$ are iid $k \times 1$ random vectors with $E(\underset{\sim}{X}) = \underset{\sim}{\mu}$ and $\text{cov}(\underset{\sim}{X}) = \Sigma_{\underset{\sim}{X}}$, then

$$\sqrt{n}\left(\overline{\underset{\sim}{X}}_n - \underset{\sim}{\mu}\right) \xrightarrow{D} N_k\left(\underset{\sim}{0}, \Sigma_{\underset{\sim}{X}}\right).$$

46) a) If estimator $\underset{\sim}{g}\left(\underset{\sim}{T}_n\right) \xrightarrow{P} \underset{\sim}{g}(\underset{\sim}{\theta})$ $\forall\, \underset{\sim}{\theta} \in \Theta$, then $\underset{\sim}{g}\left(\underset{\sim}{T}_n\right)$ is a consistent estimator of $\underset{\sim}{g}(\underset{\sim}{\theta})$.

b) If $n^{\delta}\left(\underset{\sim}{g}\left(\underset{\sim}{T}_n\right) - \underset{\sim}{g}(\underset{\sim}{\theta})\right) \xrightarrow{D} \underset{\sim}{X}$, then $\underset{\sim}{g}\left(\underset{\sim}{T}_n\right) \xrightarrow{P} \underset{\sim}{g}(\underset{\sim}{\theta})$.

c) If $\underset{\sim}{X}_n \xrightarrow{P} \underset{\sim}{X}$ then $\underset{\sim}{X}_n \xrightarrow{D} \underset{\sim}{X}$.

d) $\underset{\sim}{X}_n \xrightarrow{P} \underset{\sim}{g}(\underset{\sim}{\theta})$ iff $\underset{\sim}{X}_n \xrightarrow{D} \underset{\sim}{g}(\underset{\sim}{\theta})$.

47) 040 Continuous Mapping Th. Let g be a continuous function $g: \mathbb{R}^k \to \mathbb{R}^d$. If $\underset{\sim}{X}_n \xrightarrow{D} \underset{\sim}{X}$, then $g\left(\underset{\sim}{X}_n\right) \xrightarrow{D} g\left(\underset{\sim}{X}\right)$.

48) 041 know a) If $\sqrt{n}\left(\underset{\sim}{T}_n - \underset{\sim}{\mu}\right) \xrightarrow{D} N_p\left(\underset{\sim}{0}, \Sigma\right)$ and A is a $q \times p$ constant matrix, then

$$A\sqrt{n}\left(\underset{\sim}{T}_n - \underset{\sim}{\mu}\right) \xrightarrow{D} N_q\left(A\underset{\sim}{0}, A\Sigma A^T\right)$$

b) Let $\Sigma > 0$ be positive definite.

If $\sqrt{n}(T_n - \mu) \xrightarrow{D} N_p(0, \Sigma)$ and if

C_n is a consistent estimator of Σ, then

$$D^2_{T_n}(\mu, C_n)$$

$$n(T_n - \mu)^T C_n^{-1} (T_n - \mu) \xrightarrow{D} \chi^2_p$$

c) If $\Sigma > 0$, $T_n \xrightarrow{P} \mu$ and $C_n \xrightarrow{P} \Sigma$ then

$$D^2_x(T_n, C_n) = (x - T_n)^T C_n^{-1} (x - T_n) \xrightarrow{D} D^2_x(\mu, \Sigma)$$

$$= (x - \mu)^T \Sigma^{-1} (x - \mu).$$

So $D^2_{(\lceil n(1-\alpha) \rceil)} \xrightarrow{P} D^2_{x, 1-\alpha}(\mu, \Sigma)$ where $P\left(D^2_x(\mu, \Sigma) \le D^2_{x, 1-\alpha}\right)$ often $\chi^2_{p, 1-\alpha}$

$= 1 - \alpha.$

Ch 2 James et al §5.2 Prediction Regions and Bootstrap

13 047 Consider predicting a future test value Y_f given a $p \times 1$ vector x_f and training data $(y_1, x_1), \ldots, (y_n, x_n)$. A large sample $100(1-\delta)\%$ prediction interval PI for Y_f has the form

$[\hat{L}_n, \hat{U}_n]$ where $P[\hat{L}_n \le \gamma_k \le \hat{U}_n] \to 1-\delta$ as $n \to \infty$. (SL 11)

A large sample $100(1-\delta)\%$ confidence interval for an unknown parameter θ is $[\hat{L}_n, \hat{U}_n]$ where $P[\hat{L}_n \le \theta \le \hat{U}_n] \to 1-\delta$ as $n \to \infty$.

2] 05/1,55 Consider predicting a $p \times 1$ future test vector $\underline{x}_f$ or estimating an $r \times 1$ unknown parameter vector $\underline{\mu}$ given training data $\underline{x}_1, \ldots, \underline{x}_n$.

A large sample $100(1-\delta)\%$ prediction region is a set A_n such that $P(\underline{x}_f \in A_n) \to 1-\delta$ as $n \to \infty$.

A large sample $100(1-\delta)\%$ confidence region for $\underline{\mu}$ is a set A_n such that $P(\underline{\mu} \in A_n) \to 1-\delta$ as $n \to \infty$.

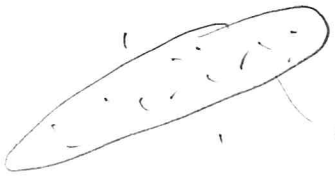
A CI and PI are special cases where $p=r=1$.

3] A large sample hypothesis test for $H_0: \underline{\mu} = \underline{\mu}_0$ vs. $H_A: \underline{\mu} \neq \underline{\mu}_0$ (with asymptotic level δ) rejects H_0 if $\underline{\mu}_0 \notin A_n$ and fails to reject H_0 if $\underline{\mu}_0 \in A_n$ a $100(1-\delta)\%$ confidence region for $\underline{\mu}$.

4] Interpretation Collect data, form $A_{n,i}$ and see if $W_i \in A_{n,i}$ for $i = 1, \ldots, k$ where the k

trials are independent from the same pop and $\underline{w}_i = \underline{x}_i$ or $w_i = \underline{\theta}$. Then the number of times $\underline{w}_i \in A_{ni} \sim bin(k, 1-\delta_n)$ where $\delta_n \to \delta$ as $n \to \infty$. So for a 95% region and $k = 100$, expect about 95 of the $\underline{w}_i \in A_{ni}$.

5) The volume of a confidence region $\to 0$ as $n \to \infty$ while the volume of a prediction region goes to that of a population region that would contain a new $\underline{x}_f$ with prob $1-\delta$.



95% pop prediction region

ex) Consider predicting the height of the next person to come through the door. [4ft, 7ft] $\approx$ 100% PI.

[5ft, 6ft] might be an 80% PI since the class has some short ladies.

6) 08H 100(1-δ)% coverage is the nominal coverage. If the actual coverage $100(1-\delta_n) > 100(1-\delta)$, the region is conservative if $100(1-\delta_n) < 100(1-\delta)$ the region is liberal. For a nominal one

being 5% conservative is "much better" than being 5% liberal. Want small volume with coverage near or higher than the nominal coverage.

7) In simulations with 5000 runs, simulated coverage (prop times $W_i \in A_{n,i}$) $\in [0.94, 0.96]$ suggests actual coverage is close to the nominal 95% coverage.

8) The bootstrap generates pseudodata $T_1^*, \ldots, T_B^*$ for a statistic T_n that estimates a parameter μ, such that, under regularity conditions, applying certain large sample $100(1-\delta)\%$ prediction regions to the bootstrap sample results in large sample $100(1-\delta)\%$ confidence regions.

9) Let $Z_1, \ldots, Z_n$ (training Z_i), Z_t (test) be iid. The shortest interval is the shortest closed interval containing at least C of the Z_i. For a small data set order the data and compute the lengths of intervals containing $= C$ cases!

$\{[\bar{z}_{(1)}, z_{(c)}], [\bar{z}_{(2)}, z_{(c+1)}], \ldots, [\bar{z}_{(n-c+1)}, z_{(n)}]\}$

$z_{(c)} - z_{(1)} \quad z_{(c+1)} - z_{(2)} \quad z_{(n)} - z_{(n-c+1)}$

$\text{Shorth}(c) = [z_{(s)}, z_{(s+c-1)}]$ is the interval with shortest length.

Find shorth(4)

ex) 0, 1, 3, 6, 9, 10, 11

$6-0=6$

$9-1=8$

$10-3=7$

$11-6=5$

$[6, 11] = \text{Shorth}(4)$

Let $\lceil x \rceil$ be the smallest integer $\geq x$. So $\lceil 7.7 \rceil = 8$, $\lceil 7 \rceil = 7$.

10] If $\frac{c_n}{n} \to 1-\delta$, the Shorth($c_n$) interval is a large sample $100(1-\delta)\%$ PI, $c_n \geq k_n = \lceil n(1-\delta) \rceil$ contains $\geq 100(1-\delta)\%$ of the training data so has $1-\delta_n < 1-\delta$ for test data. Frey (2013) showed the max undercoverage $\simeq 1.12\sqrt{\frac{\delta}{n}}$ and used $c_n = \min\left[n, \left\lceil n\left[1-\delta+1.12\sqrt{\frac{\delta}{n}}\right]\right\rceil\right]$

↓

11] The Shorth(c_n) PI estimates the $100(1-\delta)\%$ highest density region if the x_i are iid from a unimodal (bimodal) its shorth

12} 052 Let $\underbrace{x_1, \ldots, x_n}_{\text{training}}$ $\underbrace{x_f}_{\text{test}}$ be iid from a distribution with nonsingular cov. matrix Σ_x.

A large sample $100(1-\delta)\%$ nonparametric prediction region for x_f is

$$A_n = \{\bar{x} : D_x^2(\bar{x}, S) \le D_{(c)}^2\} =$$

$$\{\bar{x} : D_x(\bar{x}, S) \le D_{(c)}\} \text{ where}$$

$c = c_n \downarrow \lceil n(1-\delta) \rceil$ as $n \to \infty$.

↙ increase training data coverage by 5% if $n \le 20p$

Let $q_n = \begin{cases} \min(1-\delta+0.05, 1-\delta+\frac{p}{n}) & \delta \ge 0.1 \\ \min(1-\frac{\delta}{2}, 1-\delta-\frac{10\delta p}{n}) & \delta \le 0.1 \end{cases}$

↖ 95% PR uses 97.5% coverage of training data if $n \le 20p$

If $1-\delta < 0.999$ and $q_n < 1-\delta+0.001$ set $q_n = 1-\delta$.

Then $D_{(c)}$ is the $100 q_n$th quantile of $D_1, \ldots, D_n$

eg use $c = \lceil n q_n \rceil$. (*)

13} 049 A large sample PI for Y_f when $Y_i = m(x_i) + e_i$ finds the shorth(c) of the residuals $[r_{(s)}, r_{(s+c-1)}]$.

Then the PI is $[\hat{m}(x_f) + b_n r_{(s)}, \hat{m}(x_f) + b_n r_{(s+c-1)}]$ where

$$b_n = \begin{cases} \left(1+\frac{15}{n}\right)\sqrt{\frac{n+2d}{n-d}} & d < \frac{8n}{9} \\ 5\left(1+\frac{15}{n}\right) & d \geq \frac{8n}{9} \end{cases}$$

and d = crude estimator of model df.

Here $c = \lceil n g_n \rceil$ and g_n replaces ρ by d in (*).
This PI roughly uses the shorth of the pseudodata $\hat{Y}_f + r_i$, $i = 1, \ldots, n$.
§7.3 Bootstrapping James et al §5.2

(4) Nonparametric bootstrap: Let data $Z_1, \ldots, Z_n$ be iid and $T_n = t(Z_1, \ldots, Z_n)$. Draw a sample of given $Z_1^*, \ldots, Z_n^*$ with replacement from $Z_1, \ldots, Z_n$, compute $T_{n,1}^* = t(Z_1^*, \ldots, Z_n^*)$ and repeat B times to generate

$T_{n,1}^*, \ldots, T_{n,B}^*$.

(5) Know for EI) For any B given B bootstrap samples compute $T_1^*, \ldots, T_B^*$. Often $T_n = \bar{Y}_n$ or MED(n)

ex) range = $x_{(n)} - x_{(1)} = T$. $\{x_1, x_2\} = \{0, 4\}$

$T_n = 4 - 0 = 4$

bootstrap sample x_i^*	T_i^*
(0,0)	0
(0,4)	4
(4,0)	4

Problem 0.2.4

ex] $x_1, \ldots, x_n = 1, 2, 5, 10, 50$ $\quad MED(n) = 5 = T_n$

bootstrap sample	ordered	
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	$T_3^* = 2$

ex] If $(Y_i, \underline{x}_i)$ are iid with $Y_i = \underline{x}_i^T \underline{\beta} + e_i$

draw bootstrap sample, compute $\hat{\beta}_j^*$ $j = 1, \ldots, B$,

for the nonparametric bootstrap.

16] D63 An alternative to the above ex is the residual bootstrap. The ith residual

$r_i = Y_i - \hat{Y}_i = Y_i - \underline{x}_i^T \hat{\underline{\beta}}$.

In matrix form $\underset{\sim}{Y} = \underset{\sim}{X}\underline{\beta} + \underline{e}$. Regress $\underline{Y}$ on $\underline{X}$

to obtain $\underset{p\times 1}{\hat{\beta}}$, $\underset{n\times 1}{\underset{\sim}{r}}$, $\underset{n\times 1}{\underset{\sim}{\hat{Y}}}$. Draw a sample

of size n with replacement from the residuals

$r_1^*, \ldots, r_n^*$ to form vector $\underline{Y}_j^*$ with ith

element $Y_{ij}^* = \hat{Y}_i + r_i^*$. For $j = 1, \ldots, B$

regress $\underline{Y}_j^*$ on $\underline{X}$ to form $\hat{\underline{\beta}}_1^*, \ldots, \hat{\underline{\beta}}_B^*$.

The residual bootstrap works well with least squares

17) OSG If $D_{\underline{\mu}}(T_n, \hat{*}_T) \xrightarrow{D} D^2$ and $D_{\underline{\mu}}(T_n, \hat{*}_T) \xrightarrow{D} D^2$ and $\hat{D}^2_{1-\delta} \xrightarrow{P} D^2_{1-\delta}$ then $R_C =$

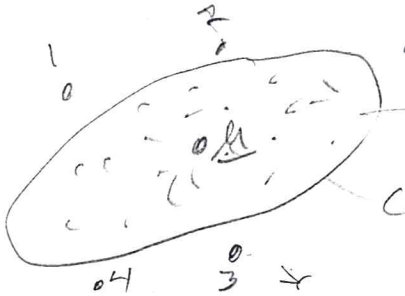
$\{\underline{w} : D^2_{\underline{w}}(T_n, \hat{*}_T) \leq \hat{D}^2_{1-\delta}\}$ is a large sample $100(1-\delta)\%$ cont. reg. for $\underline{\mu}$, and if $\underline{\mu}$ is known $R_P = \{\underline{w} : D_{\underline{w}}(\underline{\mu}, \hat{*}_T) \leq \hat{D}^2_{1-\delta}\}$ is a large sample $100(1-\delta)\%$ prediction region for a future value of the statistic T_{En}. Region R_C contains $\underline{\mu}$ iff region R_P contains T_n.

Problem usually $B=1$ is not large

18) Idea: suppose there is an iid sample $T_{1n}, \ldots, T_{Bn}$ of size B of the statistic. A large sample $100(1-\delta)\%$ prediction region for T_{En} is

$\{\underline{w} : D^2_{\underline{w}}(\underline{\mu}, \hat{*}_T) \leq D^2_{(C)}\}$ if $\underline{\mu}$ is known.

$D^2_{T_i}(\underline{\mu}, \hat{*}_T) = D^2_{\underline{\mu}}(T_i, \hat{*}_T)$.

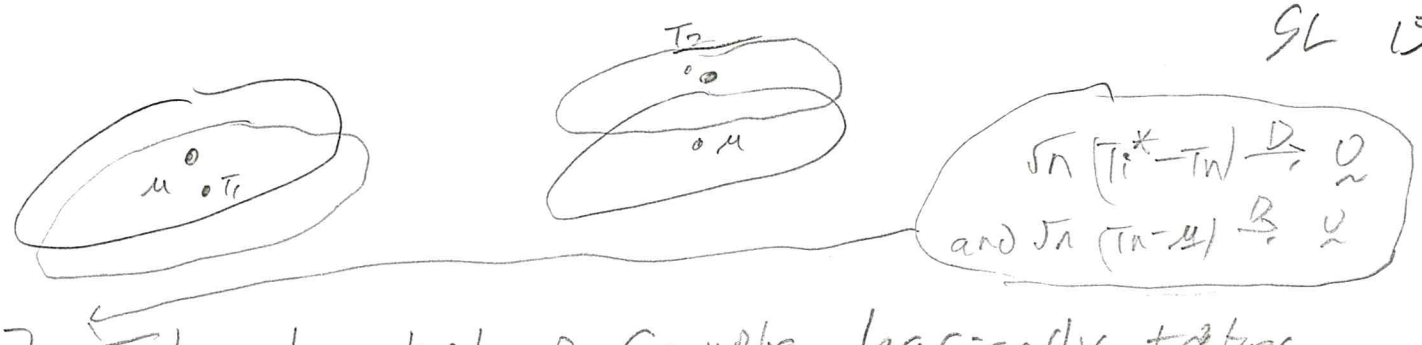


95 values of T_i in the pred region

Contains $\approx 95\%$ of the T_i if n and B are large

95 values are in 3 are out

So $\underline{\mu} \in \{\underline{w} : D^2_{\underline{w}}(T_i, \hat{*}_T) \leq D^2_{(C)}\}$ iff $T_i \in \{\underline{w} : D^2_{\underline{w}}(\underline{\mu}, \hat{*}_T) \leq D^2_{(C)}\}$



19] The bootstrap sample basically takes the cloud of T_i centered at $\underline{\mu}$ and shifts the cloud to be centered at T_n.

20] Need $B \geq 50p$ if $\underline{\mu}$ is $p \times 1$ and need n large, sometimes extremely large. If n is not large enough, undercoverage ($<$ nominal $100(1-\delta)\%$ coverage) often occurs.

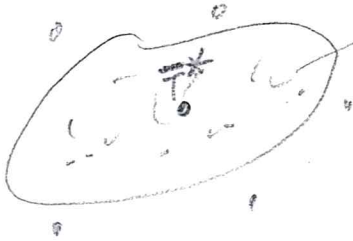
21] (b) prediction region method to test

$H_0: \underline{\mu} = \underline{c}$ vs $H_A: \underline{\mu} \neq \underline{c}$.

($\underline{\mu}$ is $r \times 1$)

make a bootstrap sample $\underset{\sim}{W}_i = \underline{\hat{\mu}}_i^* - \underline{c}$ for $i = 1, \ldots, B$. Apply the nonparametric prediction region to get a large sample $100(1-\delta)\%$ confidence region for $\underline{\mu} - \underline{c}$. Reject H_0 if $\underline{0}$ is not in the region. Fail to reject H_0 if $\underline{0}$ is in the region. If $r = 1$ apply the shorth(cn) PI to get a large sample $100(1-\delta)\%$ CI. For large n the 95% con't region covers $\sim$ 95%

of the $\underline{W}_i$ and $\simeq 5\%$ are not covered



$\simeq 99$ in 3 oct

22) 068 the large sample $100(1-\delta)\%$ confidence region for $\underline{\mu}$ is $\left\{\underline{w}: D^2(\underline{T}^*, S_T^*) \leq D^2_{(U_B)}\right\} = R$ where

$D^2_{(U_B)}$ is computed from $D_i^2 = (\underline{T}_i^* - \underline{T}^*)^T \left[S_T^*\right]^{-1} (\underline{T}_i^* - \underline{T}^*)$

for $i = 1, \ldots, B$ and $D^2_{(U_B)}$ is the $100(1-\delta)$th sample quantile of the D_i^2. Reject H_0 if $\underline{\mu_0} \notin R$.

23) Sufficient conditions:

$\underline{x}_1, \ldots, \underline{x}_n$ iid $T_n = T(\underline{x}_1, \ldots, \underline{x}_n)$ for nonparametric ~~bootstrap~~

$r_1, \ldots, r_n$ from a good estimator $\hat{\underline{\beta}}$ for the residual bootstrap where $\hat{\underline{\mu}} = A\hat{\underline{\beta}} = T_n$, $\underline{\mu} = A\underline{\beta}$.

So far only proven if $\sqrt{n}(T_n - \underline{\mu}) \xrightarrow{D} N_r(\underline{0}, \Sigma_T)$ $\Sigma_T > 0$. Conjectured to work for $\sqrt{n}(T_n - \underline{\mu}) \xrightarrow{D} X$ for many other distributions for X.

Sketch 2.3 we will come back to 0 § 2.3, 3 later.

end of material

exam 2 material

Multiple Linear Regression MLR

SL 16

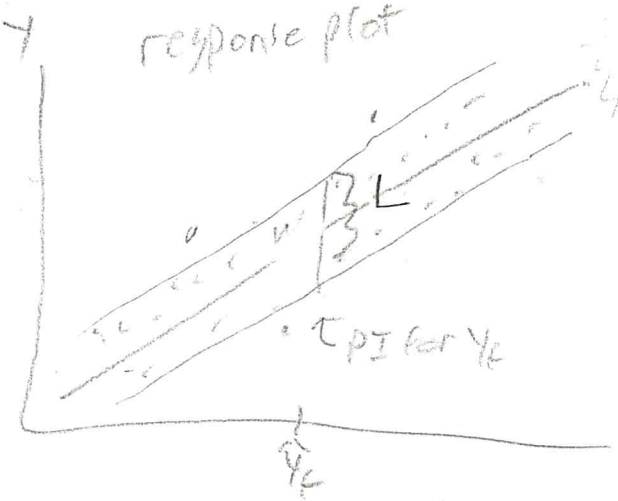
Ch 3 Ch 3, 6 part of 5

1) Also see Olive (2010, ch 2) with URL given on p. 268 (M484 material)

2) Olive: Prediction interval bands for $\frac{n}{d}$ large

pointwise

≈ 2 lines parallel to the identity line covering $100(1-\delta)\%$ of the data with smallest vertical length L

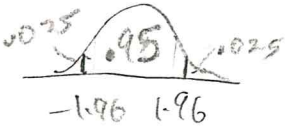


≈ 97 in 3 cases for 95% PI

for $n = 100$, a bit more than 95% of the training data are within the bands

$\hat{Y} = \underline{x}^T \underline{\hat{\beta}}$

$e_i \sim N(0,1) \Rightarrow L = 2(1.96) = 3.92$ as $\frac{n}{d} \to \infty$



3) p. 71 MLR $Y_i = \underline{x}_i^T \underline{\beta} + e_i$ ← text uses Σ_i

$$Y_1 = x_{11}\beta_1 + \cdots x_{1p}\beta_p + e_1$$
$$\vdots$$
$$Y_n = x_{n1}\beta_1 + \cdots + x_{np}\beta_p + e_n$$

or $\underset{n\times 1}{\underline{Y}} = \underset{n\times p}{\underline{X}}\,\underset{p\times 1}{\underline{\beta}} + \underset{n\times 1}{\underline{e}}$

4) p69 The ordinary least squares OLS full model

dummy variable not β from $Y = X\beta + e$

$\hat{\beta}_{OLS}$ minimizes $Q_{OLS}(\beta) = \sum_{i=1}^{n} r_i^2(\beta) = RSS(\beta)$

$= (Y - X\beta)^T (Y - X\beta)$ estimating equations

$\hat{\beta}_{OLS}$ estimates β from $Y = X\beta + e$

Good criterion if $n > np$.

Analogy maximum likelihood estimator

Y_i iid with pdf $f(y|\theta)$ (parameter)

$L(\theta) = \prod_{i=1}^{n} f(y_i|\theta)$

dummy variable but $\hat{\theta} = \arg\max L(\theta)$ estimates parameter θ

5) OLS LS CLT: $Y_i = x_i^T\beta + e_i$, the zero mean errors e_i are iid with $V(e_i) = \sigma^2$.

Assume p is fixed and $n \to \infty$, $\max(h_1, \ldots, h_n) \to 0$

$h_i = h_{ii}$ where $H = X(X^TX)^{-1}X^T = (h_{ij})$, and

$\frac{X^TX}{n} \to V^{-1}$ as $n \to \infty$ Then

$\sqrt{n}(\hat{\beta}_{OLS} - \beta) \xrightarrow{D} N_p(0, \sigma^2 V)$ and

$\frac{1}{\sqrt{n}}(\hat{\beta}_{OLS} - \beta) \xrightarrow{D} N_p(0, \sigma^2 I_p)$

$\frac{\underset{\sim}{X}^T\underset{\sim}{X}}{n}$ estimates V^{-1} so

$\left(\frac{\underset{\sim}{X}^T\underset{\sim}{X}}{n}\right)^{-1} = n(\underset{\sim}{X}^T\underset{\sim}{X})^{-1}$ estimates V

$\hat{\underset{\sim}{\beta}}_{OLS} \sim AN_p(\underset{\sim}{\beta}, MSE(\underset{\sim}{X}^T\underset{\sim}{X})^{-1})$

where $MSE = \frac{1}{n-p}\sum_{i=1}^{n} r_i^2$.

$\left(\sqrt{n}(\hat{\underset{\sim}{\beta}}_{OLS} - \underset{\sim}{\beta}) \sim AN_p\{\underset{\sim}{0}, \sigma^2 n(\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\}\right.$ multiply by $\frac{1}{\sqrt{n}}$ so

$\left.\hat{\underset{\sim}{\beta}}_{OLS} - \underset{\sim}{\beta} \sim AN_p\{\underset{\sim}{0}, (\underset{\sim}{X}^T\underset{\sim}{X})^{-1}\}\right)$

6) Let the nontrivial predictors $\underset{\sim}{U}_i = (x_{i2}, \ldots, x_{ip})^T$.

Let the $n \times (p-1)$ matrix of standardized nontrivial predictors be $\underset{\sim}{W} = (w_{ij})$ where

$\sum_{i=1}^{n} w_{ij} = 0$ and $\sum_{i=1}^{n} w_{ij}^2 = n$

$\sum_{i=1}^{n} w_{ij} = 0$: jth standardized predictor has sample mean $= 0 = \bar{w}_j$

so $\tilde{\sigma}_j^2 = \frac{1}{n}\sum_{i=1}^{n} w_{ij}^2 = 1$ ↑ biased sample variance.

Then the sample correlation matrix of the nontrivial predictors is $R_{\underset{\sim}{U}} = \frac{\underset{\sim}{W}^T\underset{\sim}{W}}{n}$.

Many MLR methods fit

$$\underset{\sim}{Z} = \underset{\sim}{W}\underset{\sim}{\eta} + \underset{\sim}{e}$$ and then find $\hat{Z}$ from data

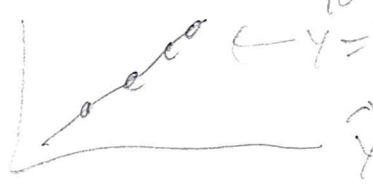
use $\hat{\underset{\sim}{Y}} = \hat{\underset{\sim}{Z}} + \bar{\underset{\sim}{Y}}$.

7) Problems i) when $n > p$ X is not invertible

but if $n \geq p$ $\hat{\underset{\sim}{Y}} = H\underset{\sim}{Y} = X(X^TX)^{-1}X^TY = I_n\underset{\sim}{Y} = \underset{\sim}{Y}$.

regardless of how bad the predictors are.

"Overfitting"
"fitting noise"



PI for Y_h is $[\hat{Y}_{hL}, \hat{Y}_{hU}]$ of 0% coverage.

ii) If $n < p$ $\hat{\underset{\sim}{Y}} = \underset{\sim}{Y}$ or program fails,

iii) Need $n \geq 5p$ with $J \geq 5$ and preferably $J \geq 10$. If $n < 5p$ the model is overfitting (not enough data to estimate $\hat{\beta}_{p\times 1}$ well).

8)

x-variable	coef	SE	shortest 95% CI for β_i
x_1 = intercept = constant	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$[\hat{L}_1, \hat{U}_1]$
x_2	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$[\hat{L}_2, \hat{U}_2]$
⋮			
x_p	$\hat{\beta}_p$	$SE(\hat{\beta}_3)$	$[\hat{L}_p, \hat{U}_p]$

The large sample 95% CI

for β_i is $\hat{\beta}_i \pm 1.96\, SE(\hat{\beta}_i)$.

(consider $H_0: \beta_i = 0$ $H_A: \beta_i \neq 0$

If $0 \in$ CI for β_i, then fail to reject H_0 and conclude X_i is not needed in the MLR model given the other predictors are in the model.

If $0 \notin$ CI for β_i, reject H_0 and conclude X_i is needed in the MLR model. see HW4 B)

9.1, 9.3, 9.5, 9.32 J 9.1 Forward Selection

9] Variable selection is a search for a subset of predictors that can be deleted

{ without important loss of information if $n \geq 10p$
{ so that the model I is good for prediction if $n < 5p$.

I = remaining predictors

10] I_{min} is the model, among the p candidates, that minimizes C_p if $n \geq 10p$ EBIC if $n < 10p$.

label could be X_1

I_j	model	X_2	X_3	X_4	X_5
I_1			*		
I_2	1		*	*	
I_3	2		*	*	
I_4	3	*	*	*	
I_5	4	*	*	*	*

$\hat{\beta}_{I_j,0}$ if $\hat{\beta}_i = \hat{\beta}_{j \times 1}$

$(\hat{\beta}_1, 0, \hat{\beta}_3, 0, 0)^T$

$(\hat{\beta}_1, 0, \hat{\beta}_3, \hat{\beta}_4, 0)^T$

$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, 0)^T$

$(\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4, \hat{\beta}_5)^T = \hat{\beta}_{OLS}$ of full model

out C_p $C_p(I_2)$ $C_p(I_3)$ $C_p(I_4)$ $C_p(I_5)$

The models also contain X_1 = constant = 1

11) For $n \geq 10p$ find the OLS full model residuals for the residual bootstrap. Then bootstrap $\hat{\beta}_{I_{min},0}$. (see HW4A).

$\hat{\beta}_{I_{min},0}$ is $p \times 1$

	coef	SE	95% shorth CI for β_j
X1 = int = const	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$[L_1, U_1]$
X2	0		$[L_2, U_2]$
X3	0		$[L_3, U_3]$
X4	$\hat{\beta}_4$	$SE(\hat{\beta}_4)$	$[L_4, U_4]$
X5	$\hat{\beta}_5$	$SE(\hat{\beta}_5)$	$[L_5, U_5]$

missing if $\hat{\beta}_{I_{min},j} = 0$

number of coef

$T_n = \hat{\beta}_{I_{min},0}$ OLS SE's not accurate unless

$T_n = \hat{\beta}_{I,0}$ i) I_{min} is chosen before seeing the data for $\hat{\beta}_{I_{min},0}$

$T_n = \hat{\beta}_{EN,0}$ or ii) $W = \frac{X^T X}{n} \to V^{-1} = \text{diag}(d_1, \ldots, d_p)$, all $d_i > 0$

Suppose the nontrivial predictors are iid, have mean, finite variance, or the predictors are orthogonal. Then

$$W = (w_{ij}) \text{ with } w_{ij} = \frac{\underline{v}_i^T \underline{v}_j}{n} = \frac{\sum_{k=1}^{n} x_{ki} x_{kj}}{n} \xrightarrow{P} E(X_i X_j)$$

where $X = [\underline{v}_1 \cdots \underline{v}_p]$. If $\underline{v}_i \perp \underline{v}_j$ then $\underline{v}_i^T \underline{v}_j = 0$ for $i \neq j$

If $X_i \perp X_j$ for $i, j \geq 1$, $i \neq j$ then $w_{ij} \to EX_i EX_j = 0 \cdot 0$

If $X_i \equiv 1$ then $w_{ij} \to \bar{x}_j \xrightarrow{P} EX_j = 0$. So ii) holds.

Note that if $\underset{3 \times 1}{\hat{\beta}_{I_{min}}} = (\hat{\beta}_1, \hat{\beta}_4, \hat{\beta}_5)^T$, then

$\hat{\beta}_{I_{min},0} = (\hat{\beta}_1, 0, 0, \hat{\beta}_4, \hat{\beta}_5)^T$: pad $\hat{\beta}_{I_{min}}$ with 0s to

12} OTH A model for variable selection is

$$Y = \underline{x}^T\underline{\beta} + e = \underline{x}_S^T\underline{\beta}_S + \underline{x}_E^T\underline{\beta}_E + e = \underline{x}_S^T\underline{\beta}_S + e \quad (*)$$

where $\underline{x} = (\underline{x}_S^T, \underline{x}_E^T)^T$ is $p \times 1$, $\underline{x}_S$ is $q_S \times 1$ and $\underline{x}_E$ is $(p - q_S) \times 1$. Given $\underline{x}_S$ is in the model, $\underline{\beta}_E = \underline{0}$ and E denotes the subset of terms that can be eliminated given subset S is in the model.

13} S is unknown. Let $\underline{x}_I$ be the vector of terms from a candidate subset I and let $\underline{x}_O$ be the vector of remaining predictors (out of the candidate subset.

Then $Y = \underline{x}_I^T\underline{\beta}_I + \underline{x}_O^T\underline{\beta}_O + e$. If $S \subseteq I$ and (*) holds, then

$$\underline{x}^T\underline{\beta} = \underline{x}_S^T\underline{\beta}_S = \underline{x}_S^T\underline{\beta}_S + \underline{x}_{I|S}^T\underline{\beta}_{I|S} + \underline{x}_O^T\underline{\beta}_O = \underline{x}_I^T\underline{\beta}_I$$

where $\underline{x}_{I|S}$ denotes predictors in I that are not in S.

14} $Y = \underline{x}^T\underline{\beta} + e$ = full model

$Y = \underline{x}_I^T\underline{\beta}_I + e$ = submodel

[illegible] is a submodel.

15) Underfitting occurs if $S \not\subseteq I$ so $\underset{p\times 1}{x_I}$ is missing important predictors. Underfitting occurs if $\underset{p\times 1}{x_I}$ with $d = k < a_S$. Overfitting occurs

if $n < 5k$ or if $S \subseteq I$ but $S \neq I$.

$n < 5k$: not enough data to estimate the k parameters well.

Overfitting is serious if $n < 5k$ but not "much of a problem" if $n > 10p$ or $n > 20p$.

Underfitting is a serious problem.

$$Y = x_I^T \beta_I + e_I$$

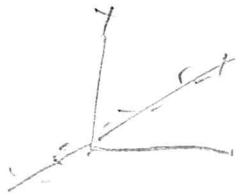
$\text{var}(e_I) > \text{var}(e) = \sigma^2$

may not be constant could depend on case i.

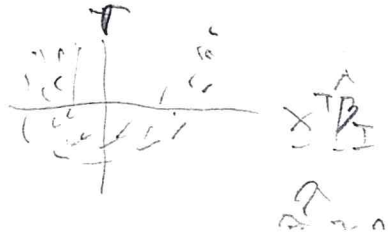
and the model may no longer be linear. (check with response and residual plots.

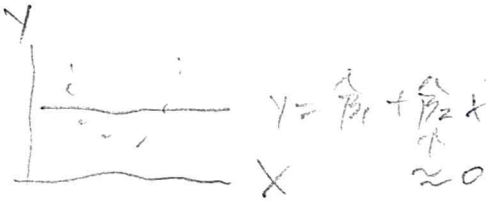
ex] $Y = \beta_1 + \beta_2 x + \beta_3 x^2 + e$

$SP = x^T\beta$ is a hyperplane that is a parabolic in x



If $SP(I) = \beta_1 + \beta_2 x$ = linear x





16) Forward Selection is a shrinkage method:

2^p models are produced and except for the full model, some $|\hat{\beta}_i|$ are shrunk to 0. Lasso and elastic net are also shrinkage methods. Ridge regression is a shrinkage method but $|\hat{\beta}_i|$ is not shrunk to 0. Shrinkage methods that shrink $\hat{\beta}_i$ to are also variable selection methods.

17) The pop MLR model is sparse if a_S is small. The pop model is dense if $\frac{n}{a_S} < J$ where $J=5$ or 10, say.

β_S is $a_S \times 1$

The fitted model $\hat{\beta} = \hat{\beta}_{I_{min,0}}$ is sparse if d = # nonzero coefficients is small.

The fitted model is dense if $\frac{n}{d} < J$ where $J=5$ or 10.

18) 097

I_j		# models fitted
I_1	$x_1^* = x_1$	0 (or 1: often do not fit this model)
I_2	x_1^*, x_2^*	$p-1$
I_3	x_1^*, x_2^*, x_3^*	$p-2$
⋮		
I_p	$x_1^*, \ldots, x_p^*$	1

$p(p-1)$

Forward selection forms a sequence of submodels

$I_1, \ldots, I_m$ eg $m = \min(p, \lceil \frac{n}{5} \rceil)$

I_1 uses $x_1^* = x_1$ a constant but no nontrivial predictors.

To form I_2 consider all models I with 2 predictors including x_1^*, Compute

$Q_2(I) = SSE(I) = RSS(I) = \underline{r}^T(I)\underline{r}(I) = \sum_{i=1}^{n} r_i^2(I)$

$= \sum_{i=1}^{n} (Y_i - \hat{Y}_i(I))^2$. (SSE stands for sum of squared errors, RSS is more accurate.)

Let I_2 minimize Q_2 and use x_1^*, x_2^*.

In general to form I_j, consider all models I with j predictors including $x_1^*, \ldots, x_{j-1}^*$.

Let I_j minimize $Q_j(I) = RSS(I)$ and use predictors $x_1^*, \ldots, x_j^*$.

[9] If $m = \lceil \frac{n}{5} \rceil \approx \frac{n}{5}\left(\frac{2p - \frac{n}{5}}{2}\right) \approx \frac{n(2p-n)}{2}$ models are fit. There are fast updating formulas for OLS for adding 1 predictor and branch and bound algorithms shorten time, but forward selection

20) 095 Need to choose the final model from the sequence of M models $I_1, \ldots, J_M$.

Let $\underline{x}_I$ and $\hat{\beta}_I$ be axl. For a given data set, p, n, and $\hat{\sigma}^2$ act as constants. A criterion below may add a constant or be divided by a positive constant without changing the subset I_{min} that minimizes the criterion.

Let Criterion $C_S(I)$ have the form

$$C_S(I) = SSE(I) + a k_n \hat{\sigma}^2.$$

$C_p(I) = AIC_S(I)$ uses $k_n = 2$ while $BIC_S(I)$ uses $k_n = \log(n)$. If $n \geq Jp$, $J \geq 5$ preferably $J \geq 10$, $\hat{\sigma}^2 = MSE = \sum_{i=1}^{n} \frac{r_{i,OLS}^2}{n-p}$.

The following criterion still need $\frac{n}{p}$ large

$$AIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + 2a$$

$$BIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + a \log(n),$$

The EBIC criterion may work when $\frac{p}{n}$ is not large.

$$EBIC(I) = BIC(I) + 2 \log\left[\binom{p}{a}\right]$$

$$\binom{p}{a} = \frac{p!}{(p-a)!\,a!} \quad \text{so} \quad \log\binom{p}{a} = \log(p!) - \log[(p-a)!] - \log[a!]$$

21) Fix p_1 and let I_{min} minimize C_p or AIC.

Let $m = p_1$. The probability that I_{min} underfits $\to 0$ as $n \to \infty$.

If $S \subseteq I$ then $\hat{\beta}_{I,0}$ is a $\sqrt{n}$ consistent estimator of β if $Y = x^T\beta + e = x_S^T\beta_S + e$.

Since there are at most 2^p regression models I that contain S and the prob that I_{min} picks one of these models goes to 1, $\hat{\beta}_{I_{min},0}$ is a $\sqrt{n}$ consistent estimator of β.

Hence the large sample $100(1-\delta)\%$ PI (2.7) works for OLS forward selection if p is fixed.

$\hat{\beta}_{I_{min}}$ is $d \times 1$ and $\hat{m}(x) = x^T\hat{\beta}_{I_{min},0} = x_{I_{min}}^T \hat{\beta}_{I_{min}} = \hat{m}(x_{I_{min}})$.

22) $\hat{\beta}_{I_{min},0}$ is not asymptotically normal, so residual bootstrap inference has not yet been proven to work, but in simulations, bootstrap inference was superior to that of the OLS full model if $a_S < p$ $\frac{n}{p}$ is large and $\beta \geq 5\sigma p$ is large.

Exception $\hat{\beta}_{I_{min},S}$ is asymptotically normal if $\frac{X^TX}{n} \to \text{diag}(d_1, \ldots, d_p)$ with all $d_i > 0$.

ex 8.1-8.3, 13, $n=100$, $p=4$, $\beta=(1,1,0,0)^T$

So $\underline{\beta}_2 = \begin{pmatrix}\beta_1\\ \beta_2\end{pmatrix}$. $\underline{x} = (1\ \underline{0}^T)^T$. Let $m=n-1$.

$\underline{w}_i \sim N_{n-1}(\underline{0}, I)$, $\underline{v}_i = A\underline{w}_i \sim N_{n-1}(\underline{0}, A^2)$

where $A = A^T = (a_{ij})$, $a_{ij} = \begin{cases} \psi, & 0 \le \psi < 1 \quad i \ne j \\ 1 & i = j \end{cases}$

So $\Sigma_0 = (\sigma_{ij})$ with $\sigma_{ij} = \begin{cases} 1+(n-1)\psi^2 & i=j \\ 2\psi+(n-2)\psi^2 & i \ne j \end{cases}$

So $\rho = \text{cor}(x_i, x_j) = \dfrac{2\psi+(n-2)\psi^2}{1+(n-1)\psi^2}$ for $i \ne j$, $i,j > 1$.

If $\psi = 0$, $\rho = 0$, if $\psi = \frac{1}{\sqrt{cp}}$, $\rho \to \frac{1}{c+1}$ as $p \to \infty$.

If ψ is close to 1 or .420 (fixed) and p large, then ρ gets close to 1 and the nontrivial predictor vectors cluster about a line in the direction of $\underline{1}$.

The prediction region method was used to test $H_0: \begin{pmatrix}\beta_3\\ \beta_4\end{pmatrix} = \begin{pmatrix}0\\ 0\end{pmatrix}$,

$e_i \sim N(0,1) \to$ pop CI len $= 0.392$ if $\psi = 0$

len of $\{\underline{0}, D_{(4)}\}$ / $\sqrt{\chi^2_{2,.95}} \approx 2.45$

model	ψ	cov/len	β_1	β_2	β_3	β_4	test	
reg	0	cov	.9456	.9474	.9496	.9474	.9442	
		len	.3961	.3997	.3988	.3992	2.4503	different digits so reg's are not comparable not χ^2
vs	0	cov	.9472	.9470	.9980	.9980	.9936	
		len	.3964	.3991	.3246	.3233	2.6936	
reg	0.9	cov	.9432	.9512	.9500	.9498	.9442	
		len	.3963	3.2621	3.2613	3.2611	2.4505	
		cov						

Testing $\underset{\sim}{\beta}_0 = \underset{\sim}{0}$ has smaller volume (length) for bootstrap FS because $\beta_i^* = 0$ often for β_i a component of $\underset{\sim}{\beta}_0$.

23) 083 To my knowledge, this is the 1st simulation where bootstrap FS $\hat{\ }$ inference is as good or better than OLS full model inference if $\frac{n}{p}$ is large.

When $\frac{n}{p}$ is small OLS full model inference is bad and should not be used (eg interpolates the data if $n = p$).

24) 0132 Suppose $Y = \underset{\sim}{x}^T \underset{\sim}{\beta} + e$, $n \geq 10p$ $B \geq 50p$

and $\beta_{2i}^* = \beta_{3i}^* = \beta_{6i}^* = 0$ for $i = 1, \ldots, B$

β_1^*	β_2^*	β_3^*	β_4^*	β_5^*	β_6^*
	0	0			0
	⋮	⋮			⋮
	0	0			0

Consider testing $H_0 \underset{\sim}{\beta}_0 = \begin{pmatrix} \beta_2 \\ \beta_3 \\ \beta_6 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ vs $H_A \underset{\sim}{\beta}_0 \neq \underset{\sim}{0}$

Conjecture: fail to reject H_0 and method can be used after looking at the data. (after the bootstrap).

Idea prediction region method covers $\geq 100(1-\delta)\%$ of $\underset{\sim}{\beta}_{0i}^*$ so prediction region confidence region is $\{\underset{\sim}{0}\}$ for

25) 0133 Another variable selection model

is $\underset{\sim}{x}^T\underset{\sim}{\beta} = \underset{\sim}{x}_{S_i}^T \underset{\sim}{\beta}_{S_i}$ for $i = 1, \ldots, J$

so that there are $J \geq 2$ nonnested "true" sub models where $\underset{\sim}{\beta}_{S_i}$ is $a_{S_i} \times 1$.

If x_1, x_2, x_3 are in S_1 and x_1, x_2, x_4 in S_2 then x_3 and x_4 are likely both often selected and omitted by forward selection in the B bootstrap samples. So omitting all predictors with $\beta_{ij}^* = 0$ would result in underfitting (using model x_1, x_2).

26) Criteria like $AIC_{(S)}$, $BIC_{(S)}$, and EBIC attempt to minimize $\frac{1}{n}\sum_{i=1}^{n} E\left(Y_{i,new} - \hat{Y}_{i,new}(I)\right)^2$

$= \frac{1}{n} E\left[\| \underset{\sim}{Y}_{new} - \hat{\underset{\sim}{Y}}_{new}(I) \|_2^2\right]$ ← based on training data

where $Y_{i,new}$ uses $\underset{\sim}{x}_t = \underset{\sim}{x}_i$ for $i = 1, \ldots, n$.

Using the PI length and coverage with cross validation may be better than EBIC when $\frac{n}{p}$ is small. (covered later)

Ø §3.3 J §6.3.1, 6.7.1 PCR

27) 0102 Suppose A $p \times p$ is symmetric so $A = A^T$

Then A has eigenvalue λ with eigenvector $\underline{x} \neq \underline{0}$ if $A\underline{x} = \lambda \underline{x}$. Let $\underline{e}$ be an eigenvector of A with unit length $\|\underline{e}\|_2 = \sqrt{\underline{e}^T\underline{e}} = 1$. Then $\underline{e}$ and $-\underline{e}$ are eigenvectors with unit length, Then A has p eigenvalue eigenvector pairs $(\lambda_1, \underline{e}_1), \ldots, (\lambda_p, \underline{e}_p)$ such that $\underline{e}_i^T \underline{e}_j = 0$.

Symmetric $A > 0$ is positive definite if $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ and $A \geq 0$ is positive semidefinite if $\lambda_p \geq 0$.

28) PCR Principal components Regression

for $\underline{Y} = \underline{X}\underline{\beta} + \underline{\varepsilon}$ uses $\underline{Z} = \underset{(p-1)\times 1}{\underline{W}}\,\underline{\eta} + \underline{\varepsilon}$

and finds $\hat{\underline{\beta}}$ and $\hat{\underline{Y}}$ using $\underline{\eta}$ and $\underline{\hat{\eta}}$. $R_{\underline{U}} = \frac{\underline{W}^T\underline{W}}{n}$

has eigenvalue eigenvector pairs $(\hat{\lambda}_1, \hat{\underline{e}}_1), \ldots, (\hat{\lambda}_k, \hat{\underline{e}}_k)$

where $k = \min(n, p-1)$. Let $\underline{W}_1, \ldots, \underline{W}_n$ be the standardized vectors of the non trivial predictors with $\underline{W} = \begin{pmatrix} \underline{W}_1^T \\ \vdots \end{pmatrix}$ Then the k Principal

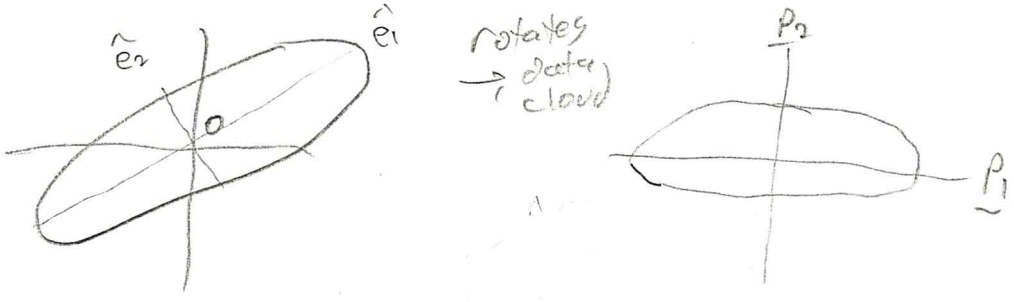
components corresponding to the jth cases $\underline{w}_j$ are $P_{j1} = \hat{e}_1^T \underline{w}_j, \ldots, P_{jK} = \hat{e}_K^T \underline{w}_j$.
(v_1, v_K)

The ith principal component consists of $(\hat{e}_i^T \underline{w}, \ldots, \hat{e}_i^T \underline{w}_n)^T = \underline{P}_i$

- PCR does OLS regression of $\underline{z}$ on $\underline{P}_1$, then $\underline{P}_1, \underline{P}_2, \ldots, \underline{P}_K$, resulting in K models.
($\underline{z}$, $\underline{P}_K$)

10 fold CV (0§3.11) is used to select the model.

29} Suppose $n \geq 10p$ so $K = p$. The OLS full model $= PCR_K$. Note that $\bar{\underline{w}} = 0$ and the hyperellipsoid $\{\underline{w} \mid D_{\underline{w}}^2(0, R_w) \leq h^2\} = \{\underline{w} \mid \underline{w}^T R_w^{-1} \underline{w} \leq h^2\}$ is centered at $\underline{0}$ with axes given by the eigenvectors $\hat{e}_i$ where the half length in the direction $\hat{e}_i$ is $h\sqrt{\lambda_i}$. $\underline{P}_1$ is obtained by projecting the $\underline{w}_i$ on the longest axis, $\underline{P}_2$ on the next longest axis, ..., P_K on the shortest axis.



30} Problems i) In general $\hat{\beta}_{OLS}$ is an inconsistent

estimator of $\underline{\beta}$ unless $P(i \to P-1) =$

$P\left(\underline{\hat{\beta}}_{PCR_{J_i}} = \underline{\hat{\beta}}_{OLS}\right) \to 1$.

ii) There is no reason why $\underline{P}_1, \underline{P}_2, \ldots, \underline{P}_K$ should decrease in importance for predicting y.

O §3.4 J §6.3.2 6.7.2 PLS

3) Partial least squares uses PLS components $\underline{P}_1, \ldots, \underline{P}_{P-1}$. Let $V_j = \underline{\gamma}_j^T \underline{X}$ and model

J_i contain $V_1, \ldots, V_i$. for $i = 1, \ldots, m$.

If $m = P-1$, PLS = OLS full model.

PLS uses $\underline{Z} = \underline{W}\underline{M} + \underline{E}$ then gets $\underline{\hat{\beta}}$ and $\underline{\hat{Y}}$ from $\underline{\hat{M}}, \underline{\hat{Z}}$ and $\underline{\bar{Y}}$. Unlike PCR, PLS uses Y in forming the PLS components: want components highly correlated with y.

3a) The 1st PLS component does OLS on Z and W_i without a constant to produce γ_{i1} for $i = 1, \ldots, P-1$. Then $\underline{\gamma}_1 = (\gamma_{11}, \ldots, \gamma_{P-1,1})^T$.

3b) $\underline{\hat{\beta}}_{PLS}$ is not a consistent estimator of $\underline{\beta}$ unless $\frac{P}{n} \to 0$.

O §3.5 J §6.2.1, 6.6.1 RR

343 $\underline{Y} = \underline{X}\underline{\beta} + \underline{e}$ and

$\underline{Z} = \underline{W}\underline{\eta} + \underline{e}$ is used to fit ridge regression.

Then $\hat{\underline{Z}}$, $\hat{\underline{\eta}}$ and $\bar{Y}$ are used to find $\hat{\underline{\beta}}$ and $\hat{Y}_i$.

Let $\lambda \geq 0$. Then the ridge regression estimator $\hat{\eta}_R$ minimizes the ridge regression criterion $Q_R(\underline{\eta}) = \frac{1}{a}(\underline{Z}-\underline{W}\underline{\eta})'(\underline{Z}-\underline{W}\underline{\eta}) + \frac{\lambda_n}{a}\sum_{i=1}^{p-1}\eta_i^2$ ($\underline{\eta}^T\underline{\eta}$)

$= \frac{1}{a} RSS_{\underline{W}}(\underline{\eta}) + \frac{\lambda_n}{a}\|\underline{\eta}\|_2^2$ with $a = 1, 2, n, 2n$ common. If $\lambda_n = 0$, $\hat{\underline{\eta}}_{RR} = \hat{\underline{\eta}}_{OLS}$. As $\lambda_n \to \infty$, $\hat{\underline{\eta}}_{RR} \to \underline{0}$ and $\hat{\underline{Y}} \to \bar{Y}$ (RR is a shrinkage method.)

$\hat{\underline{\eta}}_R = (\underline{W}^T\underline{W} + \lambda_n I_{p-1})^{-1}\underline{W}^T\underline{Z}$.

$\hat{\underline{\eta}}_{OLS} = (\underline{W}^T\underline{W})^{-1}\underline{W}^T\underline{Z}$ ← OLS full model if the inverse exists.

Warning: the literature typically uses $\lambda = \frac{\lambda_n}{a}$.

35} $\underline{W}^T\underline{W}$ is symmetric and square so $\underline{W}^T\underline{W} \geq 0$ with eigenvalues $\psi_1 \geq \cdots \geq \psi_p \geq 0$. If $\psi_p = 0$ then $(\underline{W}^T\underline{W})^{-1}$ does not exist.

Let $(\psi, \underset{\sim}{g})$ be an eigenvalue eigenvector pair of $\bar{W}^T\bar{W} = nR_{\underset{\sim}{v}}$. Then $\left(\bar{W}^T\bar{W} + \lambda_{in} I_{p-1}\right)\underset{\sim}{g}$

$= \bar{W}^T\bar{W}\underset{\sim}{g} + \lambda_{in}\underset{\sim}{g} = \psi\underset{\sim}{g} + \lambda_{in}\underset{\sim}{g} = (\psi + \lambda_{in})\underset{\sim}{g}$

So $(\underbrace{\psi + \lambda_{in}}_{>0 \text{ if } \lambda_{in}>0}, \underset{\sim}{g})$ is an eigenvalue eigenvector pair of $\bar{W}^T\bar{W} + \lambda_{in} I_{p-1}$ $\overbrace{>0 \text{ if } \lambda_{in} > 0.}^{\text{positive definite}}$

Hence $\left(\bar{W}^T\bar{W} + \lambda_{in} I_{p-1}\right)^{-1}$ exists $\forall\ \lambda_{in} > 0$, even if $\bar{W}^T\bar{W}$ is singular or ill conditioned.

36) Usually a grid of $M \approx 100$ λ_{in} values $0 \le \lambda_1 < \lambda_2 < \lambda_3 < \dots < \lambda_M$ is used where $\lambda_i = \lambda_{ini}$.

(5.3.11) 10 fold CV is used to select $\lambda_s = \hat{\lambda}_{in}$ for the final model. Model selection from M models, not variable selection.

37) 0108 Let augmented matrices

$$\bar{W}_A = \begin{pmatrix} \bar{W} \\ \sqrt{\lambda_{in}}\, I_{p-1} \end{pmatrix}, \quad \underset{\sim}{z}_A = \begin{pmatrix} \underset{\sim}{z} \\ \underset{\sim}{0} \end{pmatrix} \text{ where } \underset{\sim}{0} \text{ is } (p-1)\times 1,$$

For $\lambda_n > 0$, the OLS estimator from regressing $\underline{Z}_A$ on W_A is

$\hat{\underline{\gamma}}_A = (W_A^T W_A)^{-1} W_A^T \underline{Z}_A = \hat{\underline{\gamma}}_R$. Since

$W_A^T \underline{Z}_A = W^T \underline{Z}$ and $W_A^T W_A = W^T W + \lambda_n I_{p-1}$.

38] OIII) $\hat{\underline{\gamma}}_R = W^T \underbrace{(W W^T + \lambda_n I_n)^{-1}}_{n \times n} \underline{Z}$

$= \underbrace{(W^T W + \lambda_n I_{p-1})^{-1}}_{(p-1)\times(p-1)} W^T Z$

$A^{-1}_{n \times n}$ has $O(n^3)$ complexity so use 1st formula if $n < p-1$ and 2nd formula otherwise.

39] Suppose $n > p$ and $(W^T W)^{-1}$ exists.

Then $\hat{\underline{\gamma}}_R = \overbrace{(W^T W + \lambda_n I_{p-1})^{-1}}^{A_n} \underbrace{W^T W}_{I_{p-1}} \overbrace{(W^T W)^{-1} W^T Z}^{\hat{\underline{\gamma}}_{OLS}}$

So $\hat{\underline{\gamma}}_R = A_n \hat{\underline{\gamma}}_{OLS}$. HW 4 showed (problem 3.3)

$A_n = B_n = I_{p-1} - \lambda_n (W^T W + \lambda_n I_{p-1})^{-1}$!

40) Suppose p is fixed $\frac{W^TW}{n} \xrightarrow{P} V^{-1} \left(= S_0\right)$

and the OLS CLT holds so that

$$\sqrt{n}\left(\hat{\underline{\eta}}_{OLS} - \underline{\eta}\right) \xrightarrow{D} N_{p-1}\left(\underline{0}, \sigma^2 V\right).$$

If $\frac{\lambda_n}{n} \to 0$ then $\frac{W^TW + \lambda_n I_{p-1}}{n} \xrightarrow{P} V^{-1}$

and $n\left(W^TW + \lambda_n I_{p-1}\right)^{-1} \xrightarrow{P} V$. Also

$$A_n = \frac{n}{n} A_n = \left(\frac{W^TW + \lambda_n I_{p-1}}{n}\right)^{-1} \frac{W^TW}{n} \xrightarrow{P} V V^{-1} = I_{p-1}.$$

41) 0 1 2) RR CLT: Assume p is fixed

and LS CLT holds for $\underline{Z} = W\underline{\eta} + \underline{e}$ as in 39).

a) If $\frac{\lambda_n}{\sqrt{n}} \xrightarrow{P} 0$, then $\sqrt{n}\left(\hat{\underline{\eta}}_R - \underline{\eta}\right) \xrightarrow{D} N_{p-1}\left(\underline{0}, \sigma^2 V\right)$.

So OLS full model and RR are asymptotically equivalent

b) If $\frac{\lambda_n}{\sqrt{n}} \xrightarrow{P} \tau \geq 0$ then

$$\sqrt{n}\left(\hat{\underline{\eta}}_R - \underline{\eta}\right) \xrightarrow{D} N_{p-1}\left(\underbrace{-\tau V \underline{\eta}}_{\text{inferior to OLS unless this term} = \underline{0}}, \sigma^2 V\right)$$

Proof) $\hat{\underline{\eta}}_R = B_n \hat{\underline{\eta}}_{OLS} = \left(I_{p-1} - \hat{\lambda}_n \left(W^TW + \hat{\lambda}_n I_{p-1}\right)^{-1}\right) \hat{\underline{\eta}}_{OLS}$

Hence $\sqrt{n}\left(\hat{\underline{\eta}}_R - \underline{\eta}\right) = \sqrt{n}\left[\hat{\underline{\eta}}_R - \hat{\underline{\eta}}_{OLS} + \hat{\underline{\eta}}_{OLS} - \underline{\eta}\right]$

$= \sqrt{n}\left(\hat{\underline{\eta}}_{OLS} - \underline{\eta}\right) - \sqrt{n}\,\frac{\hat{\lambda}_{1n}}{n}\, n\left(W^TW + \hat{\lambda}_{1n} I_{p-1}\right)^{-1} \hat{\underline{\eta}}_{OLS}$

$\frac{n}{n} = 1$; $\frac{\hat{\lambda}_{1n}}{n} \xrightarrow{P} \tau$; $n\left(W^TW + \hat{\lambda}_{1n} I_{p-1}\right)^{-1} \xrightarrow{P} V$; $\hat{\underline{\eta}}_{OLS} \xrightarrow{P} \underline{\eta}$

$\xrightarrow{D} N_{p-1}(\underline{0}, \sigma^2 V) - \tau V \underline{\eta} \sim N_{p-1}(-\tau V \underline{\eta}, \sigma^2 V)$.

$-\tau V \underline{\eta}$ is $\underline{0}$ if $\tau = 0$.

□

Knight and Fu (2000) have a harder proof.

42) $\hat{\underline{\eta}}_R$ is $\sqrt{n}$ consistent if $\frac{\hat{\lambda}_{1n}}{\sqrt{n}} \xrightarrow{P} \tau$

and $\hat{\underline{\eta}}_R$ is a consistent estimator of $\underline{\eta}$

if $\frac{\hat{\lambda}_{1n}}{n} \xrightarrow{P} 0$ as $n \to \infty$.

43) Problems i) RR large sample theory is worse or as good as that of OLS full model if $\frac{n}{p}$ is large.

ii) 10 fold CV does not appear to guarantee that $\frac{\hat{\lambda}_{1n}}{\sqrt{n}} \xrightarrow{P} 0$ or $\frac{\hat{\lambda}_{1n}}{n} \xrightarrow{P} 0$.

iii) RR tends to undertit if $\frac{n}{p}$ is more than about 20 and the predictors are highly correlated

44) RR is a lot better than OLS full model if a) X^TX is singular or ill conditioned or b) $\frac{n}{p}$ is small.

O §3.6 J §6.2.2 §6.6.2 lasso

45) $\underline{y} = \underline{X}\underline{\beta} + \underline{e}$ for $\underline{z} = \underline{W}\underline{\eta} + \underline{e}$

The lasso estimator $\hat{\underline{\eta}}_L$ minimizes the criterion

$$Q_L(\underline{\eta}) = \frac{1}{n}(\underline{z} - \underline{W}\underline{\eta})^T(\underline{z} - \underline{W}\underline{\eta}) + \frac{\lambda_n}{n}\sum_{i=1}^{p-1}|\eta_i|$$

$\leftarrow L_1$ norm

$$= \frac{1}{n} RSS_{\underline{W}}(\underline{\eta}) + \frac{\lambda_n}{n}\|\underline{\eta}\|_1.$$ Lasso is a shrinkage method and a variable selection method some $\hat{\eta}_i = 0$. Like ridge regression, the lasso criterion is convex so fast algorithms to compute $\hat{\underline{\eta}}_L$ exist.

46) A grid of $m = 100$ λ_n values is used

$0 \le \lambda_1 < \lambda_2 < \cdots < \lambda_m$ and $\lambda_i = \lambda_{ni}$.

10 fold CV is used to select $\lambda_S = \hat{\lambda}_n$.

λ_m is the smallest value of λ such that $\hat{\underline{\eta}}_{\lambda_m} = \underline{0}$. Hence $\hat{\underline{\eta}}_{\lambda_i} \neq 0$ for $i < m$.

47) By Karush Kuhn Tucker (KKT) conditions for convex optimality (Math 471 see),

$$-\underline{W}^T(\underline{z} - \underline{W}\hat{\underline{\eta}}_L) + \frac{\lambda_n}{2}\underline{s}_n = \underline{0}$$ where $s_{in} \in \{-1, 1\}$.

Thus $\hat{\underline{\eta}}_1 = (\underline{W}^T\underline{W})^{-1}\underline{W}^T\underline{z} - n(\underline{W}^T\underline{W})^{-1}\frac{\lambda_n}{2}\underline{s}_n$

48) Lasso CLT: Assume p is fixed and the OLS CLT holds for $\underline{Z} = \underline{W}\underline{\eta} + \underline{e}$.

a) If $\hat{\lambda}_n/\sqrt{n} \xrightarrow{P} 0$, then $\sqrt{n}(\hat{\underline{\eta}}_L - \underline{\eta}) \xrightarrow{D} N_{p-1}(\underline{0}, \sigma^2 V)$.

b) If $\hat{\lambda}_n/\sqrt{n} \xrightarrow{P} \tau \geq 0$ and $\underline{s}_n \xrightarrow{P} \underline{s} = \underline{s}_\eta$, then

$$\sqrt{n}(\hat{\underline{\eta}}_L - \underline{\eta}) \xrightarrow{D} N_{p-1}(-\tfrac{\tau}{2} V\underline{s}, \sigma^2 V).$$

Proof $\sqrt{n}(\hat{\underline{\eta}}_L - \underline{\eta}) = \sqrt{n}(\hat{\underline{\eta}}_L - \hat{\underline{\eta}}_{OLS} + \hat{\underline{\eta}}_{OLS} - \underline{\eta})$

$$= \sqrt{n}(\hat{\underline{\eta}}_{OLS} - \underline{\eta}) - \sqrt{n}\,\frac{\hat{\lambda}_n}{2n} n(\underline{W}^T\underline{W})^{-1}\underline{s}_\eta \xrightarrow{D}$$

$$N_{p-1}(\underline{0}, \sigma^2 V) - \frac{\tau}{2} V\underline{s} \sim N_{p-1}(-\tfrac{\tau}{2} V\underline{s}, \sigma^2 V) \quad \square$$

49) $\hat{\underline{\eta}}_L$ is $\sqrt{n}$ consistent if $\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{P} \tau$ and $\hat{\underline{\eta}}_L$ is a consistent estimator of $\underline{\eta}$ if $\frac{\hat{\lambda}_n}{n} \xrightarrow{P} 0$ as $n \to \infty$,

50) At most n coefficients $\hat{\eta}_i \neq 0$ even if $p > n$.

51) Problems i) Lasso large sample theory is worse or as good as that of the OLS full model if $\frac{n}{p}$ is large. ii) 10 fold CV does not appear to guarantee that $\frac{\hat{\lambda}_n}{n} \xrightarrow{P} 0$ or $\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{P} 0$. iii) Lasso tends to underfit if $a_S \neq 0$ and the predictors are highly correlated.

iv) RR can be better than lasso if $a_s > n$.

52) $\frac{n}{p}$ large and iii) is bad. Program should include $\lambda_1 = 0$ if $n \geq 5p$.

Also include a value like $\lambda_2 \approx \frac{\sqrt{n}}{\log(n)}$.

53) Lasso can be a lot better than the OLS full model if

a) X^TX is singular or ill conditioned

b) $\frac{n}{p}$ is small,

with $M = 100$ lasso can be much faster than forward selection if n and p are large.

0 § 3.7

54) The relaxed lasso estimator

$\underset{j \times 1}{\hat{\beta}_{RL}}$ is OLS fit to the j variables, including a constant, that have $\hat{\beta}_{iL} \neq 0$.

So $\hat{\beta}_{RL}$ is an alternative to forward selection. Let X_A denote the matrix corresponding to the j "active" lasso variables.

Then $\hat{\beta}_{RL} = (X_A^T X_A)^{-1} X_A Y$

55) Relaxed lasso should be $\sqrt{n}$ consistent when lasso is $\sqrt{n}$ consistent.

56) $\underline{Y} = \underline{X}\underline{\beta} + \underline{e}$ $\underline{Z} = \underline{W}\underline{\eta} + \underline{e}$

The elastic net estimator $\underline{\hat{\eta}}_{EN}$ minimizes the criterion $RSS_W(\underline{\eta}) + \lambda_1 \|\underline{\eta}\|_2^2 + \lambda_2 \|\underline{\eta}\|_1$ (*)

where $\lambda_1 = (1-\alpha)\lambda_n$, $\lambda_2 = 2\alpha\lambda_n$ and $0 \le \alpha \le 1$.

Note that $\alpha = 1$ corresponds to lasso using $2\lambda_n$ and $\alpha = 0$ corresponds to RR.

57) Let $\underline{W}_A = \begin{pmatrix} \underline{W} \\ \sqrt{\lambda_1}\, I_{p-1} \end{pmatrix}$, $\underline{Z}_A = \begin{pmatrix} \underline{Z} \\ \underline{0} \end{pmatrix}$.

Then $\underline{\hat{\eta}}_{EN}$ can be obtained from the lasso of $\underline{Z}_A$ on $\underline{W}_A$. See proof on 0121.

58) By KKT conditions for convex optimality of (*), $2\underline{W}^T\underline{W}\,\underline{\hat{\eta}}_{EN} - 2\underline{W}^T\underline{Z} + 2\lambda_1 \underline{\hat{\eta}}_{EN} + \lambda_2 \underline{s}_n = \underline{0}$

or $(\underline{W}^T\underline{W} + \lambda_1 I_{p-1})\,\underline{\hat{\eta}}_{EN} = \underline{W}^T\underline{Z} - \frac{\lambda_2}{2}\underline{s}_n$ or

$$\underline{\hat{\eta}}_{EN} = \underline{\hat{\eta}}_{RR} - n\left(\underline{W}^T\underline{W} + \lambda_1 I_{p-1}\right)^{-1} \frac{\lambda_2}{2n}\underline{s}_n. \quad s_i \in \{1, 3\}.$$

If $\hat{\lambda}_{1n}/\sqrt{n} \xrightarrow{P} \tau$ and $\hat{\alpha} \xrightarrow{P} \psi$, then $\frac{\hat{\lambda}_1}{\sqrt{n}} \xrightarrow{P} (1-\psi)\tau$ and $\frac{\hat{\lambda}_2}{\sqrt{n}} \xrightarrow{P} 2\psi\tau$.

59} EN CLT Assume p is fixed and the LSCLT holds. a) If $\hat{\lambda}_{in}/\sigma_n \xrightarrow{P} 0$, then

$\sqrt{n}\,(\hat{\underline{\beta}}_{EN} - \underline{\beta}) \xrightarrow{D} N_{p+1}(\underline{0}, \sigma^2 V)$ (asy equiv to OLS).

b) If $\frac{\hat{\lambda}_n}{\sqrt{n}} \xrightarrow{P} \tau$, $\hat{\alpha} \xrightarrow{P} \psi \in [0,1]$ and $\underline{\hat{S}} \xrightarrow{P} \underline{S} = \underline{S}_m$, then

$$\sqrt{n}\,(\hat{\underline{\beta}}_{EN} - \underline{\beta}) \xrightarrow{D} N_{p+1}\left(-V[(1-\psi)\tau\underline{\beta} + \psi\tau\underline{S}],\ \sigma^2 V\right).$$

See proof on 0123 use $\sqrt{n}(\hat{\underline{\beta}}_{EN} - \underline{\beta}) =$

$\sqrt{n}\,(\hat{\underline{\beta}}_{EN} - \hat{\underline{\beta}}_{RR} + \hat{\underline{\beta}}_{RR} - \underline{\beta})$ and RR CLT.

60} The enet function uses 10 fold CV and a grid of α values $\{\underbrace{0}_{\text{RR}}, \frac{1}{am}, \frac{2}{am}, \ldots, \frac{am}{am} = 1\}$ (lasso)

$am \geq 1$, $am+1$ values

so enet takes about $(am+1)$ times as long as lasso as lasso or RR. The default is $am = 10$.

Both RR and lasso have problems with highly correlated predictors, so EN likely does, too.

0§3.9 61} PI for Y_f is $\left[\hat{Y}_f + b_n \hat{\xi}_{\delta_1},\ \hat{Y}_f + b_n \hat{\xi}_{1+\delta_2}\right]$

asymptotically adds Y_f to the shorth applied to the residuals.

$b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+2d}{n-d}}$ if $d \leq \frac{8n}{9}$, $b_n = 5\left(1 + \frac{15}{n}\right)$ otherwise.

where d = crude estimate of df.

If $n \geq 10p$ $d = p$ works ok for OLS, FS, PCR, PLS, RR L, RL, EN. FS, PCR, PLS, L, RL use variables $x_1^*, \ldots, x_d^*$ and a better value for d is d = # variables used, including a constant = $\#\hat{\beta}_i \neq 0$. EN and RR have a better $d = tr(H)$, but it is hard to get d from the software.

FS, L, RL have $x_i^* = x_j = (0, \ldots, 0, 1, 0, \ldots, 0)\, x$ (1 in j^{th} position)

PCR and PLS have $x_i^* = \gamma_i^T x$, some linear combination of the predictors x.

62] If $\frac{n}{p}$ is small, FS with EBIC, lasso and RL can work well under strong regularity conditions for sparse POP models.

63] PIs work badly if x_f is not like the training data $x_1, \ldots, x_n$, eg if x_f is an outlier or not in a covering hyperellipsoid (or hypersphere) for $x_1, \ldots, x_n$. Extrapolation

D §3.10 J §5.1 Choosing from M models

64] If $\frac{n}{p} \geq 20$ could always pick OLS full model. Then inference is easy.

65] Program uses C_p if $n \geq 10p$ and EBIC if $n < 10p$ for FS.

66] For k-fold cross validation (k-fold CV), randomly divide the training data into k groups or folds of approx equal size $n_j \approx \frac{n}{k}$ for $j = 1, \ldots, k$. Leave out the 1st fold, fit the method to the $k-1$ remaining folds, then compute some criterion for the 1st fold. Repeat for folds $2, \ldots, k$.

67] For $\underline{Y} = X\beta + e$ MLR, compute $\hat{Y}_i(j)$ for each Y_i in the fold j left out.

Then $MSE_j = \frac{1}{n_j} \sum_{i=1}^{n_j} (Y_i - \hat{Y}_i(j))^2$ and

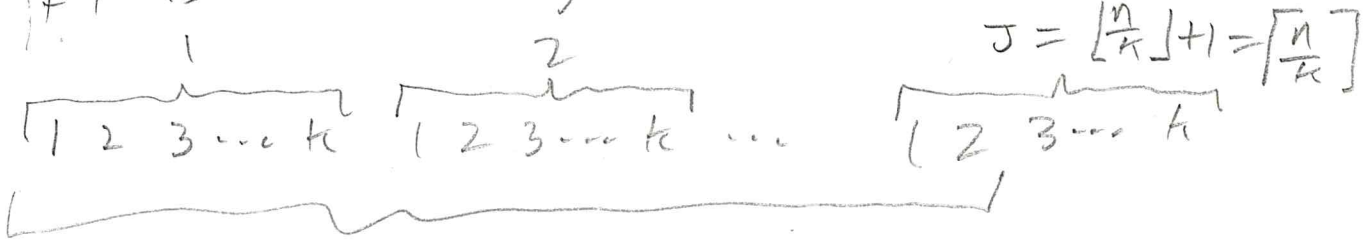
the overall criterion is $CV_{(k)} = \frac{1}{k} \sum_{j=1}^{k} MSE_j$.

If each $n_j = \frac{n}{k}$, then $CV_{(k)} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_{i(j)})^2$.

Pick model I that minimizes $CV_{(k)}(I)$. Usually $k = 10$ or

68] Jp250 code is not very good.

It is better to generate



get vector +folds
n×1
...

Then select a random permutation of +folds.

ex] n=26 k=5 (cases)

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
4	2	3	5	3	3	1	5	2	2	5	1	2	1	3	4	2	1	5	5	1	4	1	4	4

(folds)

Usually the cases are not given, see E2rev q1).

cases 7, 12, 14, 18, 21 and 23 are in fold 1

the other folds have 5 cases.

69] If $n \geq 5p$, lasso, RR, EN should use $\lambda_1 = 0$ and $\lambda_2 \simeq \frac{\sqrt{n}}{\log(n)}$. They may use $\lambda_i = \frac{\lambda_M i}{100}$, $i = 1, \ldots, 100$ where $\lambda_M \overset{\text{lasso}}{=} 2' \max\left(\frac{S_j^T z}{n}\right)$, j^{th} column of W

So λ_i may not be small enough for RR, L or EN CLT's to hold.

λ_M of EN is possible if S_j is highly correlated with z.

7.0} For FS, PCR, and PLS,

I_1 , I_2 ... I_p are well defined
x_1^* $x_1^* x_2^*$ $x_1^* \cdots x_p^*$

For RR, lasso and EN, I_i uses $\frac{\lambda_i}{2n} = \frac{\lambda_{n,i}}{2n}$, but 10 fold CV uses $K = n - n_j \approx .9n$ cases when fold j is left out. Maybe use $\frac{\lambda_n}{2n} \approx \frac{\lambda_K}{2K} \approx C$ are comparable, so $\lambda_K \approx \frac{K}{n} \lambda_n$. It is not clear what the code does.

7.0} Modify 10 fold CV to compute PI coverage and ave PI length. So n PIs are made using $\lambda_i = \lambda_i$ for $i = 1, \ldots, n$. The coverage is the prop of times the n PIs contained Y_i. Choose model I_d with the shortest ave PI length such that the nominal $100(1-\delta)\%$ PI had coverage $\geq c_n = \max(1-\delta - \frac{1}{3\sqrt{n}}, 1-\delta-0.02)$.

If no model I_i^0 has coverage $\geq c_n$, choose the model with the largest coverage. See 0119, Fig 3.1.

Read 0§3.11: all $\hat{\beta}_{0i}^* = 0 \Rightarrow \{0\}$ is the cont region

§3.12 Simulations 7.0} $\hat{\beta}_1, \ldots, \hat{\beta}_p$ test for FS, lasso, RR

73) Compare methods with PIs:

Want: PI len near pop length

e_i	$N(0,1)$	t_3	$Exp(1)-1$	$U(-1,1)$	$.9N(0,1)+.1N(0,100)$
pop len	3.92	6.365	2.996	1.90	13.490

(1) For RR used same d as lasso.

74) For $\frac{n}{p}$ large OLS forward sel PIs work,

PLS and PCR simulated well. Need $\frac{n}{p} \approx 100$ to get close to pop len. Lasso, RR, (RL) had trouble if $k \geq 19$ $\psi \geq 0.5$. k = # nontrivial active pop predictors = $q_s - 1$.

$\beta = (1, \ldots, 1, 0, \ldots, 0)^T$ ($k+1$ ones, $p-k-1$ zeros)

nominal 95% 5000 runs

n	p	ψ	k	cov/len	FS	L	RL	RR	PLS	PCR	
200	20	0	19	cov	.979	.977	.979	.979	.955	.979	PLS best given cov .94
				len	4.961	4.964	4.961	5.046	4.321	4.961	
400	20	0.9	19	cov	.967	.975	.960	.973	.955	.954	PCR best
				len	4.512	10.609	4.562	10.663	4.002	3.977	

75) For $\frac{n}{p} < 5$ FS with EBIC and lasso worked ok if k is small (sparse model)

100	200	0	1	cov	.965	.976	.927	.958	.662 (severe undercoverage)	.992	FS best given cov .90
				len	4.427	4.976	4.225	6.161 (too low)	2.770	12.412	
400	400	.9	19	cov	.935	.964	.956	.963	.946	.948	PLS best
				len	4.369	47.361	4.853	48.021	4.291	4.476	

$\psi = .9$ is favorable for PCR, PLS: 1 dominant principal component

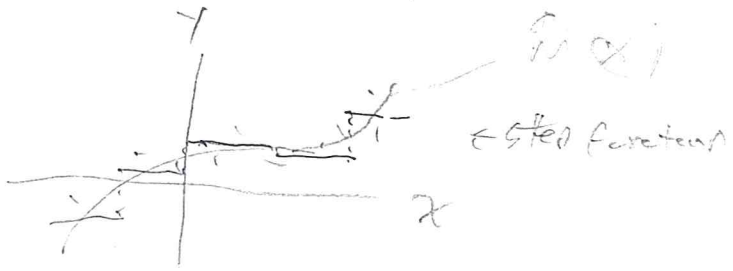
76) For a dense pop model $\frac{n}{ag}$ or $\frac{n}{M}$ is not small for

$\begin{pmatrix} Y_i \\ U_i \end{pmatrix} \sim N_p(\mu, \Sigma)$ every subset I_j follows MLR model and exam 2

O ch4 J ch7

J §7.1-7.3

1) Suppose $Y = m(x) + e$.



There are several flexible scatterplot smoothers $\hat{m}(x)$ with $df > 2$ including loess, lowess, smoothing splines and the additive error GAM $Y = \text{sat} + \text{sat}$, polynomial regression uses $Y = \sum_{i=1}^{J} B_i x^{i-1} + e$

$= B_1 + B_2 x + B_3 x^2 + \cdots + B_J x^{J-1} + e.$

we could use step functions: divide x into J intervals with $\approx \frac{n}{J}$ cases compute $\bar{Y}$ in each interval

Basis functions use $Y = \sum_{i=1}^{J} B_i b_i(x) + e$ where $b_1(x) \equiv 1$.

PI 0(2.7) often works well.

2) §7.4 we could divide X into $k+1$ intervals

$[a_0, a_1] (a_1, a_2], \ldots, (a_{k-1}, a_k] (a_k, a_{k+1}]$

where $a_1, a_2, \ldots, a_k$ are called knots.

Fit a cubic polynomial to each interval.

3) we get a cubic spline by having $\hat{m}(x)$ continuous with 2nd derivatives (so the curve $\hat{m}(x)$ is smooth). A cubic spline with k knots can be modeled as $Y = B_1 + B_2 b_2(x) + \cdots + B_{k+4} b_{k+4}(x) + e$

$1, x, x^2, x^3, h(x,\xi_1), \ldots, h(x,\xi_m)$ where

$\xi_1, \ldots, \xi_m$ are the knots and

$$h(x,\xi_j) = (x-\xi_j)_+^3 = \begin{cases} (x-\xi_j)^3, & x > \xi_j \\ 0, & \text{else.} \end{cases}$$

Programs use several sets of knots at percentiles and CV.

JS 7.5 4) A smoothing spline $\hat{g}$ minimizes

use ftn df(CV) or GCV and various values of λ

$$\underbrace{\sum_{i=1}^{n} (Y_i - g(x_i))^2}_{RSS(g)} + \lambda \int [g''(t)]^2 \, dt$$

$\underline{\hat{Y}} = H_\lambda \underline{Y}$ and $d = tr(H_\lambda)$. Then $\hat{g}(x) = \hat{m}(x)$.

5) JS 7.6 describes lowess and loess

JS 7.7 CH4

6) 4 important regression models are AER

1) $Y = SP + e$, $\begin{cases} SP = m(x) \\ \text{response plot and PIs like MLR} \end{cases}$

2) $Y|\underline{x} \sim \text{poisson}(e^{SP})$ poisson reg

3) $Y|\underline{x} \sim \text{bin}\left(1, \frac{e^{SP}}{1+e^{SP}}\right)$ binary reg

4) $Y_i | x_i \sim \text{bin}\left(m_i, \frac{e^{SP_i}}{1+e^{SP_i}}\right)$ binomial reg

3) is a special case of 4) with $m_i = 1$

If $SP = \underline{x}^T\underline{\beta}$ the model tends to be a generalized linear model GLM.

If $SP = \beta_1 + \sum_{j=2}^{p} S_j(x_j) = AP$ = additive predictor, the model tends to be a generalized additive model GAM. The GLM is a special case with $S_j(x_j) = \beta_j x_j$. A GAM is flexible while a GLM is inflexible.

7] For $n \geq 5p$, $J \geq 5$ and preferably $J \geq 10$ AIC is used for forward selection for a GLM. Also plot $\underline{x}_i^T \hat{\underline{\beta}}_{OLS}$ vs $\underline{x}_i^T \hat{\underline{\beta}}_{GLM}$. If the correlation is high do MLR forward selection and fit GLM to I_{min} variables using C_p if $n \geq 10p$.

8] Lasso-EN, 0.64.62 type criterion for binary, binomial?, and poisson regression (add family = binomial or family = poisson to the cv.glmnet arguments). A relaxed lasso GLM fits the GLM to the predictors with nonzero lasso coefficients. Or do MLR lasso and fit GLM to I_{min} variables. Using MLR forward selection with EBIC and applying [illegible] variables led to severe underfitting.

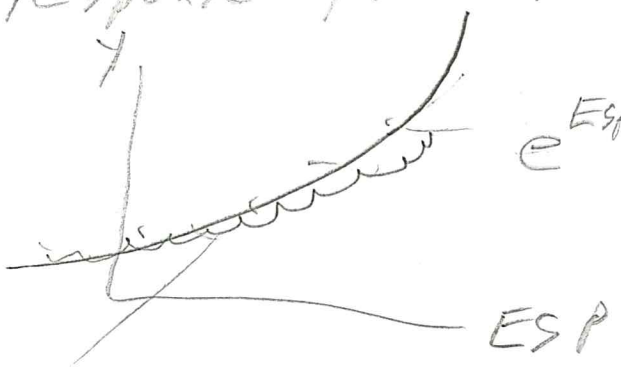
9) $EAP = ESP = \hat{\beta}_1 + \sum_{j=2}^{p} \hat{S}_j(x_j)$.

Often smoothing splines are used iteratively to find $\hat{S}_j$ for $j = 2, \ldots, p$. If a plot of x_j vs $\hat{S}_j$ is linear, $\hat{\beta}_j x_j$ can be used. If a plot of x_j vs $\hat{S}_j$ is a quadratic add x_j^2, cubic add x_j^3 and x_j

So the GAM can be useful for checking the GLM and for suggesting predictor transformations for the GLM.

10) Poisson regression response plot

$\widehat{E[Y|x]} = e^{ESP}$

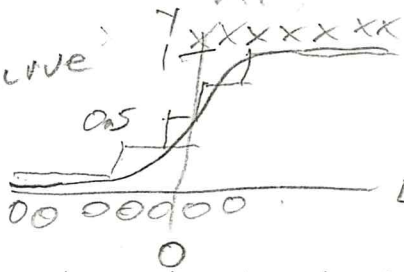


(lowess should track e^{ESP} except possibly for large ESP.

11) Binary regression response plot

(for binomial regression replace Y_i by $\frac{Y_i}{m_i}$ for a similar plot)

$\widehat{E(Y|x)} = \frac{e^{ESP}}{1+e^{ESP}}$ = logistic curve



0 if $Y = 0$
1 if $Y = 1$

Add a step function with step height = $\bar{Y}$ for cases in interval. Want step function to track logistic curve closely

Och 5 JCh 4 Classification Discriminant Analysis

1) In <u>supervised classification</u>, there are $G \geq 2$ known groups and m test cases to be classified. Each case is predicted to be in exactly one group based on its measurements $\underline{W}_i$.

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 i) had a heart attack or ii) did not have a heart attack.

ex) Person applies for credit card based on (salary, credit rating)^T and is classified as acceptable or not acceptable.

2) Suppose there are $G \geq 2$ groups or populations with pdf $f_j (\geq 1$ for $j = 1, \ldots, G$ and $\underline{x}$ is $p \times 1$, so if $\underset{p \times 1}{\underline{x}}$ comes from pop j, then $\underline{x}$ has pdf $f_j (\geq 1$,

Assume there is training data consisting of a random sample of n_j cases $\underline{x}_{1j}, \ldots, \underline{x}_{n_j j}$ for each group. Let $(\bar{\underline{x}}_j, S_j)$ be the sample

mean and covariance matrix for each group.
Let $\underline{W}_i$ be a new $p \times 1$ test random vector from 1 of the G groups, but the group is unknown. Usually there are $\underline{W}_1, \ldots, \underline{W}_m$ = test data, and discriminant analysis or classification attempts to allocate the $\underline{W}_i$ to the correct groups.

3) Let $\underline{\tilde{W}}_i$ be a random vector and $\underline{w}_i$ the observed random vector. Let $Y_i = j$ if w_i comes from the jth group for $j = 1, \ldots, G$.
Then $\pi_j = P(Y = j)$ and the posterior probability that $Y = k$ or $\underline{w}_i$ belongs to group k is

$$P_k(\underline{w}_i) = P\left(Y = k \mid \underline{\tilde{W}}_i = \underline{w}_i\right) = \frac{\pi_k f_k(\underline{w}_i)}{\sum_{j=1}^{G} \pi_j f_j(\underline{w}_i)}.$$

4) i) The maximum likelihood discriminant rule allocates $\underline{w}_i$ to group a (predicts $Y = a$) if $\hat{f}_a(\underline{w}_i)$ maximizes the $\hat{f}_j(\underline{w}_i)$ for $i = 1, \ldots, G$.

ii) The Bayesian discriminant rule allocates

$\underline{w}_i$ to group a if $\hat{p}_a(\underline{w}_i)$ maximizes

$$\hat{p}_k(\underline{w}_i) = \frac{\hat{\pi}_k \hat{f}_k(\underline{w}_i)}{\sum_{j=1}^{G} \hat{\pi}_j \hat{f}_j(\underline{w}_i)} \quad \text{for } k = 1, \ldots, G.$$

iii) The pop Bayes classifier allocates $\underline{w}_i$ to group a if $p_a(\underline{w}_i)$ maximizes

$p_k(\underline{w}_i)$ for $k = 1, \ldots, G$.

5) J p38-39, 139 The Bayes classifier has the lowest possible expected test error rate out of all classifiers using the same p predictors $\underline{w}$. Problem: usually the π_j and f_j are unknown.

6) The maximum likelihood rule and Bayesian discriminant rule are equivalent if $\hat{\pi}_j \equiv \frac{1}{G}$ for $j = 1, \ldots, G$.

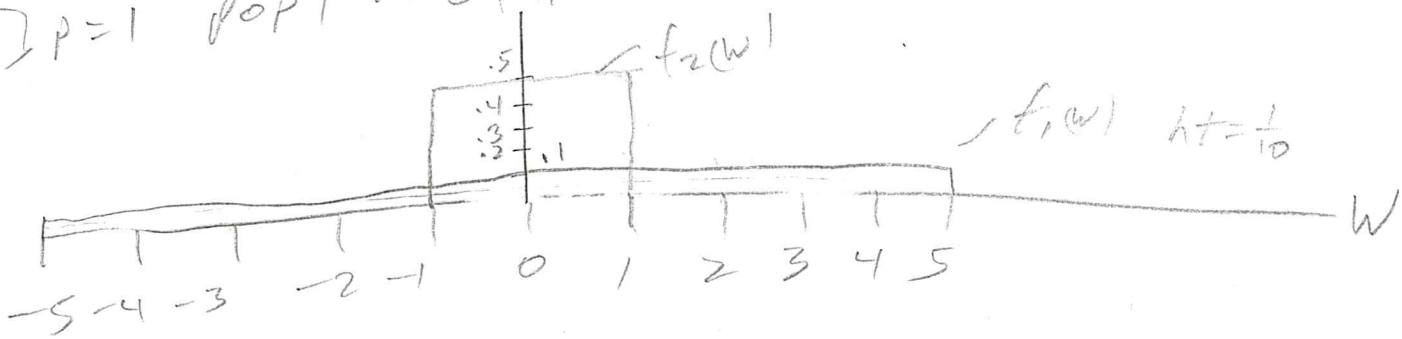
7) Several discriminant rules can be modified to incorporate π_j and costs of correct and incorrect classification. We will assume that the costs of correct allocation are unknown or 0 and that the costs of incorrect allocation are unknown

the probabilities π_j are unknown or $\pi_j \equiv \frac{1}{G}$ for $j = 1, \ldots, G$.

8) A regularized estimator attempts to use $\hat{d}$ such that $\frac{n}{\hat{d}}$ is large. (Lasso and RR with $\lambda > 0$ were regularized MLR estimators.)

9) Given $p = 1$ and a graph of $f_1, \ldots, f_G$, give the maximum likelihood rule.

ex) $p = 1$ pop 1 $\sim U(-5, 5)$ pop 2 $\sim U(-1, 1)$



ML rule: allocate w to pop 2 if $-1 < w < 1$
allocate w to pop 1 if $-5 < w < -1$ or $1 < w < 5$.

10) The pooled covariance matrix is

$$S_{pool} = \frac{1}{n-G} \sum_{j=1}^{G} (n_j - 1) S_j.$$ This estimator is good if the G groups have the same population covariance matrix Σ_x.

Σ and S_j $(p \times p)$ have $p + (p-1) + \cdots + 1 = \frac{p(p+1)}{2}$ unknown parameters

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

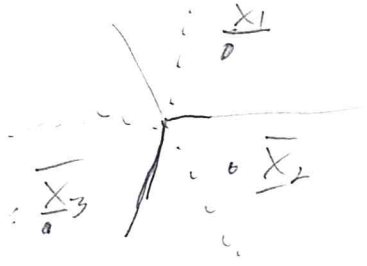
The df d is: often = # estimated parameters.

So S_{pool} is regularized compared to using $S_1, \ldots, S_G$ separately.

11) *The linear discriminant analysis (LDA) rule allocates $\underline{w}$ to the group with the largest value of

$$d_j(\underline{w}) = \bar{\underline{x}}_j^T S_{pool}^{-1} \underline{w} - \frac{1}{2} \bar{\underline{x}}_j^T S_{pool}^{-1} \bar{\underline{x}}_j = \hat{\alpha}_j + \hat{\underline{\beta}}_j^T \underline{w}.$$

LDA is the most used rule, and basically separates the G groups with G hyperplanes



12) *The quadratic discriminant analysis (QDA) rule allocates $\underline{w}$ to the group with the largest value of

$$Q_j(\underline{w}) = -\frac{1}{2} \log|S_j| - \frac{1}{2} (\underline{w} - \bar{\underline{x}}_j)^T S_j^{-1} (\underline{w} - \bar{\underline{x}}_j)$$

$$= -\frac{1}{2} \log|S_j| + D_{\underline{w}}^2(\bar{\underline{x}}_j, S_j).$$

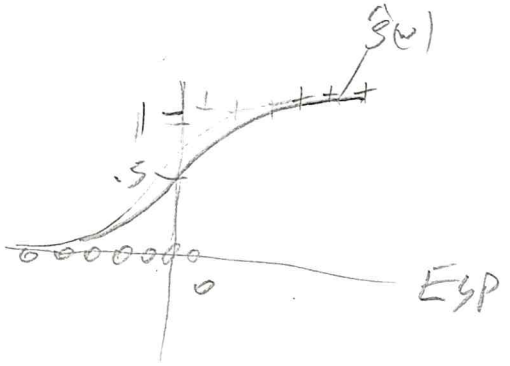
QDA df >> LDA df.

13) * Let $G=2$ with groups 0 and 1.

Let $\hat{p}(\underline{w}) = P(\underline{w} \in \text{group } 1) = P(Y=1 \mid \underline{w})$. Let $\hat{p}(\underline{w})$ be the binary logistic regression estimator of $p(\underline{w})$.

Then $ESP = \hat{\alpha} + \hat{\underline{\beta}}^T \underline{w}$ and $\hat{p}(\underline{w}) = \frac{e^{ESP}}{1+e^{ESP}} =$

$\frac{\exp(\hat{\alpha} + \hat{\underline{\beta}}^T \underline{w})}{1+\exp(\hat{\alpha} + \hat{\underline{\beta}}^T \underline{w})}$. The LR discriminant rule allocates $\underline{w}$ to group $\begin{cases} 1 & \text{if } \hat{p}(\underline{w}) \geq 0.5 \ (ESP \geq 0) \\ 0 & \text{if } \hat{p}(\underline{w}) < 0.5 \ (ESP < 0) \end{cases}$



The misclassification rate (training data) $\approx \min(\hat{p}, 1-\hat{p})$. $ESP > 2$ or $ESP < -2$ will have low error (misclassification) rate if the LR model is good.

14) * Training data: $\underline{x}_{1j}, \ldots, \underline{x}_{n_j j}$ group j.

Use a discriminant analysis method to classify the training data (optimistic compared to test data: estimated error rate is too low). If m_j of the n_j group j cases are correctly classified, the apparent error rate for group j is $1 - \frac{m_j}{n_j}$.

Let $m_A = \sum_{j=1}^{G} m_j$. Then the

apparent error rate AER $= 1 - \frac{m_A}{n}$.

15) leave one out cross validation CV: For $i = 1, \ldots, n$ leave case i out, 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 is $1 - \frac{m_C}{n}$.

(6) 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 from the cases not left out. Let m_L be the number of cases correctly classified. The "leave a subset out" error rate is $1 - \frac{m_L}{n_V}$.

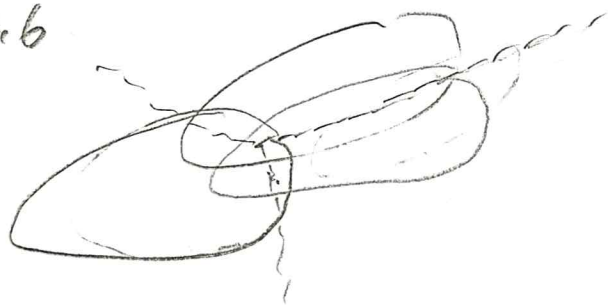
17) k-fold CV. Let m_K be the number of cases correctly classified. Then the k-fold CV error rate is $1 - \frac{m_K}{n}$.

18) * The K-nearest neighbors (KNN) method identifies the k cases in the training data closest to the test case $\underline{w}$. Suppose m_j of the k cases are from group j. The the k nearest neighbor (KNN) discriminant rule allocates $\underline{w}$ to group a if m_a maximizes m_j for $j=1,\ldots,G$.

(If there is a tie, we might use the group, among the tied groups, that has the smallest sum of distances.)

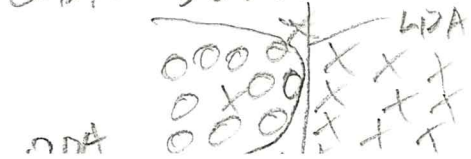
19) LDA roughly makes G hyperellipsoids $\{D_j^2(\bar{x}_i, S_p) \le k^2\}$ of the same volume and shape that cover most of the training data. The hyperplanes minimize overlap of the hyperellipsoids.

See J Fig 4.6



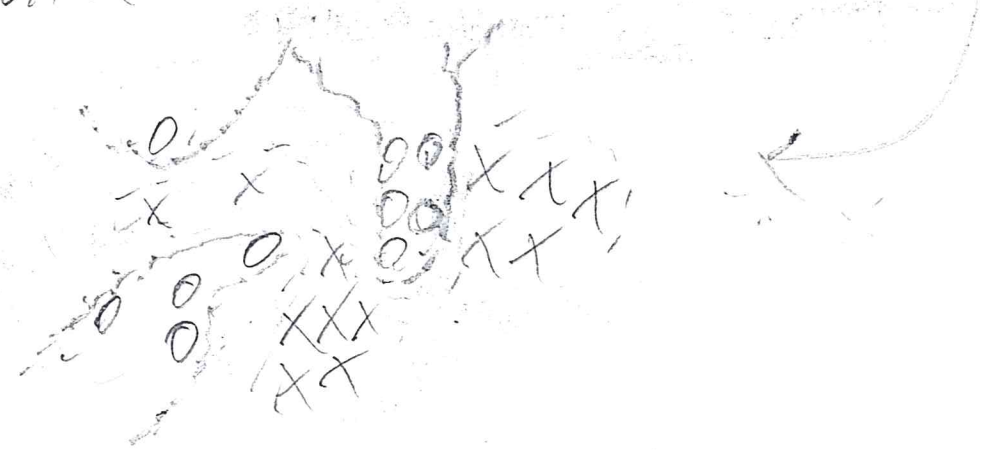
20) For $p=2$ LDA boundaries are lines
QDA boundaries are quadratics.

See J F 4.9



21] KNN is flexible see 5 Fig 2.15 2.16

22] Adding terms like x_i^2 & $x_i x_j$ to LDA increases the LDA df and makes LDA more like QDA.



23] For Fisher's discriminant analysis (FDA), let $\hat{W}$ be a $p \times p$ dispersion matrix and let $\hat{B}$ be a $p \times p$ symmetric matrix used to measure variability between classes. Let the eigenvalue eigenvector pairs of $\hat{W}^{-1}\hat{B}$ be $(\hat{\lambda}_1, \hat{e}_1), \ldots, (\hat{\lambda}_p, \hat{e}_p)$ with $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$. Then $\max_{a \neq 0} \frac{a^T \hat{B} a}{a^T \hat{W} a} = \hat{\lambda}_1$ with $a = \hat{e}_1$. Similarly $(\hat{\lambda}_i, \hat{e}_i)$ achieves the max among all unit vectors orthogonal to $\hat{e}_1, \ldots, \hat{e}_{i-1}$.

24) Let (T_i, C_i) be an estimator of multivariate location and dispersion for the ith group, eg $(T_i, C_i) = (\bar{x}_i, S_i)$. Let $\bar{T} = \frac{1}{G}\sum_{i=1}^{G} T_i$ and $\hat{B}_T = \sum_{i=1}^{G} (T_i - \bar{T})(T_i - \bar{T})^T$ so $\frac{\hat{B}_T}{G-1}$ is the sample covariance matrix of $T_1, \ldots, T_G$. Let $\hat{W}_T = \sum_{i=1}^{G} C_i$.

Let $\hat{W}_L = G\hat{\Sigma}_{pool}$. Other choices are possible.

25) The FDA discriminant rule allocates $\underline{w}$ to group a that minimizes $|\hat{e}_i^T \underline{w} - \hat{e}_i^T T_i|$ for $i = 1, \ldots, G$.

26) Let the ith group have pop mean and covariance matrix $(\underline{\mu}_i, \Sigma_i)$. Let $\underline{\bar{\mu}} = \frac{1}{G}\sum_{i=1}^{G} \underline{\mu}_i$, $B = \sum_{i=1}^{G} (\underline{\mu}_i - \underline{\bar{\mu}})(\underline{\mu}_i - \underline{\bar{\mu}})^T$ and $W = \sum_{i=1}^{G} \Sigma_i$. Then the between group variability $b(\underline{a}) = \underline{a}^T B \underline{a} = \sum_{i=1}^{G} |\underline{a}^T (\underline{\mu}_i - \underline{\bar{\mu}})|$, the within group variability $w(\underline{a}) = \underline{a}^T W \underline{a} = \sum_{i=1}^{G} \underline{a}^T \Sigma_i \underline{a} = \sum_{i=1}^{G} \mathrm{Var}(\underline{a}^T \underline{X}_i)$ where $\underline{X}_i$ is a RV from the pop corresponding to group i.

Then $\max_{\underline{a} \neq \underline{0}} \frac{b(\underline{a})}{w(\underline{a})} = \max_{\underline{a} \neq \underline{0}} \frac{\underline{a}^T B \underline{a}}{\underline{a}^T W \underline{a}} = \lambda_1$ with $\underline{a} = \underline{e}_1$

where λ_1 is the largest eigenvalue of $W^{-1}B$.

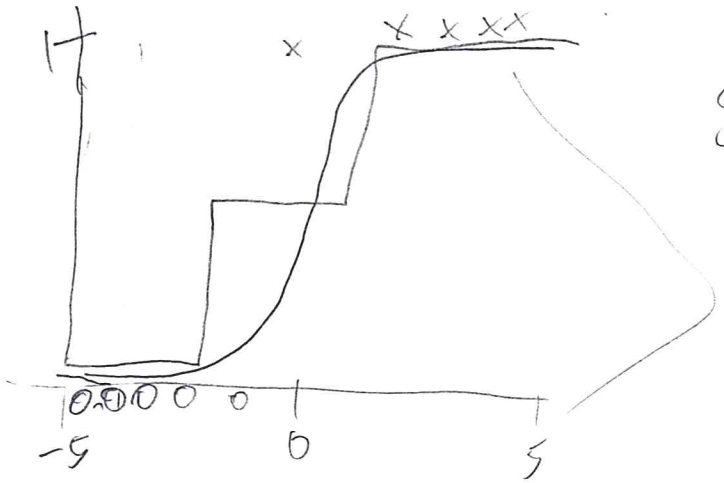
27) FDA attempts to approximate 26), and will likely work well with $\hat{W}_A^{-1}$ and $\hat{B}_T$ if $\hat{\underline{a}}^T \hat{W}_A \underline{a}$ can be made small.

28) If $G = 2$, $(T_i, C_i) = (\bar{x}_i, s_i)$, $\hat{B} = \hat{B}_T$ and $\hat{W} = 2\,S_{pool}$, then LDA and FDA are equivalent. The D_LDA program is actually an FDA program.

29) For kNN if $k=1$ AER=1 but if $\frac{n_j}{n} \to \pi_j$ then the test error rate L_n of $k=1$ kNN converges in probability to L where $L_B \leq L \leq 2L_B$ and L_B is the test error rate of the Bayes classifier. If $k = k_n$ satisfies $k_n \to \infty$ and $\frac{k_n}{n} \to 0$ as $n \to \infty$ the the kNN test error rate $L_n \xrightarrow{P} L_B$. The leave one out CV error rate $\hat{L}_n$ is a good estimator of L_n.

30) If $\underset{p\times 1}{W}$ is $p \times 1$ where p is large often W_{I_m} $k \times 1$ will do better at classification than W for some subset I_m of the p predictors.

31) LR response plots

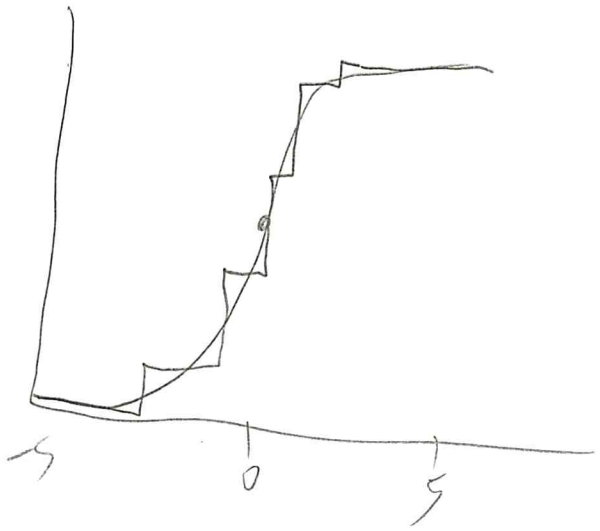


good repose plot

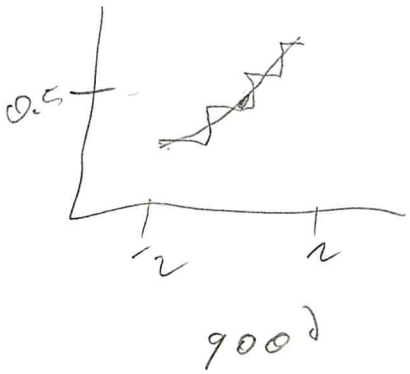
group 1 and 0 cases are well separated

good response plot

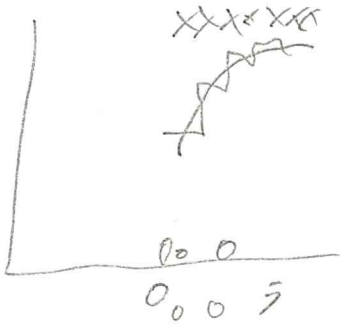
perfect classification of training data



good response plot

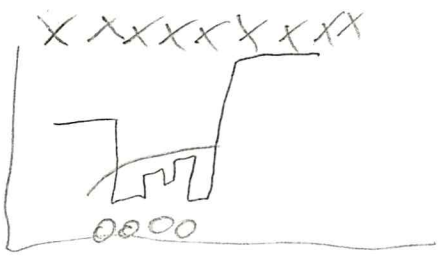


good

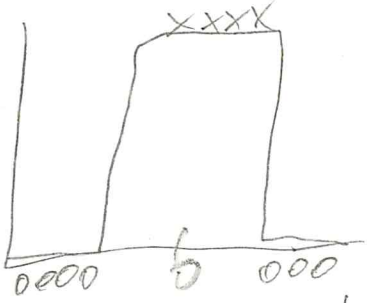


good

good



bad



bad plot although good classification is possible but

32) LR output

label	coef estimate
constant	$\hat{\alpha}$
x_1	$\hat{\beta}_1$
⋮	⋮
x_p	$\hat{\beta}_p$

std err Est/Err pval

0 = F 1 = M Gender

ex

label	Estimate
constant	−19.7762
head measurements: circum	0.0244688
head measurements: length	0.0371472

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

a) Find ESP for $\underline{x}$.

$$ESP = \hat{\alpha} + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = -19.7762 + 0.0244688(550) + 0.0371472(200) = \underline{1.1105}$$

b) Is $\underline{x}$ classified in group 0 or group 1?

group 1 since ESP > 0

c) Find $\hat{\pi}(\underline{x}) = \hat{\pi}_{ESP}$

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

33] variable selection: find a subset (SL 41
of variables that does a good job of
classification

crude forwards selection $x_1, \ldots, x_p$ are nontrivial predictors

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

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

⋮

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

⋮

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

Final model: might be the one with the smallest AER
if $n >> p$ so there is NO overfitting.

crude backward elimination

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

2) Delete $w_p = x_{j_1}$ such that the model with $p-1$ variables minimizes the AER. $w_1, \ldots, w_{p-1}$ are in the model.

3) Delete $w_{p-1} = x_{j_2}$ such that the model with $p-2$ variables minimizes the AER. $w_1, \ldots, w_{p-2}$ are in the model

!

k) $W_1, \ldots, W_{p-k+2}$ are in the model.

Delete $W_{p-k+2} = X_{j,k-1}$ such that the model with $p-k+1$ variables minimizes the AER.

$W_1, \ldots, \underbrace{W_{p-k+1}}_{p-(k-1)}$ are in the model.

⋮

p) W_1 and W_2 are in the model, Delete W_2 such that the model with W_1 minimizes the AER.

34) Other Criteria can be used.

proc stepdisc in SAS does variable selection, see OH 5.8 and HW 8 for crude variable selection with LDA and QDA.

Ch 6 Regularizing a correlation matrix R,

1) Classification is a multivariate analysis technique. Many multivariate analysis techniques are based on the covariance matrix $S = \hat{\Sigma}_x$ or correlation matrix $R = \hat{\rho}_x$, If $\underline{X}$ is $p \times 1$ Σ_x has $p + p + 1 + \cdots + 1 = \frac{p(p+1)}{2}$ parameters and ρ_x

has $(p-1)+(p-2)+\cdots+1 = p(p+1)/2$ parameters.

We want $n \geq 10p$ to use R or S. If $n \leq 5p$, these matrices are being overfit: the degrees of freedom is too large for the sample size n, and the matrices may be ill conditioned.

2) Also many multivariate techniques need $\hat{\Sigma}_x^{-1}$ or $\hat{\rho}_x^{-1}$. S and R need $n > p$ to be invertible.

3) Plug in estimators are often used.

ex "LDA" uses $\hat{\mu}_j^T \hat{\Sigma}_{pool}^{-1} \underline{w} - \frac{1}{2} \hat{\mu}_j^T \hat{\Sigma}_{pool} \hat{\mu}_j$

where the usual LDA has $(\hat{\mu}_j, \hat{\Sigma}_{pool}) = (\bar{x}_j, S_{pool})$.

4) A regularized estimator attempts to use well conditioned $\hat{\Sigma}_x$ or $\hat{\rho}_x$.

5) A common technique is to use $\hat{\Sigma}_x = S_d = \text{diag}(S)$ or $\hat{\rho}_x = R_d = \text{diag}(R) = I_p$.

(like using $\hat{\underline{\eta}}_M = \underline{0}$ for lasso, RR or EN).

6) Let $S = (s_{ij})$ and $D = \text{diag}(\sqrt{s_{11}}, \ldots, \sqrt{s_{pp}}) = \text{diag}(\sqrt{s_{ii}})$

Then $S = DRD$ and $R = D^{-1} S D^{-1}$

7) Let $S^{-1} = (s^{ij})$ and $R^{-1} = (r^{ij})$

Let $\underline{x}_{(ij)}$ be the vector $\underline{x}$ with x_i and x_j deleted $i \neq j$. Then partial correlation

$r_{ij\,\underline{x}_{(ij)}}$ between x_i and x_j is found by regressing x_i and x_j on $\underline{x}_{(ij)}$ get the two sets of residuals then find the correlation

$r_{ij\,\underline{x}_{(ij)}}$ of the two sets of residuals.

$$r_{ij\,\underline{x}_{(ij)}} = \frac{-s^{ij}}{\sqrt{s^{ii}\, s^{jj}}} = \frac{-r^{ij}}{\sqrt{r^{ii}\, r^{jj}}}$$

and $R^{-1} = D S^{-1} D$. Also $r^{ii} = VIF_i = \frac{1}{1 - R_i^2}$

where R_i^2 is the squared multiple correlation from regressing x_i on $\underline{x}_{(i)}$ that omits predictor i.

8) Graphical lasso (Glasso) takes $\underline{\underline{S}} \geq 0$ as an input and returns $\hat{\Sigma}_G \geq 0$. Many elements

($\underline{\underline{S}} \geq 0$: positive definite; positive semi definite)

9] Could use a robust estimator as an input.

10] There are a couple of books written on regularized versions of $\hat{\Sigma}$, $\hat{\Sigma}^{-1}$, R, R^{-1}.

11] RR regularizes $W^TW = nR$.

12] Let $\delta \geq 0$ and $\underset{p\times p}{R} = (r_{ij})$. A simple way to regularize R is to use

$$R_\delta = \frac{1}{1+\delta}(R + \delta I_p) = (t_{ij}) \quad \text{with } t_{ii} = 1$$

and $t_{ij} = \frac{r_{ij}}{1+\delta}$, for $i \neq j$, $R_\delta = kR + (1-k)I_p$

where $k = \frac{1}{1+\delta} \in (0,1]$. Note that each correlation r_{ij} is divided by the same factor $1+\delta$.

If λ_i is the ith eigenvalue of R, then $\frac{\lambda_i + \delta}{1+\delta}$ is the ith eigenvalue of R_δ. The eigenvectors of R and R_δ are the same since if $R\underline{x} = \lambda_i \underline{x}$ then

$$R_\delta \underline{x} = \frac{1}{1+\delta}(R + \delta I_p)\underline{x} = \frac{1}{1+\delta}(\lambda_i + \delta)\underline{x}.$$

13] R can be further regularized by setting $t_{ij} \doteq 0$ if $|t_{ij}| \leq \tau$ where $\tau \in [0,1]$ should have $\tau < 0.5$. Denote the resulting matrix by

$R(\delta,\tau)$. Then $R_\delta = R(\tau,0)$. The default is $\tau = 0.05$. Using τ is known as thresholding.

14) A regularized covariance matrix

$$S(\delta,\tau) = D\, R(\delta,\tau)\, D \quad \text{with } D = \text{diag}\left(\frac{1}{s_{ii}}\right).$$

(see 6)

15) The condition number of a symmetric $A \geq 0$ ($p \times p$) is $\text{cond}(A) = \frac{\lambda_1(A)}{\lambda_p(A)} \geq 1$ where $\lambda_1(A) \geq \cdots \geq \lambda_p(A) > 0$ are the eigenvalues of A, A well conditioned matrix has $\text{cond}(A) \leq C$ for some number C such as 50 or 500. $R_\delta > 0$ for $\delta \geq 0$ and well conditioned if $\text{cond}(R_\delta) = \frac{\lambda_1 + \delta}{\lambda_p + \delta} \leq C$

or $\delta = \max\left(0, \frac{\lambda_1 - C\lambda_p}{C-1}\right)$ if $1 < C \leq 500$.

So $C = 50$ suggests $\delta = \max\left(0, \frac{\lambda_1 - 50\lambda_p}{49}\right)$.

16) For $R(\delta,\tau)$ compute I_p and $R(\delta,0)$, $R(\delta,0.05)$ for $C = 50, 100, 200, 300, 400$, and 500. compute R if $R > 0$.

As $\delta \to \infty$, $R_\delta \to I_p$ and I_p corresponds to $C = 1$.

So $S(\alpha,0) = :\text{diag}(S_{ii}) .= S_\delta$. See 5).

17) To get a regularized analog of R^{-1}, let

$$R_d = \begin{cases} R & \text{cond}(R) \le 500 \\ R(S=\alpha_0,0) & \text{else} \end{cases}$$

Let $A = R_d^{-1}$ the analog of R^{-1} to be regularized.

A acts like a covariance matrix. Let

$D_A = \text{diag}(\sqrt{A_{ii}})$. Then a generalized correlation matrix $R_I = D_A^{-1} A D_A$ is regularized with

$R_{I,d} = R_I(\delta,0)$ and $R_{I,t} = R_I(\delta,T)$. Then the regularized analogs of the inverse correlation matrix are $R_{INV,d} = D_A R_{I,d} D_A$ and

$R_{INV,t} = D_A R_{I,t} D_A$.

ex) $R = \begin{bmatrix} 1 & .4 \\ .4 & 1 \end{bmatrix}$, $R_{\delta=1} = \begin{bmatrix} 1 & .2 \\ .2 & 1 \end{bmatrix}$

$1+\delta = 2$

$R[.1,.2] = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, $R[1,.3] = R_{\delta=1}$

Ch 7 § 10.3, 10.4 Cluster analysis

1) Cluster Analysis is an example of unsupervised learning; there are no labelled responses (groups for classification).

2) Discriminant analysis is called supervised classification while clustering is called unsupervised classification; we try to find groups but the groups are unknown both for the training data and the test data. Sometimes there is no test data.

3) K means clustering:

i) partition the n training data cases into k initial groups and find the mean of each group. Alternatively choose k initial seed points. These are groups of size 1 so the mean is equal to the seed point.

ii) Compute distances between each training data case and each mean. Assign each case to the cluster whose mean is nearest.

iii) Recalculate the mean of each cluster.

iv) Go to ii) and repeat until no changes occur.

4) 2 problems i) There could be more or less than k clusters.

ii) 2 initial seed points could actually belong to the same cluster.

(could use several random sets of k seeds for several values of k, but need a method to choose the final group of clusters and k.)

5) Hierarchical clustering needs a distance.

Single linkage is the minimum distance between cases in cluster i and cases in cluster j.

Complete linkage is the maximum distance

"

Sometimes the average distance is used.

i) Start with $m = n$ clusters so each case forms a cluster. Compute the distance matrix for the n clusters. Let $d_{U,V}$ be the smallest distance. Combine clusters U and V into a single cluster and set $m = n-1$.

ii) Repeat i) with the new m. Continue

until there is a single cluster.

9ii) Plot the results in a dendogram. Use the dendogram to select k reasonable clusters of cases.

ex) (JW p 558-561) Consider the numbers 1-10 in 11 languages.

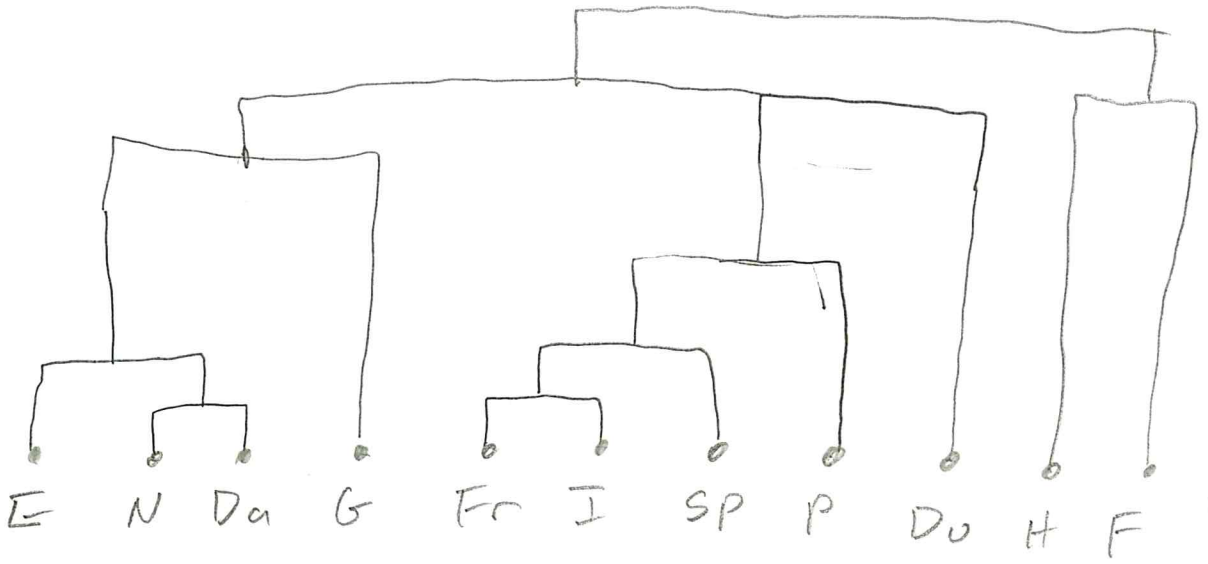
one en en een ein in uno uno jeden egy ytsi e

Look at 1st letter of each word. Two different languages are concordant if they have the same 1st letter.

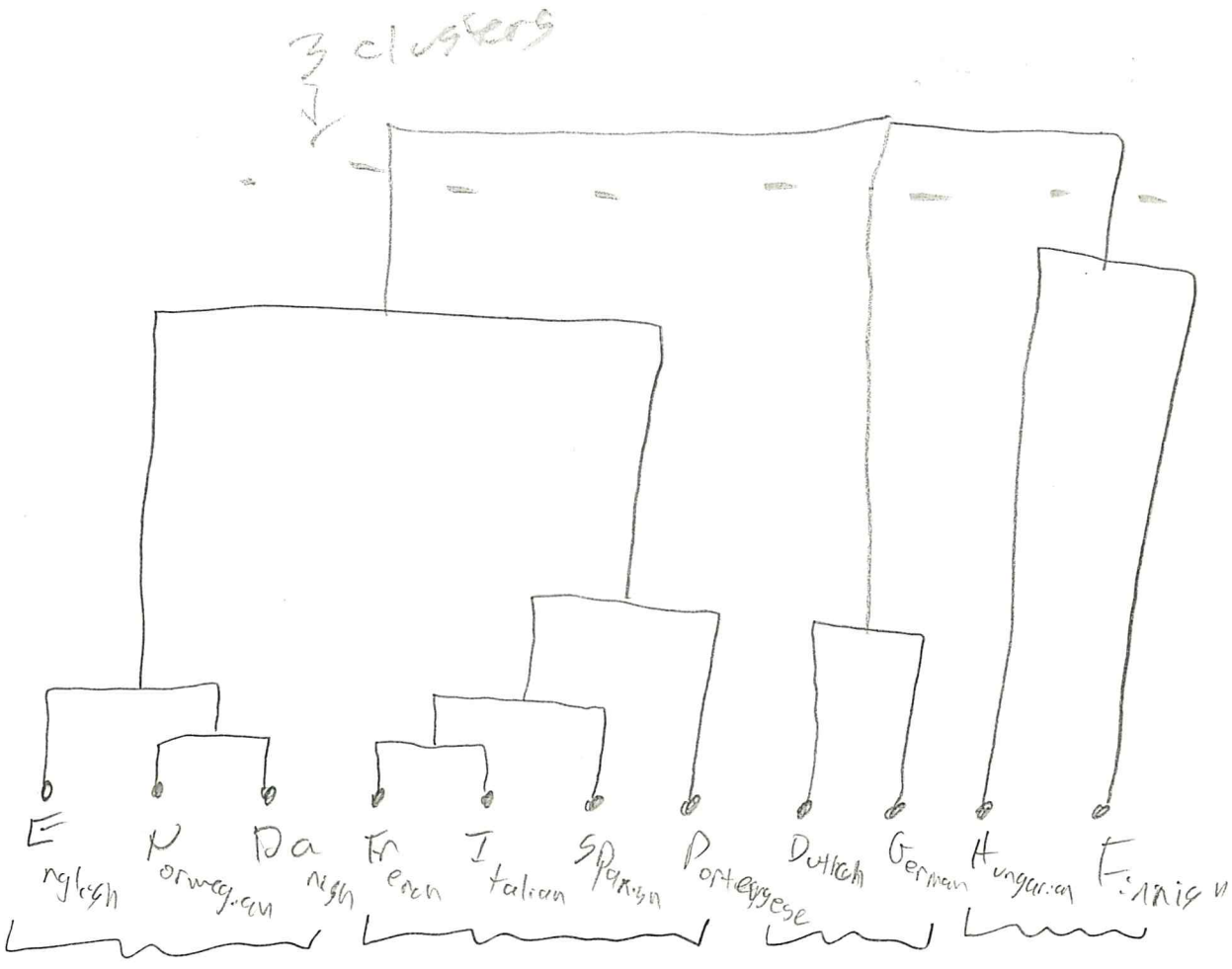
concordant 1st letters

	E	N	Da	Du	G	F	Sp	I	P	H	Fi
E	10										
N	8	10									
Da	8	9	10								
Du											
F											
Sp											
I											
P											
H											
Fi	1	1	1	. . .					1	2	10

10 — concordant matrix
= distance matrix



complete linkage



single linkage

vertical height represents closeness of clusters

could cluster with a horizontal line

could use 4 clusters

could use 3 clusters.

6) Dendrograms are hard to use with large data sets. Need an automated method to choose the final k clusters for large data sets.

→ often take training data with known groups to check the cluster analysis

J ch 8 Regression and Classification Trees

1) A regression tree is a flexible method for $Y = m(X) + \varepsilon$ or for $Y_i = m(X_i) + \sigma_i \varepsilon_i$. A classification tree is a flexible method for classification. Both methods produce graphs called trees that look like dendrograms. Each branch has a label like $\underbrace{X_i > 7.56}_{X_i \text{ quantitative.}}$ or $\underbrace{X_j = a, c}_{X_j \text{ a factor taking on levels } a, b, c, d, e, f \text{ say}}$

Unless told otherwise, if the condition is true, go to the left of the branch. Otherwise go to the right of the branch. ^ The bottom of the tree has leaves that give $\hat{Y} = \hat{Y}(x)$. $\hat{Y}$ is a number for a regression tree and

Some software switches this. Our text does.

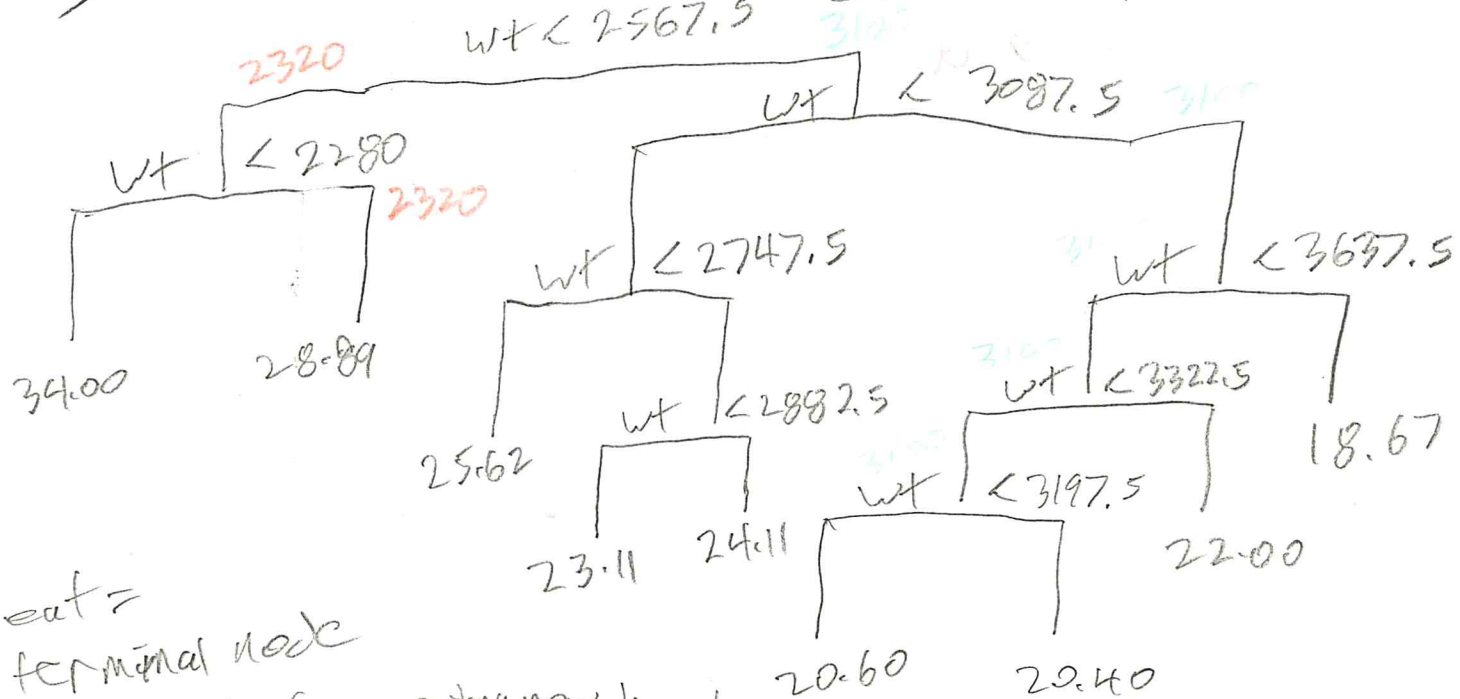
$\hat{y}$ is a label for the classification group for a classification tree. For example if $Y = \begin{cases} F = 0 \\ M = 1 \end{cases}$, we could have $\hat{y} \in \{0,1\}$ or $\hat{y} \in \{F, M\}$.

2) Know (Splus2000 Guide to Statistics Vol I 372-374)

Given a tree and $\underline{x}$ values, find $\hat{y}$, see HW10 c), D),

ex) Y = gas mileage X = weight of car

∈ root = top node



leaf = terminal node

split = rule for creating new branches

a) Predict mileage if wt = 2320

$\hat{y} = 28.89$

b) Predict milage if wt = 3100 $\hat{y} = 20.60$

ex) (Berk 2008 p145-147) Prison data

y = no misconduct, minor misconduct, or major misconduct
0 1 2

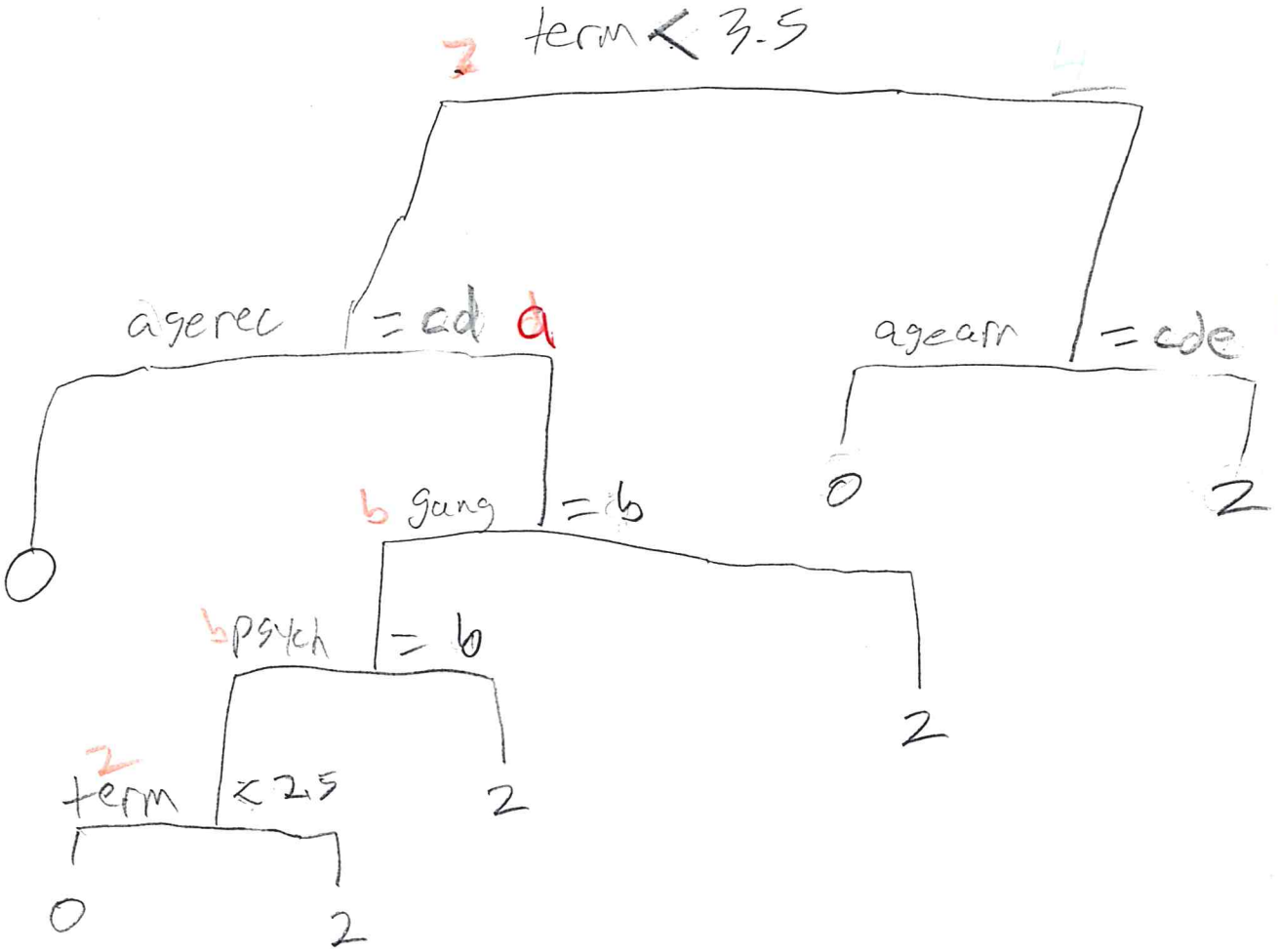
x_1 = term = sentence length in years

x_2 = agerec (enter prison): a: 16-20 b: 21-26, c: 27-35, d: 236

x_3 = agearr (1st arrest): a: 0-17, b: 18-21, c: 22-29, d: 30-35, e: 236

x_4 = gang: a = gang activity b = no gang activity

x_5 = psych: a = mental illness b = no mental illness



a) classify if x_1 = term = 2 x_2 = agerec = d

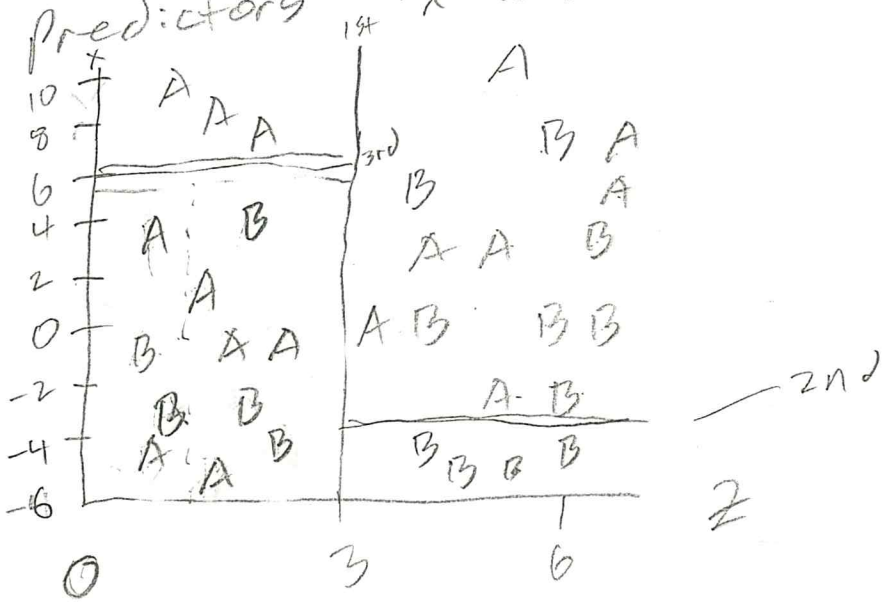
x_4 = gang = b, x_5 = psych = b

y = 0 red

b) classify if x_1 = term = 4; x_3 = agearr = b

3) Trees that use recursive partitioning for classification and regression trees use the CART algorithm

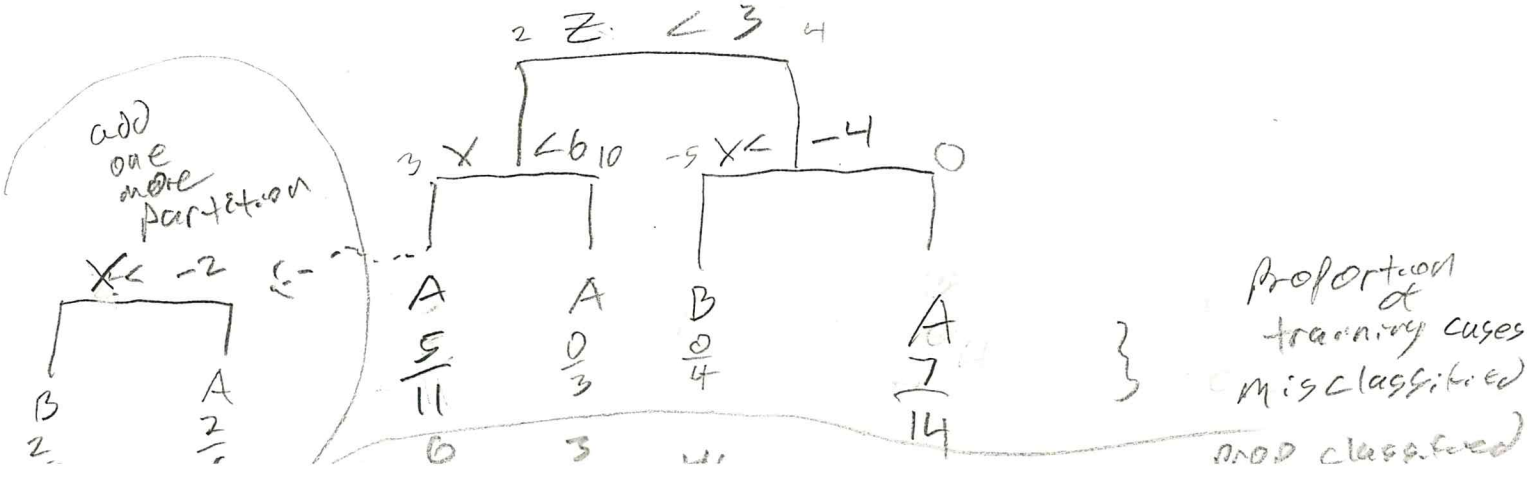
ex) Y = A or B (classification) with Predictors X and Z



A single vertical line at Z=3 is the 1st partition,

The horizontal line at X=6 is the 2nd partition.

" X=-4 3rd "



4) Trees i) give prediction rules that can be rapidly and repeatedly evaluated)

ii) are useful for screening predictors (variable selection, interactions))

iii) can be used to assess the adequacy of linear models

iv) can be used to summarize large multivariate data sets

5) If Y is a factor, (has levels 1,...,K) classification rules are of the form if $x_1 \leq 2.3$ and $x_3 \in \{A, B\}$ then Y is most likely to be in level 5.

If Y is numerical, regression rules are of the form if $x_2 \leq 2.3$ and $x_4 \in \{C, D, F\}$ and $x_5 \leq 3.5$ then $\hat{y} = 4.75$.

6) Trees can be easier to interpret when some predictors are numerical and some categorical

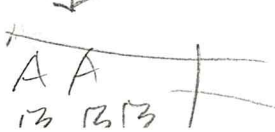
Trees are invariant to monotone (increasing or decreasing) transformations of the predictor variable X_i

Trees handle missing values better than MLR. Trees beat MLR if there is nonadditive behavior. Trees can handle complex unknown interactions. For classification Y can have more than 2 levels.

7) In growing a tree, the binary partitioning algorithm recursively splits the data in each node until either the node is homogeneous (0 roughly training data misclassifications for a classification tree, $Y \simeq$ constant for a regression tree) or the node contains too few observations (default ≤ 5).

8) The <u>deviance</u> is a measure of node homogeneity and deviance = 0 for a perfectly homogeneous node.

9) Often { use the mean of the region for $\hat{Y}$, regression tree
mode, classification tree



10] J p306-7 a) Divide the predictor space

(= set of possible values for $\underbrace{X_1, \ldots, X_p}_{\text{training data}}$) into

J distinct and nonoverlapping regions $R_1, \ldots, R_J$.

b) For every observation that falls in region R_j, make the same prediction

$\hat{y}_{R_j} \equiv \begin{cases} \bar{y}_j & = \text{sample mean of training data } y \text{ in } R_j, \text{ regtree} \\ \text{mode}_j = & \text{class tree} \end{cases}$

11) The R_j are high dimensional boxes.

Choose R_j so $RSS = \sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$

is small.

$I(y_i \neq \hat{y}_{R_j})$ } class

12) Let $\{\underline{X} \mid X_j < s\}$ be the region in the predictor space such that $X_j < s$ where $\underline{X} = (X_1, \ldots, X_p)'$.

s is small

Define 2 ~~half hyperplanes~~ regions

$R_1(j,s) = \{\underline{X} \mid X_j < s\}$, $R_2(j,s) = \{\underline{X} \mid X_j \geq s\}$ and seek

"cutpoint" s and j to minimize

$\sum_{i : \underline{x}_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i : \underline{x}_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$.

This can be done "quickly" if p is small, SL 50 (could use order statistics).

Then repeat the process looking for the best predictor and best cutpoint in order to split the data further so as to minimize the RSS within each of the resulting regions. Only split one of the regions. So now there are regions R_1, R_2, and R_3. Continue this process until a stopping criterion is reached such as no region contains more than 5 obs's (and stop splitting if the region is homogeous eg all Y in region belong to class k).

13] (For classification) such trees are usually not competitive with earlier techniques Bagging, random forests and boosting makes trees more competitive.

14] Trees use regions $R_1, \ldots, R_J$. If J is too large, the tree overfits. One strategy is to grow a large tree T_0 with J_0 regions

[5] J p308 cost complexity pruning = weakest link pruning

Let $T \subseteq T_0$, $\alpha \geq 0$, and $|T|$ = # terminal nodes of tree T_0. Each terminal node corresponds to a region (hyper rectangle) R_j. Let R_m be the region corresponding to the mth terminal node and $\hat{y}_{R_m}$ be the predicted response for R_m,

For each value of $\alpha > 0$, there corresponds a subtree $T \subseteq T_0$ such that $\sum_{m=1}^{|T|} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$ is as

- $I(y_i \neq \hat{y}_{R_m})$ for class tree

(*) small as possible. Note that $\alpha = 0$ has $T = T_0$ and (*) = RSS(T_0) = training data RSS for T_0.

[6] Much like lasso, as α increases there is a sequence of nested subtrees

(**) $T_0 \supseteq T_{\alpha_1} \supseteq T_{\alpha_2} \supseteq \cdots \supseteq T_{\alpha_m}$. Branches get "pruned" from T_0 in a nested and predictable fashion,

[8] a) Build tree T_0, stopping when each terminal node has ≤ 5 obs.

c) Use K-fold CV to choose $\alpha = \alpha_{j*}$.

For each $i \in 1, \ldots, K$

i) Repeat steps a) and b) on all but the ith fold.

ii) Evaluate the mean square prediction error on data in the left out fold i as a function of α.

Average the results for each value of α and pick α_{j*} to minimize the average error.

d) Use tree $T_{\alpha_{j*}}$ from b),

For a regression tree

$$MSE_i = \frac{1}{n_i} \sum_{j=1}^{n_i} \left(y_j - \hat{y}_j(i)\right)^2$$ for data y_j in left out fold i

$$CV_{(K)} = \frac{1}{K} \sum_{i=1}^{K} MSE_i$$

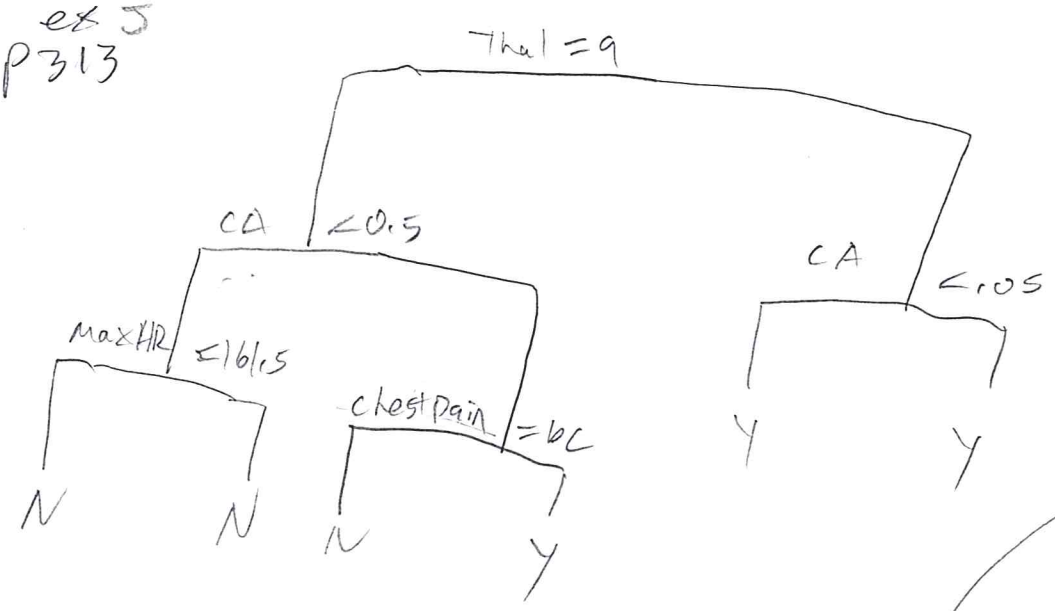
For a classification tree use

$$MSE_i = \frac{1}{n_i} \sum_{j=1}^{n_i} I\left(y_j \neq \hat{y}_j(i)\right) = \text{proportion}$$

misclassified in the ith fold.

If $n_i \equiv \frac{n}{K}$ then $CV_{(K)} = \begin{cases} \frac{1}{n} \sum_{j=1}^{n} \left(y_{ji} - \hat{y}_j(i)\right)^2 & \text{reg tree} \\ \frac{1}{n} \sum_{j=1}^{n} I\left(y_{ji} \neq \hat{y}_j(i)\right) & \text{class tree} \end{cases}$

ex 5
p313



← subtree
6 terminal nodes

Big tree T_0 added many more branches and had 18 terminal nodes.

303 patients with chest pain

Y = heart disease

N = no heart disease

Thal = Thalium stress test a = normal

Chest pain a = typical aginal b = atypical aginal
c = non-aginal pain d = asymptomatic

maxHR max-mum heart rate? CA calcium in blood?

18) For a tree T_α, $Y_i = m(\underline{x}_i) + \sigma_i \varepsilon_i$ and

$$\hat{Y} = \hat{m}(\underline{x}_i) = \sum_{m=1}^{J_\alpha} c_m I(\underline{x}_i \in R_m)$$ where T_α

uses regions $R_1, \ldots, R_{J_\alpha}$.

19) Trees can handle categorical variables (factors, qualitative variables) without creating indicators = dummy variables.

20] Bagging was used before.

Compute $T_1^*, \ldots, T_B^*$ with the bootstrap and the sample mean $\bar{T}^* = \frac{1}{B}\sum_{i=1}^{B} T_i^*$ is the bagging estimator.

21] For a regression tree, draw a sample of size n with replacement from $x_1, \ldots, x_n$. Fit the tree and find $\hat{f}_1^*(x)$. Repeat to get $\hat{f}_1^*(x), \ldots, \hat{f}_B^*(x)$. Then the bagging estimator

$$\hat{f}_{bag}^*(x) = \frac{1}{B}\sum_{i=1}^{B} \hat{f}_i^*(x).$$

The trees are not pruned so terminate when each terminal node has 5 or fewer obs's.

22] For classification draw a sample of size n_i from each group with replacement.

Let $\hat{f}_i^*(x) = j_i(x) \in \{1, \ldots, G\}$ where Y takes on levels $1, \ldots, G$. Compute $\hat{f}_1^*(x), \ldots \hat{f}_B^*(x)$ and let $m_k = \#\{j_i(x) = k\}$ for $k = 1, \ldots, G$. Take

$\hat{f}^*(x) = \max_k m_k$

23) (5 p197) The prob that x_j is not in the bootstrap sample $\rightarrow e^{-1} \approx .3679 \approx \frac{1}{3}$ as $n \to \infty$.

$$\frac{1-\frac{1}{n}}{1} \quad \cdots \quad \frac{1-\frac{1}{n}}{n}$$

$$\left(1-\frac{1}{n}\right)^n \to e^{-1}$$

prob x_j is not the ith obs $= 1-\frac{1}{n}$ since each obs has $\frac{1}{n}$ chance of being selected for the ith position.

24) For each bootstrap sample b

let $x_{i_1}, \ldots, x_{i_{k_b}}$ be the k_b obs's not used. These are the out of bag (OOB) obs's. Predict $\hat{y}$ for each OOB obs. Doing this for all B bootstraps produces about $\frac{B}{3}$ predictions for each x_i. Let $\hat{y}_{i0} = \begin{cases} \text{ave } \hat{y}_i & \text{reg tree} \\ \text{mode level} & \text{class tree} \end{cases}$

$$\text{OOB MSE} = \begin{cases} \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_{i0})^2 & \text{reg} \\ \frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_{i0}) & \text{class} \end{cases}$$

The OOB MSE is "virtually equivalent" to the leave one out CV estimate, for sufficiently large B.

25) Bagging typically gives better accuracy

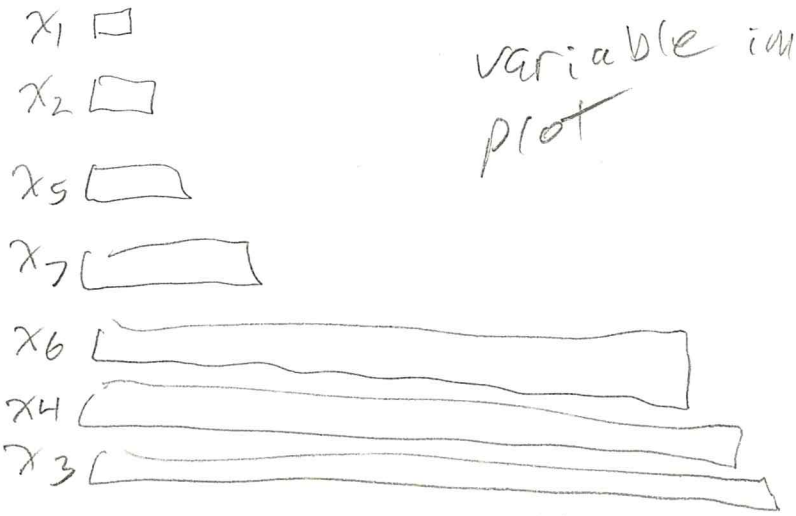
26) For classification trees, let

$\hat{p}_{mk}$ = proportion of training obs's in R_m from the kth class.

Gini's index = $\sum_{k=1}^{C} \hat{p}_{mk}(1-\hat{p}_{mk})$

is small if all $\hat{p}_{mk}$ are close to 0 or 1.

27) JP 319 For bagging with B trees a measure of variable importance can be computed for each variable using splits for each variable.



variable importance plot

x_3, x_4, x_6 are the most important predictors for the tree

28) If $Y = \alpha + \sum_{j=1}^{p} f_j(x_j) + \epsilon$ or $Y = m(x) + \epsilon$ with $m(x) = g(\alpha + \beta^T x)$ then slicing the e[illegible]

For [illegible] $\hat{f}_j(x_j)$ [illegible] $\hat{\alpha} + \hat{\beta}^T x$

is more effective than partitioning the predictor space with hyper boxes R_k.

$Y = g(\alpha + \beta^T X) + \varepsilon$

curse of dimensionality $\mathbb{R}^p$

If $Y = g(w)$ and you plot w vs Y, then you see g.
So if $Y = g(ESP) + \varepsilon$ you see g" up to noise of $ESP \approx SP = w$ and ε_i.

My idea. Predict

29) If $Y = m(x) + \varepsilon$, we can make prediction intervals for Y_f with the regression tree using $\hat{Y} = ESP = \hat{m}(x)$ and $r = Y - \hat{Y}$ as before.

§8.2.2 Random Forests

30) For random forests the bootstrap is used, but each time a split is considered, a random sample of $m = \lfloor \sqrt{p} \rfloor$ predictors is chosen as split candidates. Random forests produces bootstrap trees that are less correlated than the bagged trees (that use $m = p$) and the random forests estimator tends to have smaller test error and OOB error

than the bagging estimator.

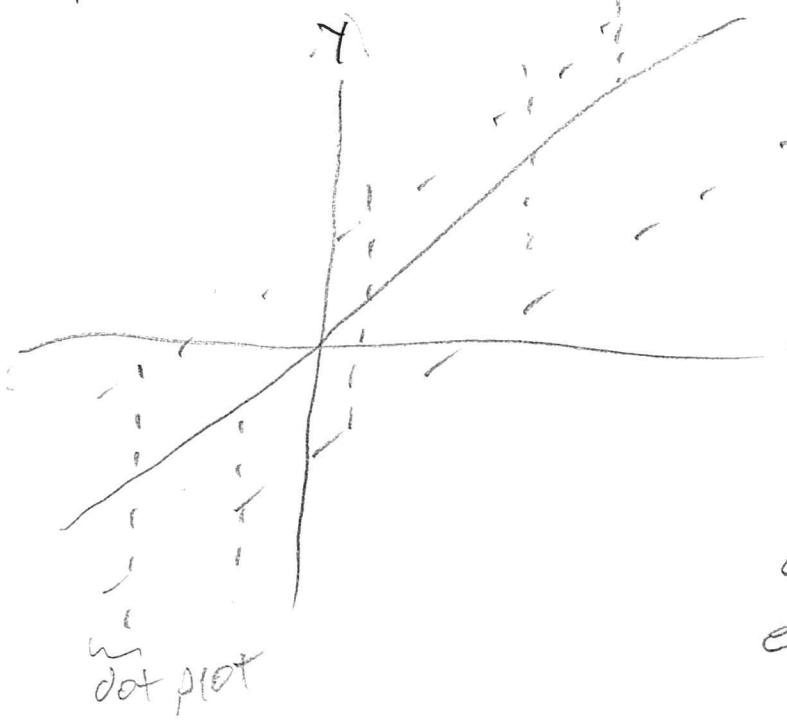
31) B around a few hundred seems to work.

32) If there is a single strong predictor bagged trees tend to use that predictor in the 1st split. For random forests, the strong predictor is not considered for $\frac{p-m}{p}$ splits, on average. Trees from random forests also tend to be less correlated than bagged trees if there are many correlated predictors,

JP321

33) For classification, the null classifier has $\hat{y} \equiv d$ where d is the dominant class. So if k% of obs's belong to the dominant class, then the test error $= \frac{100-k}{100} \leq 1 - \frac{1}{G}$ where there are G groups since $k \geq 100\frac{1}{G}$. Classifiers that do not beat the null classifier are very bad.

34) Since a tree uses J_α regions, the response plot of ESP $= \hat{Y} = \hat{m}(x)$ vs Y looks like

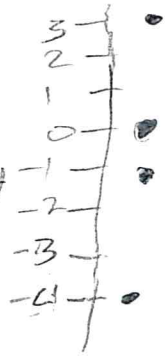


The plotted points scatter about the identity line, but there is a dot plot of n_m cases with $\hat{Y} = \hat{Y}_{Rm}$ for each of the J_α regions

(One way Anova models have a similar dot plot, but each dot plot crosses the identity line at $\hat{Y}_i = \overline{Y_i}$.)

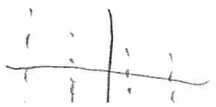
35) A dot plot of $Z_1, \ldots, Z_m$ consists of an axis and m points corresponding to the values of Z_i

ex $Z_1 = 3$ $Z_2 = -1$ $Z_3 = 0$, $Z_4 = -4$

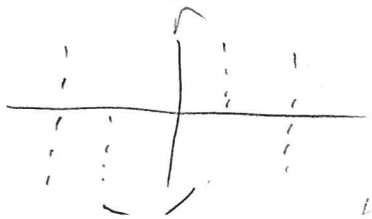


← dot plot

36) The residual plot consists of dot plots that scatter about the $r = 0$ line.



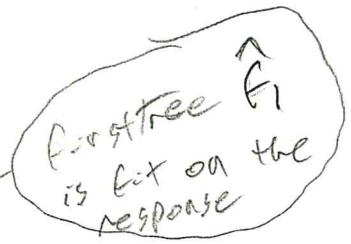
← nonconstant variance $Y_i = m(x_i) + \sigma_i e_i$



← Fit

37] JG8.2.3 The third technique for improving trees is boosting. Like bagging, boosting can be applied to many statistical methods, not just trees.

38] Boosting for regression trees



i) Set $\hat{f}(x) \equiv 0$ and $r_i = y_i$, $i = 1, \ldots, n$ ← training data

ii) For $b = 1, \ldots, B$

a) fit tree $\hat{f}_b$ with d splits ($d+1$ terminal nodes) to the training data (X, r)

b) update $\hat{f}(x)$ by adding a shrunken version of the new tree: $\hat{f}(x) \leftarrow \hat{f}(x) + \lambda \hat{f}_b(x)$ and update the residuals $r_i \leftarrow r_i - \lambda \hat{f}_b(x)$.

iii) The boosted model

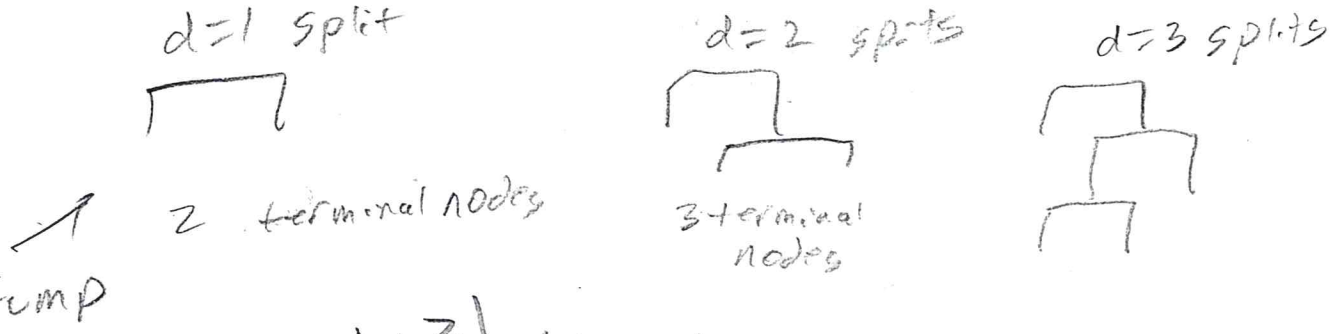
$$\hat{f}(x) = \sum_{b=1}^{B} \lambda \hat{f}_b(x).$$

39] The tree is fit to updated residuals rather than y. This slowly improves $\hat{f}$ in areas where it does not perform well, and λ slows the learning process further.

As a rule of thumb, iterative techniques that learn slowly tend to perform well.

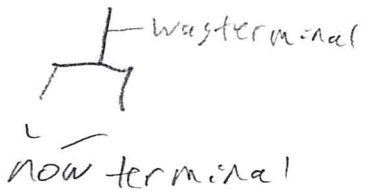
40) Boosting can overfit if B is too large.

k fold CV is used to select B.

41) 

A tree with d-1 splits has d terminal nodes

Then a tree with d splits adds the split to one of the nodes replacing the terminal node with 2 terminal nodes.



Hence a tree with d splits has d+1 terminal nodes.

42) often d=1 or 2 is used. A d=1 tree is called a stump. The value d is called the interaction depth.

43) The value λ tends to be .01 or .001. very small λ tends to need very large B for good performance.

4) JP 332 Using the d=1 stumps leads to an additive model

$$\hat{f}(x) = \sum_{j=1}^{p} \hat{f}_j(x_j).$$

I can't prove this yet.

Each stump splits one variable

Let $r^{(0)} = y$, $\hat{f}^{(b)} = \hat{f}^{(b-1)} + \lambda \begin{pmatrix} \hat{f}_b(x_1) \\ \vdots \\ \hat{f}_b(x_n) \end{pmatrix}$ with $\hat{f}^{(0)} = 0$

$$r^{(b)} = r^{(b-1)} - \lambda \begin{pmatrix} \hat{f}_b(x_1) \\ \vdots \\ \hat{f}_b(x_n) \end{pmatrix}.$$

JP 332

5) For a binary classification tree with $Y = 0$ or 1, for a fixed value of x (a test data point), the bootstrap produces B estimates of $P(Y=1|x)$.

Two common ways to get $\hat{Y}|x$ are

a) $\hat{Y}|x$ = mode class of 0 or 1

b) average the prob $\hat{Y}|x = \begin{cases} 0 \\ 1 \end{cases}$ if $\hat{P}(Y=1|x) < 0.5$ / $>$

ex $\hat{P}(Y=1|x) \overset{B=10}{=} .1, .15, .2, .2, .55, .6, .6, .65, .7, .75$

class 0 0 0 0 1 1 1 1 1 1

a) gives $\hat{Y}|x = 1$ ... $.1 + .75 = 4.5 > .45$ so $\hat{Y}|x = 1$

A note on the residual bootstrap for regression

Suppose $\underline{Y} = X\underline{\beta} + \underline{e}$ $\quad e_i$ iid with $V(e_i) = \sigma^2$

$\mathrm{cov}(\underline{e}) = \sigma^2 I_n = \mathrm{cov}(\underline{Y})$

*) $\sqrt{n}(\hat{\underline{\beta}} - \underline{\beta}) \xrightarrow{D} N_p(\underline{0}, \sigma^2 \underline{W})$. (Efron 1982 P36)

The residual bootstrap has

$$\underline{Y}^* = X\hat{\underline{\beta}} + \underline{r}^w$$

is a general linear model with respect to the bootstrap sample since the r_i^w are iid from the distribution of $\underline{\varepsilon}$ which is the empirical distribution of the residuals $r_1, \ldots, r_n$ (each with probability $\frac{1}{n}$).

So $E(r_i^w) = \sum_{i=1}^n \frac{1}{n} r_i = \bar{r} = 0$ since $\sum_{i=1}^n \hat{r}_i = 0$ for OLS residuals.

So $V(r_i^w) = E(r_i^w)^2 = \sum_{i=1}^n \frac{1}{n} r_i^2 = \frac{n-p}{n} \frac{1}{n-p} \sum_{i=1}^n r_i^2 = \frac{n-p}{n} MSE$

and $E(\underline{r}^w) = \underline{0}$, $\mathrm{Cov}(\underline{Y}^*) = \mathrm{Cov}(\underline{r}^w) = \frac{n-p}{n} MSE\, I_n$

$E(\underline{Y}^*) = X\hat{\underline{\beta}} = H\underline{Y}$

$\sigma_n^2 I_n - \sigma^2 I_n \to \underline{0}$ as $n \to \infty$

So by standard linear model theory with $\frac{X^T X}{n} \to W^{-1}$

**) $\sqrt{n}(\hat{\underline{\beta}}^* - \hat{\underline{\beta}}) \xrightarrow{D} N_p(\underline{0}, \sigma^2 W)$. $\mathrm{Cov}(\hat{\underline{\beta}}^*) = \frac{n-p}{n} MSE (X^T X)^{-1}$

Here $\hat{\underline{\beta}}^* = (X^T X)^{-1} X^T \underline{y}^*$. $\hat{\underline{\beta}}^*_{I_j} = (X_{I_j}^T X_{I_j})^{-1} X_{I_j}^T \underline{Y}^*$

Ch 9 Support Vector Machines (SVMs) SL 57

1) Logistic regression is used a lot in biostatistics and epidemiology where the focus is statistical inference. Support vector machines are used in machine learning where the goal is classification accuracy.

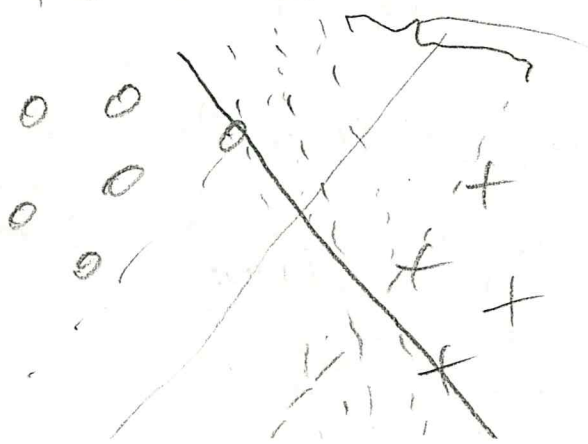
2) When $p >> n$ there is often a hyperplane that perfectly separates 2 groups. The launching point for SVMs was finding the optimal separating hyperplane.

3) For 2 groups let $SP = \beta_0 + \underline{\beta}'\underline{x}$. ($\beta_0$ is a scalar, $\underline{\beta}$ is $p \times 1$)

Classify x in group 1 if $ESP > 0$ and group -1 if $ESP < 0$ (just like LR, but code group 0 as -1).

So the classifier $\hat{C}(x) = \text{sign}(ESP)$

4)

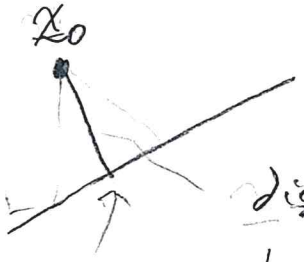


margin between classes

The estimated "optimal separating hyperplane" ESP has the largest margin on the training data.

5) Let $ESP = \hat{\beta}_0 + \underline{\hat{\beta}}^T \underline{x}$ and $f(\underline{x}) = SP = \beta_0 + \underline{\beta}^T \underline{x}$

so $\hat{f}(\underline{x}) = ESP$. $y_i \in \{-1, 1\}$.

6) 

distance of $\underline{x}_0$ from the decision boundary

$$= \frac{\beta_0 + \underline{\beta}^T \underline{x}_0}{\|\underline{\beta}\|_2} = \frac{f(\underline{x}_0)}{\|\underline{\beta}\|_2}$$

Projection of $\underline{x}_0$ onto hyperplane must be $\beta_0 + \underline{\beta}^T \underline{x}_0$

7) Think of the hyperplane $\beta_0 + \underline{\beta}^T \underline{x}_i =$

$\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ dividing $\mathbb{R}^p$ into 2 halves. A separating hyperplane has $SP > 0$ if $x \in$ group 1, < 0 group -1

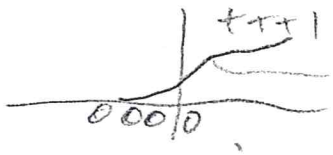
So $y_i \, SP_i = y_i (\beta_0 + \underline{\beta}^T \underline{x}_i) > 0$ for $i = 1, \ldots, n$.

PhD idea? maybe with bootstrap. Probably not SVM uses $\beta_0 + \beta^T w$ where w are functions of $\underline{x}$

8) Think of a binary classifier that uses ESP as a binary regression

$Y|x \sim bin(n=1, s(\underline{x}))$ where $s(\underline{x}) = s(SP) = P(Y = 1|\underline{x})$

where $Y = -1$ has been recoded as $Y = 0$. but $s(SP)$ is unknown. Response plot



* The bootstrap with n_i cases selected with replacement from each group is likely useful if $n >> p$.

9) CART splits $\mathbb{R}^p$ with regions $R_m \in \mathbb{R}^p$

SVM splits $\mathbb{R}^p$ into 2 regions using ESP ER so there is dimension reduction. The SVM split tries to make the 2 "halves" or partitions as homogeneous as possible.

10) The hyperplanes parallel to ESP that form the boundaries of the margin are called fences. The fences pass through at least 2 training data cases. These cases form the support set S. (of support vectors.) It turns out that

$$\hat{\beta}_M = \sum_{i \in S} \hat{\alpha}_i x_i.$$

↑ optimal margin classifier

11) Wide data has $p >> n$. If $n \leq p+1$ there is a separating hyperplane unless there are exact predictor ties across the class barrier (whatever that means).

12) Let M be the margin. The optimal margin classifier $\hat{\beta}_{0M}, \hat{\beta}_M$ maximizes M subject

to $y_i \cdot f_i = y_i(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}) \geq M \quad \forall i = 1, \ldots, n.$ (*)

This is called a hard margin classifier since no cases from either group can pass the fences of the classifier. Equivalently $\min_{\beta_0, \underline{\beta}} \|\underline{\beta}\|_2$ subject to (*).

13) A soft margin classifier allows cases from either group to pass the fences or to be misclassified. This classifier solves

minimize $\|\underline{\beta}\|_2$ subject to
$\beta_0, \underline{\beta}$

$y_i(\beta_0 + \underline{x}_i'\underline{\beta}) \geq 1 - \varepsilon_i, \quad \varepsilon_i \geq 0 \quad$ for $i = 1, \ldots, n$

and $\sum_{i=1}^{n} \varepsilon_i \leq B$.

slack variables are used in linear programming.

14) This minimization is equivalent to

$$\text{minimize}_{\beta_0, \underline{\beta}} \sum_{i=1}^{n} \underbrace{\left[1 - y_i(\beta_0 + \underline{x}_i'\underline{\beta})\right]_+}_{\text{hinge loss} = \begin{cases} 0 & \text{if } \underline{x}_i \text{ is on the correct side} \\ \text{cost of } \underline{x}_i \text{ being on the wrong side of the margin} \end{cases}} + \lambda \|\underline{\beta}\|_2^2$$

where $[w]_+ = \begin{cases} w & w \geq 0 \\ 0 & w \leq 0. \end{cases}$

This is similar to ridge regression.

15) A support vector machine that uses $\underline{x}_i$ minimizes the above loss criterion. For separable data

$(\hat{\beta}_{0SVM}, \hat{\underline{\beta}}_{SVM}) \to (\hat{\beta}_{0M}, \hat{\underline{\beta}}_{M})$ as $\lambda \to 0$.

16] It turns out that $\hat{\underline{\beta}}_{SVM} = \sum_{i \in S} \hat{\gamma}_i \underline{x}_i$

and $\hat{\beta}_{0,SVM} + \underline{x}^T \hat{\underline{\beta}}_{SVM} = \hat{\beta}_{0SVM} + \sum_{i \in S} \hat{\gamma}_i \langle \underline{x}, \underline{x}_i \rangle$

where $\langle \underline{x}, \underline{x}_i \rangle = \underline{x}^T \underline{x}_i$. This quantity can be computed using the $n \times n$ Gram matrix $\underset{n \times n}{\mathbb{X}\mathbb{X}^T}$ with $O(n^2 p)$ complexity. Using $\underset{p \times p}{\mathbb{X}^T \mathbb{X}}$ has $O(n p^2)$ complexity. Ridge regression could also be computed this way.

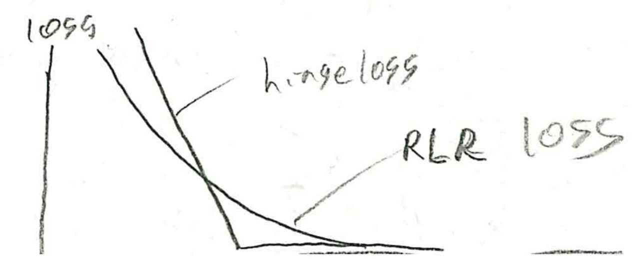
17] A lasso-SVM solves

$$\underset{\beta_0, \underline{\beta}}{\text{minimize}} \sum_{i=1}^{n} [1 - Y_i(\beta_0 + \underline{x}_i^T \underline{\beta})]_+ + \underbrace{\lambda \|\underline{\beta}\|_1}_{\text{like lasso}}$$

and does variable selection.

] "Ridged logistic regression" with $Y_i \in \{-1, 1\}$

ves $\underset{\beta_0\ \underline{\beta}}{\text{minimize}} \sum_{i=1}^{n} \log[1 + e^{-Y_i(\beta_0 + \underline{x}_i^T \underline{\beta})}] + \lambda \|\underline{\beta}\|_2^2$



19) * Truth table = confusion matrix

truth

predict	1	-1
1	18	3
-1	7	7

misclassified

correct

$$\text{error rate} = 1 - \frac{18+7}{18+7+3} = 1 - \frac{25}{28} = \frac{3}{28} = 0.1071$$

diagonal: correctly classified
off diagonal: incorrectly classified

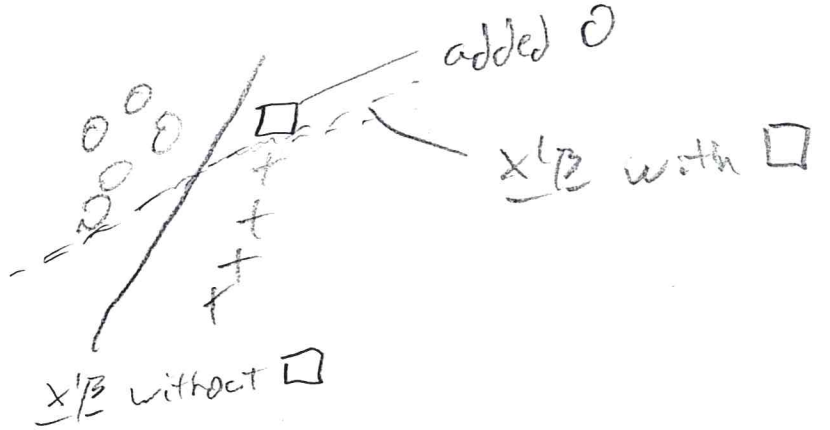
predict \ truth	a	b	c	d
a	10	0	0	12
b	0	100	11	0
c	0	5	50	0
d	6	0	0	30

corrected

$$\text{error rate} = \frac{6+5+11+12}{6+5+11+12+10+100+150+30} = \frac{34}{224} = 0.1518$$

See HW 11

20) Sometimes 1 or a few obs's shift the maximal margin hyperplane. The SVM classifier is a soft margin classifier and can do better.



18
OM
from Sol

21) SVM maximizes M = width of margin subject to

$\varepsilon_i \geq 0$, $\sum_{i=1}^{n} \varepsilon_i \leq B$, $y_i(\beta_0 + \beta^T x_i) \geq M(1-\varepsilon_i)$ for all i.

$\hat{C}(\underline{x}) = \text{sign}(\beta_0 + \underline{\beta}'\underline{x})$. (I use G for # of classes)

21) A slack variable $\varepsilon_i = 0$ if $\underline{x}_i$ is on the correct side of the margin. If $\varepsilon_i > 0$ then $\underline{x}_i$ is on the wrong side of the hyperplane.

$y_i(\beta_0 + \underline{\beta}'\underline{x}_i) \geq M$ has $\varepsilon_i = 0$ and is necessary for $\underline{x}_i$ to be on the correct side of the margin. If $y_i(\beta_0 + \underline{\beta}'\underline{x}_i) \geq M(1-\varepsilon_i)$ with $\varepsilon_i > 0$ (but not if $\varepsilon_i = 0$), then $\underline{x}_i$ is on the wrong side of the hyperplane. See 7)

22) Let the kernel function be $k(\underline{x}_i, \underline{x}_{i'})$. A linear kernel is $k(\underline{x}_i, \underline{x}_{i'}) = \underline{x}_i^T \underline{x}_{i'}$. A polynomial kernel of degree d is $k(\underline{x}_i, \underline{x}_{i'}) = (1 + \underline{x}_i^T \underline{x}_{i'})^d$. A radial kernel is $k(\underline{x}_i, \underline{x}_{i'}) =$

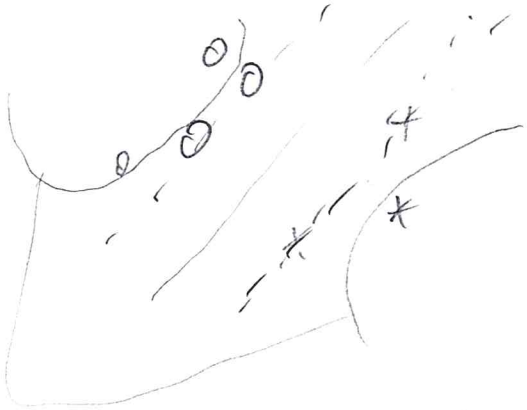
$$\exp\left[-\gamma \sum_{k=1}^{p} (x_{ik} - x_{i'k})^2\right] = \exp\left[-\gamma \|\underline{x}_i - \underline{x}_{i'}\|_2^2\right].$$

23) A SVM uses $f(\underline{x}) = \beta_0 + \sum_{i \in S} \alpha_i k(\underline{x}, \underline{x}_i)$

$$= \beta_0 + \sum_{i \in S} \alpha_i K(\underline{x}, \underline{x}_i) = SP = SP(\underline{x})$$

where S is the index of support vectors.

Note: Support vectors determine the hyperplane and margin: if they are moved the hyperplane moves too.



these points are not support vectors

24) Using $K(\underline{x}, \underline{x}_i)$ leads to nonlinear decision boundaries if K is nonlinear. The kernel is a bivariate transformation. There are $\binom{n}{2} = \frac{n(n-1)}{2}$ distinct pairs $(\underline{x}_i, \underline{x}_j)$ that are needed to estimate β_0 and the α_i.

"SVM with" $f(\underline{x}) = \hat{\beta}_0 + \sum_{i=1}^{\hat{}} \hat{\alpha}_i K(\underline{x}, \underline{x}_i) = ESP = ESP(\underline{x})$ is a competitor for QDA while the SVM with $f(\underline{x}) = \hat{\beta}_0 + \hat{\underline{\beta}}' \underline{x}$ is a competitor for LDA.

radial kernel

25) p353 If $\underline{x}$ is far from $\underline{x}_i$, then

$\|\underline{x} - \underline{x}_i\|_2^2$ is large so

$K(\underline{x}, \underline{x}_i) = \exp\{-\gamma \|\underline{x} - \underline{x}_i\|_2^2\}$ is tiny

and $\underline{x}_i$ has almost no contribution in $f(\underline{x})$.

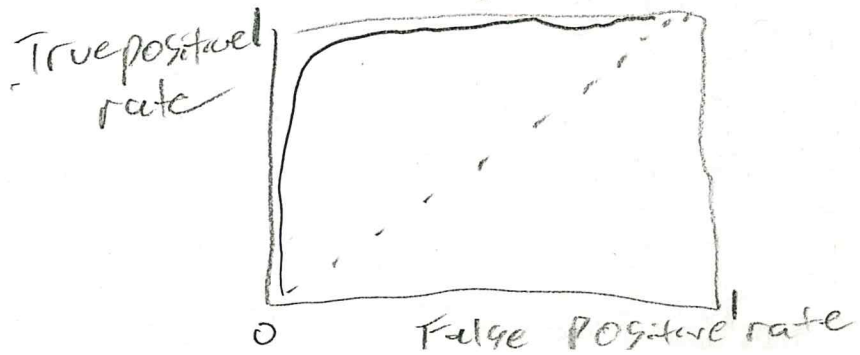
Analogy KNN.

(recever operating characteristic from quality control)

263 p147, 354 A ROC curve is used to evaluate binary classifiers. The overall performance is summarized by the area under the ROC curve (AUC). An ideal ROC curve is close to the top left corner of the plot, so the larger the AUC the better the classifier.

Predict \ truth	-1	1	total
-1	true negative TN	FN False negative	N^*
1	FP False Positive	TP true positive	P^*
total	N	P	

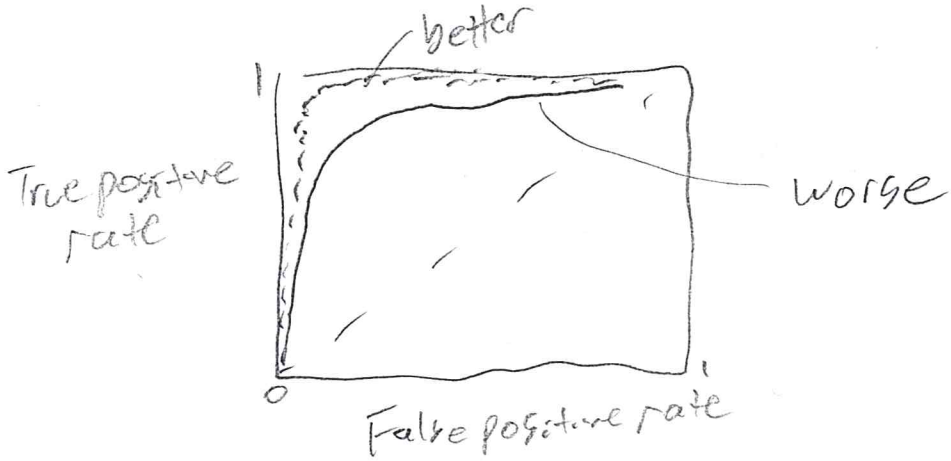
ROC Curve



$0 \leq AUC \leq 1$

A classifier with AUC=0.5 does no better than chance

27) p 354 varies γ for the radial kernel and selects $\hat{\gamma}$ with the best ROC curve



28) ROC from ~~test~~ data or validation data is better than ROC from training data.

29) The true positive rate is called the sensitivity and the false positive rate is 1 − specificity.

False positive rate = $\frac{FP}{N}$ ≃ type I error, 1 − specificity

True positive rate = $\frac{TP}{P}$ ≃ 1 − type II error, power, sensitivity, recall

Positive predicted value $\frac{TP}{P^*}$ ≃ precision, 1 − false discovery proportion

negative predicted value $\frac{TN}{N^*}$

error rate = (FP+FN)/(FP+FN+TN+TP)

30) Sometimes one error is much more important than the other. ex loan default do not default misclassifying default should be small compared to misclassifying do not default

§ 9.4 SVMs with $G > 2$ classes

R package uses this

31] The one versus one or all pairs classifier constructs $\binom{G}{2}$ binary classifiers, one for each distinct pair of groups.

Classify $\underline{x}$ with $f_{ij}(\underline{x})$ and let

m_i = # times $\underline{x}$ is predicted to be in class i.

Then $\hat{Y}(\underline{x}) = d$ where $m_d = \max(m_1, \ldots, m_G)$.

32] The one versus all classifier fits G binary SVMs: group $i = +1$ versus the $G-1$ other classes with -1 with $f_i(\underline{x})$.

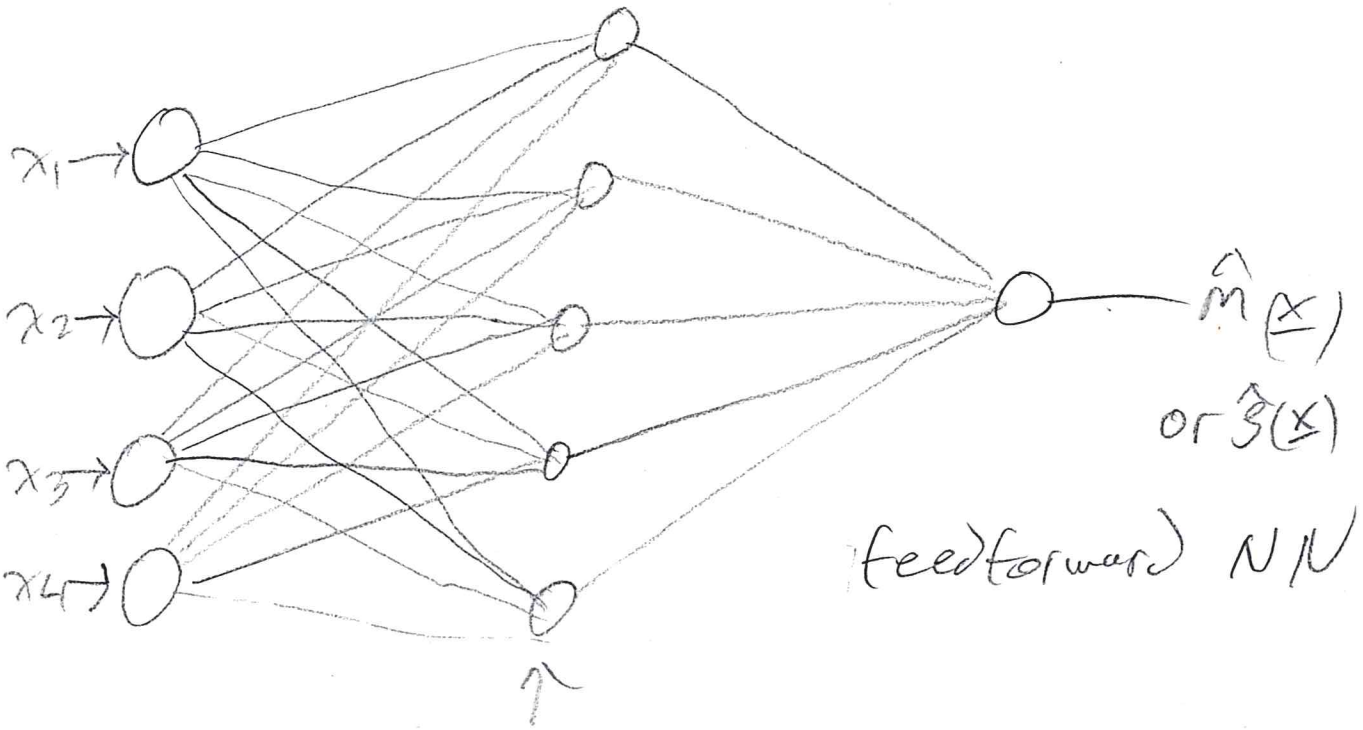
Then $\hat{Y}(\underline{x}) = d$ where $f_d(\underline{x}) = \max(f_1(\underline{x}), \ldots, f_G(\underline{x}))$.

33] Rules 31] and 32] can be applied to other binary classifiers such as logistic regression, Neural Networks not on final $Y|\underline{x} \sim \text{bin}(1, s(\underline{x}))$.

1] Suppose $Y = m(\underline{x}) + e$, or Neural networks

(NN) ... produce $\hat{m}(\underline{x})$ or $\hat{g}(\underline{x})$.

2) Input layer L_1 hidden layer L_2 output layer L_3



feedforward NN

The 5 hidden units $a_\ell(\underline{x}) =$

$$a_\ell = g\left(\beta_{\ell 0}^{(1)} + \underset{4\times 1}{\underline{x}^T \underline{\beta}_\ell^{(1)}}\right), \quad \ell = 1,\dots,5 \text{ and } a$$

single output unit $o = h\left(\eta_0^{(2)} + \underline{a}^T \underset{5\times 1}{\underline{\eta}}\right)$.

Often use a sigmoid function

$$g(t) = \frac{1}{1+e^{-t}} = \frac{1}{1+e^{-t}} \frac{e^t}{e^t} = \frac{e^t}{1+e^t}.$$

For binary regression often $h(t) = \frac{1}{1+e^{-t}}$ (sigmoid)

For $Y = m(\underline{x}) + \epsilon$, often $h(t) = t$ (identity)

So $\hat{m}(\underline{x}) = \alpha_0 + \underline{a}^T \underline{\alpha}$.

Often write $\beta_{10} = w_{10}^{(1)}$ ("biags"), $\underline{\beta_w} = \underline{w}^{(1)}$ for weights

$\alpha_0 = w_0^{(2)}$, $\underline{\alpha} = \underline{w}^{(2)}$ where

(1) = 1st layer and (2) = 2nd layer.

The NN attempts to learn $m(\underline{x})$ or $g(\underline{x})$.

3) A NN can have L_{m+2} layers

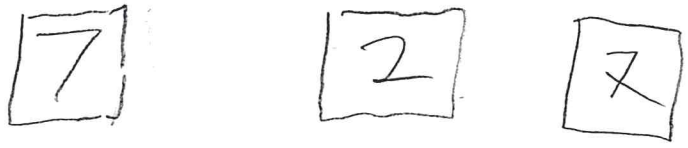
L_1 = input, L_{m+2} = output, and m hidden layers $L_2, \ldots, L_{m+1}$. Each hidden layer has K_i hidden units, $i = 2, \ldots, m+1$.

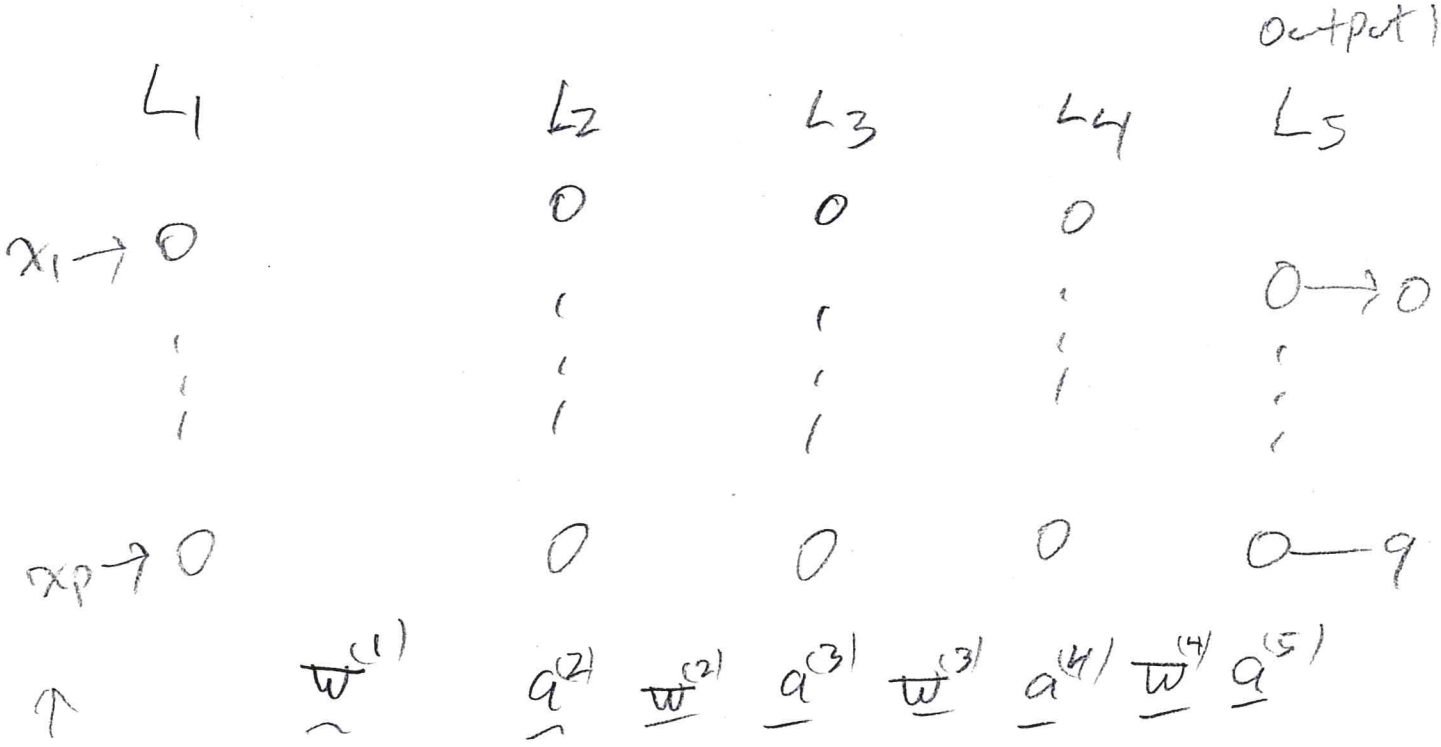
4) NNs were popular 1985 – 1995, then boosting and SVM's became popular. Around 2010 NNs made a comeback: deep learning.

5) NNs are good at optical character recognition - for example, automatic

reading of handwritten digits

Build a classifier $\hat{C}(\underline{x}) \in \{0, 1, \ldots, 9\}$ based on $\underline{x} \in \mathbb{R}^{28\times 28}$ where each digit is represented by a 28×28 greyscale image





≥ 784 digits training data

This NN has close to 4 million weights and needs to be heavily regularized.

5) Let $W^{(k+1)}$ (including bias = intercepts) be the matrix of weights that go from layer L_{k-1} to L_k. Let $\underline{a}^{(k)}$ be the

vector of activations at layer L_k,

Let $\underline{z}^{(k)} = W^{(k-1)}\,\underline{a}^{(k-1)}$ and $\underline{a}^{(k)} = g^{(k)}(\underline{z}^{(k)})$

where $g^{(k)}(\cdot)$ operates elementwise on $\underline{z}^{(k)}$ to produce vector $\underline{a}^{(k)}$ $\left(g^{(k)}: \mathbb{R}^{m_k} \to \mathbb{R}^{n_k}\right)$.

7) For G-class classification, the softmax function $g^{(G)}(z_m^{(G)}; \underline{z}^{(G)}) = \dfrac{e^{z_m^{(G)}}}{\sum_{i=1}^{G} e^{z_i^{(k)}}}$

for $m = 1, \ldots, G$. These are G probabilities that sum to one. Then

$\hat{C}(\underline{x}) = d$ where $d = \max_{m=1,\ldots,G} g^{(G)}(z_m^{(G)}, \underline{z}^{(G)})$.

8) Let w be the collection of weights.

not convex so not fast

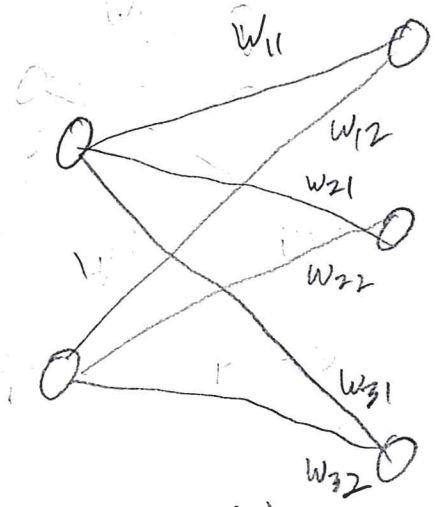
$$\min_{w} \frac{1}{n} \sum_{i=1}^{n} \left\{ L[y_i, \hat{M}(\underline{x}_i, w)] \right\} + \lambda J(w)$$

$$J(w) = \frac{1}{2} \sum_{i=1}^{k-1} \sum_{j=1}^{p_i} \sum_{\ell=1}^{p_{i+1}} \left(w_{\ell j}^{(i)}\right)^2$$ is often used for regression.

p_i = # units in ith layer

$k-1$ = # layers

↑ no intercepts.



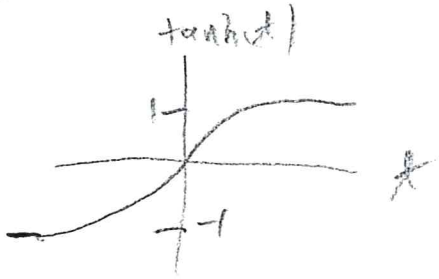
$W^{(i)}$ $\quad$ $\underset{\sim}{a}^{(i+1)}$

$P_i = 2$ $\quad$ $P_{i+1} = 3$

9) Deeper neural networks have more hidden layers. With one hidden layer, the number of hidden units determines the number of parameters. Two hidden layers works well for digit recognition.

10) In addition to the sigmoid

$$g(t) = \frac{e^t - e^{-t}}{e^t + e^{-t}} = \tanh(t) \quad \text{is used.}$$



The logistic curve (sigmoid) and $\tanh(t)$ look similar. $\tanh(t)$ ranges from -1 to 1, $\frac{\tanh(t)+1}{2}$ ranges from 0 to 1.

The rectified linear $g(t) = t_+ = \begin{cases} t & t > 0 \\ 0 & t \leq 0 \end{cases}$ is also used.

11) Deep learning ≈ neural networks with better computers. NN are used classifying natural images (data base of flowers, snakes, mammals, insects, people) using what is known as convolutional architecture

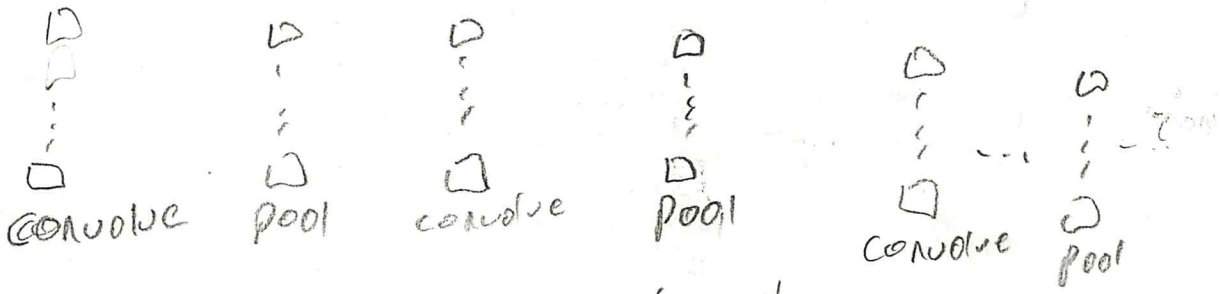
12) An autoencoder is a NN used to solve a nonlinear principal component decomposition. A single layer autoencoder solves

$$\min_{\substack{W \\ p \times p}} \sum_{i=1}^{n} \| x_i - W' g(W x_i) \|_2^2$$

13) PCA: $x_i \in \mathbb{R}^p$, $z_i = V^T x_i \in \mathbb{R}^q$, $q \leq p$ and columns of V are orthonormal. Taking

$g(t) = t$ = (identity function) gives PCA.

14) image



deep learning network

15) Gradient descent for each layer is used since the NN criterion is not convex