Chapter 1:

None

Chapter 2:

2.1 $\arg\max_k \hat{y} = \arg\min_k |t_k - \hat{y}|_2$

$$|t_k - \hat{y}|^2 = (1-\hat{y})^2 + \sum_{j \neq k} \hat{y}_j^2$$

$$= 1 - 2\hat{y} + \underbrace{|\hat{y}|^2}_{1 \text{ by assumption}}$$

$$\Rightarrow \arg\min_k 1 - 2\hat{y} = \arg\max \hat{y}$$

2.2 $b_{1 \ldots 10} \sim N\left(\begin{smallmatrix}1\\0\end{smallmatrix}, \mathbb{I}\right)$ blue

$o_{1 \ldots 10} \sim N\left(\begin{smallmatrix}0\\1\end{smallmatrix}, \mathbb{I}\right)$ orange

$N(b_i, \mathbb{I}/5)$ $N(o_i, \mathbb{I}/5)$

$$\Rightarrow p(x \mid \text{blue}) = \sum_{i=1}^{10} \frac{1}{10} N(b_i, \mathbb{I}/5)$$

$$\Rightarrow p(\text{blue} \mid x) = \frac{\left(\frac{1}{10} \sum_{i=1}^{10} N(b_i, \mathbb{I}/5)\right) \cdot \frac{1}{2}}{\frac{1}{10} \sum_{i=1}^{10} N(b_i, \mathbb{I}/5) + \frac{1}{10} \sum_{i=1}^{10} N(o_i, \mathbb{I}/5)}$$

$$p(\text{blue} \mid x) = p(\text{orange} \mid x)$$

$$\Rightarrow \sum N(b_i, \tfrac{\mathbb{I}}{5}) = \sum N(o_i, \tfrac{\mathbb{I}}{5})$$

$$\boxed{\sum \exp((x-b_i)^2 \cdot \tfrac{5}{2}) = \sum \exp((x-0_i)^2 \cdot \tfrac{5}{2})}$$

2.3 $\quad Vol(B_r^p) = v_p r^p$

$$Prob(x \sim B_1^p \notin B_r^p) = 1 - \frac{Vol\, B_r^p}{Vol\, B_1^p} = 1 - r^p$$

$$P(r,N) := Prob\left[(x_1 \cdots x_N \text{ iid} \sim B_1^p) \in B_r^p\right] = (1-r^p)^N$$

Median is when $P_N = 1/2$

$$\Rightarrow (1-r^p)^N = \tfrac{1}{2} \qquad \Rightarrow r = \left(1 - \tfrac{1}{2}^{1/N}\right)^{1/p}$$

by def of pdf

$$\int_{-\infty}^{med} = \int_{med}^{\infty}$$

2.4 $\quad$ WLOG $\quad a_i = \frac{1}{\sqrt{N}} \mathbb{1}$

$$\Rightarrow z = \tfrac{1}{\sqrt{N}} \sum x_i \sim N(0,1) \quad \text{if} \quad x_i \in N(0, \mathbb{I})$$

$\Rightarrow$ For $p=10$ avg distance^2 is

$$Mean(\Gamma(\tfrac{v}{2}, \tfrac{1}{2})_{v=p}) = p \Rightarrow \text{RMS dist} \sim \sqrt{p}$$

While along any axis its 1

2.5 $\quad$ Test point x_0, we know $Y = X^T\beta + \varepsilon$

$$\hat{y}_0 = x_0^T \hat{\beta}$$

$$\hat{y}_0 = x_0^T \beta + \sum_{i=1}^{N} \underbrace{l_i(x_0)}_{[x_0(X^TX)^{-1}X^T]_i} \varepsilon_i$$

a) $EPE(x_0) = \mathbb{E}_{y_0|x_0} \mathbb{E}_{\mathcal{T}} (y_0 - \hat{y}_0)$

$$= \underbrace{Var(y_0|x_0)}_{\text{Bayes}} + \underbrace{\left(x_0^T\beta - \mathbb{E}_{\mathcal{T}}\hat{y}_0\right)^2}_{\text{Bias}} + \underbrace{\mathbb{E}_{\mathcal{T}}\left(\hat{y}_0 - \mathbb{E}_{\mathcal{T}}\hat{y}_0\right)^2}_{\text{Var}}$$

$$= \sigma^2 + 0 + \mathbb{E}_{\mathcal{T}}\, x_0^T (X^TX)^{-1} x_0 \sigma^2$$

b) $N \to \infty \Rightarrow X^TX \to N\, Cov\, X$ (empirical cov concentrates)

$$\Rightarrow \mathbb{E}_{x_0} EPE(x_0) = \sigma^2\left(1 + \frac{1}{N}\mathbb{E}_{x_0}\left[x_0^T\, Cov(X)^{-1} x_0\right]\right)$$

$\mathbb{E}_{x_0}\, x_0 x_0^T = Cov\, x_0$

$$= \sigma^2\left(1 + \frac{1}{N} Tr\left[Cov(X)^{-1} Cov(x_0)\right]\right)$$

assuming x_0 is in-distribution $\rightarrow$

$$= \sigma^2\left(1 + \frac{1}{N} Tr[I_p]\right)$$

$$= \sigma^2\left(1 + \frac{p}{N}\right)$$

2.6 $RSS(\theta) = \sum_{i=1}^{N}\sum_{l=1}^{N_i}\left(y_{il} - f_\theta(x_i)\right)^2$

$y_{il},\ l \in 1, \dots, N_i$

$$= \sum_{i,l}\left(y_{il} - \bar{y}_i + \bar{y}_i - f_\theta(x_i)\right)^2$$

$$= \overset{\text{irre}}{\cancel{\sum_{i,l}(y_{il} - \bar{y}_i)^2}} + \sum_i N_i\left(\bar{y}_i - f_\theta(x_i)\right)^2$$

2.7 a) L.R. : $x_0(X^TX)^{-1}X^T$

kNN : $l_i(x_0;X) = \begin{cases} \frac{1}{k} & \text{if } x_i \in N_k(x_0) \\ 0 \end{cases}$

b) $$\mathbb{E}_{Y|X}\left|f(x_0) - \hat{f}(x_0)\right|^2 = \underbrace{\left(f(x_0) - \mathbb{E}_{Y|X}\hat{f}(x_0)\right)^2}_{\text{Bias}} + \underbrace{\mathbb{E}_{Y|X}\left(\hat{f}(x_0) - \mathbb{E}_{Y|X}\hat{f}(x_0)\right)^2}_{\text{Var}}$$

c) $\mathbb{E}_{Y,X} \left| f(x_0) - \hat{f}(x_0) \right|^2 = \underbrace{\left| f(x_0) - \mathbb{E}_{Y,X} \hat{f}(x_0) \right|^2}_{\text{Bias}}$

$+ \underbrace{\mathbb{E}_{Y,X} \left| \hat{f}(x_0) - \mathbb{E}_{Y,X} \hat{f}(x_0) \right|^2}_{\text{Var}}$

d) bias:

$$f(x_0) - \mathbb{E}_{Y|X} \hat{f}(x_0) = f(x_0) - \sum_i l_i(x_0; X) f(x_i)$$

Var:

$$\mathbb{E}_{Y|X} \left(\hat{f}(x_0) - \mathbb{E}_{Y|X} \hat{f}(x_0) \right)^2 = \mathbb{E}_{\varepsilon_i} \left(\sum_i l_i(x_0; X) \varepsilon_i \right)^2$$

$$= \sigma^2 \sum_i l_i(x_0; X)^2$$

$$S := \begin{pmatrix} l_1(x_0; X) \\ \vdots \\ l_N(x_0; X) \end{pmatrix}$$

$\Rightarrow$ $\text{Bias} = f(x_0) - S^T f \Rightarrow \text{Bias}^2 = f(x_0)^2 - 2 f(x_0) S^T f + f^T S S^T f$

$\text{Var} = \sigma^2 S^T S$

$$\Rightarrow \text{Bias}^2 = f(x_0)\left(f(x_0) - 2 S^T f\right) + \frac{f^T \,\text{Var}\, f}{\sigma^2}$$

for c)

bias:

$$f(x_0) - \underbrace{\int \prod dx_i\, h(x_i) \sum_i l_i(x_0; X) f(x_i)}_{\$(f; x_0)}$$

Var:

$$\sum_i \left[l_i(x_0; X) y_i - \int dx_i\, h(x_i)\, l_i(x_0; X_i)\, f(x_i) \right]$$

Do they just want

$$\text{Bias}^2 \hat{f}(x_0) + \text{Var}\, \hat{f}(x_0) = \mathbb{E}_X \left[\text{Bias}\left(\hat{f}(x_0) | X\right)^2 + \text{Var}\left(\hat{f}(x_0) | X\right) \right]$$

2.8

```
Test error rate of Linear Regression is 4.12%
Train error rate of Linear Regression is 0.58%
k-NN Model: k is 1, train/test error rates are 0.00% and 2.47%
k-NN Model: k is 2, train/test error rates are 0.58% and 2.47%
k-NN Model: k is 3, train/test error rates are 0.50% and 3.02%
k-NN Model: k is 4, train/test error rates are 0.43% and 2.75%
k-NN Model: k is 5, train/test error rates are 0.58% and 3.02%
k-NN Model: k is 6, train/test error rates are 0.50% and 3.02%
k-NN Model: k is 7, train/test error rates are 0.65% and 3.30%
k-NN Model: k is 8, train/test error rates are 0.58% and 3.30%
k-NN Model: k is 9, train/test error rates are 0.94% and 3.57%
k-NN Model: k is 10, train/test error rates are 0.79% and 3.57%
k-NN Model: k is 11, train/test error rates are 0.86% and 3.57%
k-NN Model: k is 12, train/test error rates are 0.72% and 3.57%
k-NN Model: k is 13, train/test error rates are 0.86% and 3.85%
k-NN Model: k is 14, train/test error rates are 0.86% and 3.85%
k-NN Model: k is 15, train/test error rates are 0.94% and 3.85%
```

2.9 $\mathbb{E}\, R_{tr}(\hat{\beta}) \le \mathbb{E}\, R_{te}(\hat{\beta})$

$$\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{\beta}\cdot x_i)^2\right] \le \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}(y_i-\beta x_i)^2\right]$$

$$= \mathbb{E}\, R_{tr}(\beta)$$

$$\mathbb{E}_{x,y,\tilde{x},\tilde{y}}\, R_{te}(\hat{\beta}) = \mathbb{E}_{x,y}\,\mathbb{E}_{\tilde{x},\tilde{y}}\,\frac{1}{M}\sum_{i=1}^{M}(\tilde{y}_i-\hat{\beta}\tilde{x}_i)^2$$

$$\ge \mathbb{E}_{\tilde{x},\tilde{y}}\,\frac{1}{M}\sum_{i=1}^{M}(\tilde{y}-\beta^*\tilde{x}_i)^2$$

β^* is optimal for test set MSE

$$= \mathbb{E}_{x,y}\,\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{\beta}x_i)^2$$

$$= \mathbb{E}_{x,y}\, R_{tr}(\hat{\beta})$$

Chapter 3

1. We will show that the F-statistic for adding/dropping a single term is the square of its z-score

$$\hat{\sigma}^2 = \frac{RSS_1}{N-p_1-1} \qquad (N-p-1)\hat{\sigma}^2 \sim \chi^2_{N-p-1}\cdot\sigma^2$$

$$\hat{\beta} \sim N(\beta, (X^TX)^{-1}\sigma^2) \qquad v_j = [(X^TX)^{-1}]_{jj}$$

$$z_j = \frac{\hat{\beta}_j}{\sigma\sqrt{v_j}} \sim N(0,1)$$

Show: F statistic is given by z^2

$$F := \frac{RSS_0 - RSS_1}{\frac{RSS_1}{N-p-1}}$$ bigger model

$$\min_{\beta} \ (y - X\beta)^2 + \lambda(\beta^T e_j - 0)$$

$$\Rightarrow \ \beta^0 = (X^TX)^{-1}(X^Ty - \lambda e_j)$$

$$\beta^0 \cdot e_j = 0 \Rightarrow \lambda = \frac{e_j^T (X^TX)^{-1} X^T y}{e_j^T (X^TX)^{-1} e_j}$$

$$\Rightarrow \ \beta^0 = \beta - \frac{e_j^T (X^TX)^{-1} X^T y}{e_j^T (X^TX)^{-1} e_j} (X^TX)^{-1} e_j$$

$$\Rightarrow \ RSS_0 = (y - X\beta^0)^2 = (y - X\beta' - \lambda X (X^TX)^{-1} e_j)^2$$

$$= RSS_1 - 2\lambda \underbrace{(y - X\beta^0)^T X}_{0} (X^TX)^{-1} e_j$$

$$+ \lambda^2 e_j (X^TX)^{-1} e_j$$

$$(y - X\beta^0)^T X = yX - yX(X^TX)^{-1} X^T X = 0$$

$$\Rightarrow \ RSS_0 - RSS_1 = \frac{\underbrace{(e_j (X^TX)^{-1} X^T)^2}_{(\hat{\beta}_j)}}{\underbrace{e_j (X^TX)^{-1} e_j}_{v_j}}$$

$$\Rightarrow \ \frac{RSS_0 - RSS_1}{\frac{RSS_1}{N-p_1-1}} = \frac{\hat{\beta}_j^2}{v_j} \frac{1}{\hat{\sigma}_j}$$

$$= (z_j)^2$$

2. Method 1 gives: $Var(y_0) = x_0^T \, Var\,\hat{\beta} \, x_0$

$$= x_0^T (X^TX)^{-1} x_0 \, \sigma^2$$

$$\Rightarrow y_0 = \hat{y}_0 \pm \sigma \sqrt{x_0 (X^TX)^{-1} x_0}$$

Method 2 gives: $\beta \sim N(\hat{\beta}, (X^TX)^{-1} \sigma^2)$

$$C_\beta = \{\beta \mid (\beta - \hat{\beta}) \frac{X^TX}{\sigma^2} (\beta - \hat{\beta}) \le \chi^2_{4, 0.05}\}$$

$X^T X = U^T U$ $\quad U$ is upper ∇

$\vec{v} := U(\beta - \hat{\beta})$ lies in ball of radius $r = \sigma\sqrt{\chi^2_{4,0.05}}$

$\underbrace{\beta = \hat{\beta} + \sigma\sqrt{\chi^2_{0.05}}\, U \cdot \xi}_{\text{elliptical}}$ $\quad \xi \in S^3$

$\Rightarrow$ Method 2 gives tighter bounds b.c. it bounds all β_i at the same time

3.3 a) Let $\theta = c^T y$ another estimate of $a^T\beta$

WLOG $\quad c = a(X^T X)^{-1} X^T + d$ $\leftarrow$ arbitrary, possibly X, y dep

$$\underset{y}{\mathbb{E}}[c^T y] = \underset{\varepsilon}{\mathbb{E}}\, (a(X^T X)^{-1} X^T + d)(X\beta + \varepsilon)$$

$$= a^T\beta + dX\beta$$

unbiased $\Leftrightarrow$ $dX\beta = 0 \Rightarrow dX = 0$ $\quad \forall$ vecs

$$Var(c^T y) = \sigma^2 (a(X^T X)^{-1} X^T + d)(a(X^T X)^{-1} X^T + d)^T$$

$$= \underbrace{\sigma^2 dd^T}_{\geq 0} + Var\, a^T\hat{\beta}$$

b) As before but now $d \to D$

$$\Rightarrow DX = 0$$

$$\Rightarrow Var\, C^T y = \sigma^2 DD^T + Var(\hat{\beta})$$

$$\Rightarrow Var\, C^T y \succeq Var\, \hat{\beta}$$

$$\Rightarrow \tilde{V} \succeq \hat{V}$$

3.4 $\quad X = QR \Rightarrow X^T X = R^T R$

Q is $N \times (p+1)$

$$(R^T R)^{-1} R Q^T y$$

$Q^T Q = \mathbb{I}_{p+1}$

$$= R^{-1} Q^T y$$

R is $(p+1) \times (p+1)$

calculate $q_k^T y$ to fill column vec of $Q^T y \in \mathbb{R}^{p\times 1}$

solve $R^{*} Q^T y$ by backsub

$\hat{y} = Q Q^T y$

3.5 In 3.41

$$\underset{\beta}{\text{argmin}} \sum_{i=1}^{N} (y_i - \beta_0 - X \cdot \beta)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

take $x \to x - \bar{x}_j$

$$\Rightarrow \sum_i (y_i - \beta_0 \underbrace{- \sum_j \bar{x}_j \beta_j} - \sum_j (x_{ij} - \bar{x}_j)\beta_j)^2 + \lambda |\beta|^2$$

$$\Rightarrow \beta_0^c = \beta_0 - \sum_j \bar{x}_j \beta_j$$
$$\beta_j^c = \beta_j$$

$$\frac{\partial L}{\partial \beta_0} = \sum_i y_i - N\beta_0^c - \underbrace{\sum_{ij} (x_{ij} - \bar{x}_j)\beta_j}_{0 \text{ as a sum}} = 0$$

$$\Rightarrow \beta_0^c = \bar{y}$$

$\tilde{y} = y_i - \beta_0^c$

$\tilde{x}_{ij} = x_{ij} - \bar{x}_j$

$$\Rightarrow \min_{\beta} (\tilde{X}\beta - \tilde{y})^2 + \lambda |\beta|^2$$

$$\Rightarrow \hat{\beta}_c = (X^T X + \lambda \mathbb{1})^{-1} \tilde{X}^T y$$

3.6 $P(\theta | X) \propto \exp\left[-\frac{1}{\sigma^2} \underbrace{\sum_{i=1}^{N} (x_i \cdot \beta - y_i)^2}_{\text{extensive in } N} - \frac{1}{\tau} \mathbb{1} \beta^2 \right]$

$$\Rightarrow \lambda = \frac{\sigma^2}{\tau} \qquad \Rightarrow \min_{\beta} (X\beta - y)^2 + \lambda \beta^2$$

$$\Rightarrow (X^T X + \lambda \mathbb{1})^{-1} X^T y$$

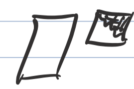
3.7 $\quad P(\beta|y) = \dfrac{P(\beta)\,P(y|\beta)}{P(y)}$

$$\Rightarrow -\log P(\beta|y) = \frac{1}{\tau^2}|\beta|^2 + \frac{1}{\sigma^2}\sum_{i=1}^{N}\Big(y_i - \beta_0 - \sum_j x_{ij}\beta_j\Big)^2 + \log Z$$

$$\lambda = \frac{\sigma^2}{\tau^2}$$

3.8 $\quad$ let $\vec{x}_j = \begin{pmatrix} x_{1j} \\ \vdots \\ x_{Nj} \end{pmatrix} \in \mathbb{R}^N$ and $\mathbb{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$ be the columns of X

$\vec{q}_j = \begin{pmatrix} q_{1j} \\ \vdots \\ q_{Nj} \end{pmatrix} \in \mathbb{R}^N$ the columns of Q

$X = Q\,R =$ 

$$\mathbb{1} = r_{00} \cdot \vec{q}_0$$

$$\Rightarrow r_{00} = \sqrt{N} \qquad \vec{q}_0 = \frac{1}{\sqrt{N}}$$

$$\Rightarrow \bar{q}_j = \sum_i q_{ij}/N = \frac{1}{\sqrt{N}} q_0^T q_j = 0$$

$$\Rightarrow \bar{x}_j = r_{0j}\, q_0 = \frac{r_{0j}}{\sqrt{N}}$$

$$\Rightarrow \tilde{x} = \vec{x}_j - \bar{x}_j \mathbb{1}$$

$$= \sum_{k\neq 0} q_k r_{kj}$$

take $\quad Q_2 = (q_1 \cdots q_p)$

$$\tilde{X} = (\tilde{x}_1 \cdots \tilde{x}_p) = U\Sigma V^T \in \mathbb{R}^{N\times p}$$

Q_2 spans the p-dim subspace spanned by the data as does $\mathrm{col}(\tilde{X})$

$$Q_2 R_2 = \tilde{X} = UDV^T$$

$$R_2 = Q_2^T U D V^T$$

$$Q_2 = U \Rightarrow \quad R_2 = \underbrace{D}_{\text{diag}_2}V^T$$

$\Rightarrow V$ is diag w elems ± 1 $\Rightarrow$ so is R

$\Rightarrow$ take all $+1$ WLOG

If $\tilde{X}$ has ortho columns

$\Rightarrow \tilde{X} = QR$ has R diag w strictly pos entries

$\Rightarrow$ this is SVD w/ $Q = U$

So $QR = SVD$ when X has orthogonal columns

3.9

Claim: $\underset{k}{\operatorname{argmax}} \; q_k^T r$

$X = Q_1 R_1$

add a predictor x_k

$\mathrm{Proj}_{X_1} x_k = \sum_i (x_k^T q_i) q_i$

$r_k = x_k - \mathrm{Proj}_{X_1} x_k$

$q_k = \frac{r_k}{|r_k|}$

$\hat{y}_1 \to \hat{y}_2 = \hat{y}_1 + (q_k^T y) q_k$

$= \hat{y}_1 + (q_k^T r) q_k$

$\Rightarrow RSS_2 = RSS_1 - (q_k^T r)^2$

$\Rightarrow$ pick max k for $(q_k^T r)^2$

3.10 Z score is $\frac{\beta_p}{\hat{\sigma}\sqrt{v}} = \frac{R_{p:}^{-1} Q^T y}{\hat{\sigma} \; R_{pp}^{-1}} = \frac{q_p^T y}{\sigma} \Leftarrow$ pick smallest z-score

for last predictor added

$R_{p:}^{-1} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ R_{pp}^{-1} \end{pmatrix}$

$\mathrm{Var}\, \beta_j = \frac{\sigma^2}{|z_p|^2} =$

3.11 $\mathrm{Tr}\, (Y - XB)^T \Sigma^{-1} (Y - XB)$

$\Sigma = I\sigma^2 \Rightarrow B = (X^T X)^{-1} X^T Y$

else $S := \sqrt{\Sigma}$ $\quad Y \to YS, \; B \to BS$

$$BS = (X^T X)^{-1} X^T Y S \Rightarrow B = (X^T X)^{-1} X^T Y$$

No closed formula for general i-dependent Σ
but still solvable by quadratic programming

3.12 $\quad \tilde{X} = \begin{pmatrix} X \\ \sqrt{\lambda}\, \mathbb{1}_{p\times p} \end{pmatrix} \Rightarrow \tilde{X}^T \tilde{X} = \begin{pmatrix} X^T & \sqrt{\lambda}\, \mathbb{1}_{p\times p} \end{pmatrix} \begin{pmatrix} X \\ \sqrt{\lambda}\, \mathbb{1}_{p\times p} \end{pmatrix}$

$Y = \begin{pmatrix} y \\ 0_p \end{pmatrix} \qquad\qquad = X^T X + \lambda \mathbb{1}$

$$\hat{\beta} = (\tilde{X}^T \tilde{X})^{-1} \tilde{X}^T y$$
$$= (\tilde{X}^T \tilde{X})^{-1} X^T y$$
$$= (X^T X + \lambda \mathbb{1})^{-1} X^T y$$

3.13 $\quad z_m = X v_m \Rightarrow$ regress y on $z_1 \dots z_m$

$z_1 \dots z_m$ orthogonal

$$\Rightarrow y^{pcr}_M = \bar{y}\mathbb{1} + \sum_{m=1}^{M} \hat{\theta}_m z_m$$

$$\hat{\theta}_m = \frac{\langle y, z_m \rangle}{\langle z_m, z_m \rangle}$$

$$\Rightarrow \beta^{pcr}(M) = \sum \hat{\theta}_m \vec{v}_m$$

$X = UDV^T \qquad z_m = X \vec{v}_m = d_m \vec{u}_m$

$$\hat{\theta}_m = \frac{\langle u_m, y \rangle}{d_m} \Rightarrow \beta^{pcr} = V \cdot D^{-1} \cdot U^T y = (X^T X)^{-1} X^T y$$

3.14 a) $z_m = \sum_j \hat{\varphi}_{mj} x_j^{(m-1)} \qquad \hat{\varphi}_{mj} = \langle y, x_j^{m-1} \rangle$

b) $\theta_m = \dfrac{\langle z_m, y \rangle}{\langle z_m, z_m \rangle}$

c) $\hat{y}^m = \hat{y}^{m-1} + z_m \theta_m$

d) $x_j^m = x_j^{m-1} - \dfrac{\langle z_m, x_j^{m-1} \rangle}{\langle z_m, z_m \rangle}$

$$\langle z_1, z_1\rangle = \sum_{i,j}^{p} \hat{\varphi}_{1i}\hat{\varphi}_{1j}\langle x_i^0, x_j^0\rangle$$

$$= \sum_{i=1}^{p} (\hat{\varphi}_{1i})^2$$

$$\langle z_1, y\rangle = \sum_j \hat{\varphi}_{1j} \underbrace{\langle y, x_j^0\rangle}_{= \varphi_{1j}} \Rightarrow \theta_1 = 1$$

$$= \sum_{i=1}^{p} (\hat{\varphi}_{1i})^2 \qquad \Rightarrow \hat{y}^1 = \hat{y}^0 + \sum_{i=1}^{p} \varphi_{1i} x_i^0$$

φ_{1j} by $x_i \perp x_j$

$$x_j^1 = x_j^0 - \frac{\langle z_1, x_j^0\rangle}{\langle z_1, z_1\rangle} z_1$$

$$= x_j^0 - \frac{\varphi_{1j}}{\sum \varphi_{1i}^2} \sum \varphi_{1j} x_i^0$$

$$\varphi_{2j} := \langle y, x_j^1\rangle = \langle x_j^0, y\rangle - \frac{\varphi_{1j}}{\cancel{\sum_i \varphi_{1i}^2}} \sum_j \varphi_{1j} \underbrace{\cancel{\langle x_i^0, y\rangle}}_{\varphi_{1i}}$$

$$= \langle x_j^0, y\rangle - \varphi_{1j} = 0$$

3.15 $\quad Corr^2(y, X\alpha)\, Var(X\alpha) = \dfrac{Cov^2(y, X\alpha)}{Var(y)}$

mth PLS $\hat{\varphi}_m$ yields

$$\Rightarrow \max_\alpha (y^T X \alpha)^2 \quad s.t. \quad |\alpha| = 1, \quad \alpha^T S \hat{\varphi}_\ell = 0$$

$$\alpha^T X^T y y^T X\alpha + \lambda(\alpha^T\alpha - 1) \qquad S = X^T X$$

$$\Rightarrow (X^T y y^T X)\alpha = \lambda\alpha$$

$$\Rightarrow \alpha = X^T y$$

$$\ell = 1 \Rightarrow \alpha_1 = \frac{X^T y}{|X^T y|_2} \propto \hat{\varphi}_1$$

Not done

$$\ell = 2 \Rightarrow \alpha_2 \propto X^T y - \frac{y^T X S X^T y}{y^T X S^2 X^T y} S X^T y$$

$$\Rightarrow \alpha_2^T S \alpha_1 = 0 \checkmark$$

Continuum regression:

$$\max_{\alpha} (y^T X\alpha)^2 (\alpha X^T X\alpha)^{\frac{r}{1-r}-1}$$

s.t. $|\alpha| = 1$ $\quad \alpha^T S \hat{\alpha}_\ell$ $\quad \ell = 1, \dots, m-1$

3.16 $\quad X^T X = \mathbb{1} \Rightarrow \hat{\beta}_j = \vec{x}_j^T y$

$$\Rightarrow \hat{\beta}_j \to \frac{\hat{\beta}_j}{1+\lambda} \quad \text{for ridge}$$

$$\hat{\beta}_j \to \hat{\beta}_j \cdot \mathbb{1}(\operatorname{rank} \beta_j \le M) \quad \text{for Subsets}$$

β_j

$$\beta^* = \underset{\beta}{\operatorname{argmin}} \tfrac{1}{2}(\hat{\beta}-\beta)^2 + \lambda|\hat{\beta}|$$

$$F = \mathcal{L}'(\beta) = \begin{matrix} (\beta-\hat{\beta}) - \lambda & \beta < 0 \\ (\beta-\hat{\beta}) + \lambda & \beta > 0 \end{matrix}$$

$$\Rightarrow \beta = \operatorname{sgn}\hat{\beta}\,(\hat{\beta} - |\lambda|)_+$$

3.17 Colab

3.18

3.19 $\quad X = UDV \Rightarrow |\beta^{ridge}|^2 = \sum_{j=1}^{p} \frac{d_j^2 (U^T y)_j}{(d_j^2+\lambda)^2}$

For LASSO use dual form

IDK for PLS

3.20

Motivation for CCA:

if $Y_k = f(X) + \varepsilon_k$
$Y_\ell = f(X) + \varepsilon_\ell$ } Pool these

Goal: Successively maximize $\mathrm{Corr}^2(Yu_m, Xv_m)$

ie Yu_1 most correlated to Xv_1

X, Y centered, look @ $\frac{Y^TX}{N}$

$$\mathrm{Corr}^2(Yu_m, Xv_m) = \frac{(u_m^T Y^T X v_m)^2}{\mathrm{Var}\, Xv_m \quad \mathrm{Var}\, Yu_m} = \frac{(u_m^T Y^T X v_m)^2}{v_m^T X^T X v_m \quad u_m^T Y^T Y u_m}$$

$$\mathcal{L} = u Y^T X v - \frac{\lambda_1}{2}(v^T X^T X v - 1) - \frac{\lambda_2}{2}(u Y^T Y u - 1)$$

$\Rightarrow Y^T X v = \lambda_1 Y^T Y u$
$X^T Y u = \lambda_2 X^T X v$ } $\Rightarrow \lambda_1 \cancel{u^T Y^T Y u}^{1} = \lambda_2 \cancel{v^T X^T X v}^{1}$
$\Rightarrow u Y^T X = \lambda_2 v (X^T X)$ $\qquad \lambda_1 = \lambda_2$

take $M = (Y^TY)^{-1/2}(Y^TX)(X^TX)^{-1/2} = \mathrm{Corr}(Y, X)$

$$M \cdot \underbrace{(X^TX)^{1/2} v}_{v_1^*} = \lambda \underbrace{(Y^TY)^{1/2} u}_{u_1^*}$$

$u (Y^TY)^{1/2} M = \lambda (X^TX)^{1/2} v \quad \Rightarrow \quad \boxed{\begin{array}{l} v = (X^TX)^{-1/2} v_* \\ u = (Y^TY)^{-1/2} u_* \end{array}}$

$M = U^* D^* V^{*T}$

u_* v_* top sing vals of covariance matrix

For $k = 2, \ldots, \min(K, p)$

$$\max_{\substack{u^T Y^T Y u = 1 \\ v^T X^T X v = 1 \\ u_i^T u_j = 0 \\ v_i^T v_j = 0}} u^T Y^T X v \Rightarrow \mathcal{L} = u^T Y^T X v - \frac{\lambda_1}{2}(v^T X^T X v - 1) - \frac{\lambda_2}{2}(u Y^T Y u - 1)$$

$$- \sum_{j<k} \alpha_j u^T u_j - \sum_{j<k} \beta_j v^T v_j$$

$$\Rightarrow \quad Y^T X v - \lambda_1 Y^T Y u - \sum_j \alpha_j u_j$$

$$X^T Y u - \lambda_2 X^T X v - \sum_j \beta_j v_j$$

$$\Rightarrow \quad u^T Y^T X v - \lambda_1 u^T Y^T Y u = 0$$
$$u^T Y^T X v - \lambda_1 v^T X^T X v = 0$$

$\Rightarrow$ same eqs

$\Rightarrow$ subsequent sing vals of $\mathrm{Corr}(Y, X)$

3.21 $$B^{rr} = \underset{\text{rank } B = m}{\operatorname{argmin}} \, \mathrm{Tr}\left[(Y - B^T X)\left(\frac{Y^T Y}{N}\right)^{-1}(Y - BX)^T\right]$$

$$Y \to Y^* = Y \Sigma^{-1/2} \Rightarrow \mathrm{Tr}\left[(Y^* - B^T X \Sigma^{-1/2})(Y^* - B^T X \Sigma^{-1/2})^T\right]$$

$$\Rightarrow \mathrm{Tr}\Big[\overset{\text{indep}}{\cancel{Y^* Y^{*T}}} + \underbrace{\Sigma^{-1/2} B^T X^T X B \Sigma^{-1/2} - 2 \Sigma^{-1/2} Y^T X B \Sigma^{-1/2}}_{\text{complete square}}\Big]$$

$$\Downarrow$$

$$\min_B \| \Sigma^{-1/2} B^T (X^T X)^{1/2} - \Sigma^{-1/2} Y^T X (X^T X)^{-1/2} \|$$

Eckhardt - Young - Mirsky

$$\min_{\mathrm{rk}\, \hat{D} = r} \| \hat{D} - D \|_F = \sum_{i=1}^{r} \sigma_i u_i v_i^T$$

in SVD: $D = U \Sigma V^T$

let $\hat{D} = \Sigma^{-1/2} B^T (X^T X)^{1/2}$ $\quad UDV^T =: \underbrace{\Sigma^{-1/2} Y^T X (X^T X)^{-1/2}}_{K \times M}$

$$\Rightarrow \quad \Sigma^{-1/2} \hat{B}^T (X^T X)^{1/2} = \sum_{i=1}^{m} d_i u_i v_i^T = U_m D V^T$$
$$= U_m U_m^T U D V$$

$$\Rightarrow \quad \hat{B}_m = (X^T X)^{-1/2} \left(\sum_{i=1}^{m} d_i v_i u_i^T\right) \Sigma^{1/2}$$

$$(X^T X)^{-1} (X^T Y) \underbrace{\Sigma^{-1/2} U_m}_{\tilde{U}_m} \underbrace{U_m^T \Sigma^{1/2}}_{\tilde{U}_m^-}$$

Equiv: $Y\Sigma^{-1/2} \Rightarrow Y^*$

$$\Rightarrow \min_{B_m} \left| B^T (X^TX)^{1/2} - \underbrace{Y^{*T} X (X^TX)^{-1/2}}_{\underbrace{U}_{K\cdot m}\Sigma V^T} \right|^2 = \underset{\text{Full rk}}{B}\ \tilde{U}_m \tilde{U}_m^-$$

$$B^T (X^TX)^{1/2} = U_m U_m^T \left(Y^{*T} X (X^TX)^{-1/2} \right)$$

$$\Rightarrow B = \underbrace{(X^TX)^{-1} X^T Y^*}_{\text{Full rank}} \underbrace{U_m U_m^T}_{\text{Proj}} \qquad U_m = \Sigma^{-1/2} U_m^* \qquad U_m^- = U_m^{*T} \Sigma^{1/2}$$

3.22 Prev was true $\forall \Sigma$

3.23 $\frac{1}{N} |\langle x_j, y\rangle| = \lambda \quad j = 1, \dots, p$

a) take $\hat\beta = (X^TX)^{-1} X^T y = (X^TX)^{-1} \lambda \frac{1}{N}$

$u = \alpha X \hat\beta \Rightarrow u$ moves a fraction α of the LS fit

$$\Rightarrow \frac{1}{N} \left| \langle X, y - \alpha X (X^TX)^{-1} X^T y \rangle \right|$$

$$= \left| \lambda - \alpha \langle X^TX\, X^TX^{-1} \rangle \frac{\langle X, y\rangle}{N} \right|$$

$$= \lambda (1-\alpha)$$

b) $$\text{Corr} = \frac{\frac{1}{N} \langle x_j, y - u_\alpha \rangle}{\sqrt{\frac{\langle x_j, x_j\rangle}{N}} \sqrt{\frac{\langle y-u_\alpha, y-u_\alpha\rangle}{N}}}$$

$$u_\alpha = \alpha X (X^TX)^{-1} X^T y$$

$$y - u_\alpha = (1-\alpha) y + \alpha (y - \hat y) \qquad (1-\alpha)^2 \langle y, y\rangle + \alpha^2 RSS$$

$$\Rightarrow \langle y - u_\alpha, y - u_\alpha \rangle = y^Ty + \alpha^2 y^T X (X^TX)^{-1} X^T y - 2\alpha y^T X (X^TX)^{-1} X^T y$$

$$\langle y - \alpha\hat y, y - \alpha \hat y\rangle$$

$$= y^T y + 2\alpha(\alpha - 2) \quad y^T \hat y$$

Needed to

$$= N + \alpha(2-\alpha)\,RSS + \alpha(\alpha-2)\,N$$

use that resp. y has $\frac{y^T y}{N} = 1$

$$= (\alpha-1)^2 N + \frac{\alpha(2-\alpha)}{2} RSS$$

$$\Rightarrow \text{Corr} = \frac{\lambda(1-\alpha)}{\sqrt{(1-\alpha)^2 + \frac{\alpha(2-\alpha)}{N} RSS}}$$

$\alpha = 0 \Rightarrow \text{Corr} = \lambda$
$\alpha = 1 \Rightarrow \text{Corr} = 0$
} always tied & decreasing

3.24 $\quad \delta_k = (X_{A_k}^T X_{A_k})^{-1} X_{A_k}^T r_k \in \mathbb{R}^{p'}$ $p' = \#$ active

$N \times p'$; residual before k is added

$$\beta_{A_k}(\alpha) = \beta_{A_k}(0) + \alpha\,\delta_k \Rightarrow \hat{f}_{A_k}(\alpha) = \hat{f}_{A_k}(0) + \alpha \underbrace{X_{A_k}\delta_k}_{u_k}$$

$\Rightarrow$ direction is $u_k = X_{A_k}\delta_k \in \mathbb{R}^N$

Claim is u_k makes smallest & equal angle w/ each col in X_{A_k}

$$\Rightarrow X_{A_k}^T u_k = \underbrace{X_{A_k}^T r_k}$$

$\vec{x}_i \cdot r_k = \vec{x}_j \cdot r_k$ by assumption of when we added x_k

also all $r_{k'} \notin A_k$ have $>$ corr by assumption

3.25 $\quad \hat{f}_k(\alpha) = \hat{f}_k(0) + \alpha\,u_k \qquad u_k = X_{A_k}\delta_k$

$$|c_a(\alpha)| = |x_a^T(y - \hat{f}_k(\alpha))|$$

$a \in A_k$
$$= x_a^T(r_k - \alpha u_k)$$
$$= x_a^T r_k - \alpha\, x_a^T u_k$$

$\leftarrow$ by 3.24 this is the same $\forall a \in A_k$

by $\nearrow$ 3.23 this is the same $\forall j \in A_k$

$= \hat{C} - \alpha A$

For $b \notin A_k$
$|x_b^T r_k| \leq |x_j^T r_k|$
$\Rightarrow$ for small α
$|c_b(\alpha)| < |c_j(\alpha)|$
until some α^*

Pick $b = \arg\max_b |c_b(\alpha)|$. Def $|c(\alpha)| = \max_{b \notin A_k} |c_b(\alpha)|$

$$|c(\alpha^*)| = |x_b^T r_k - \alpha^* x_b^T u_k| = |x_a^T r_k - \alpha^* x_a^T u_k| = |c_a(\alpha^*)|$$

$$\Rightarrow \frac{(x_b \mp x_a)^T r_k}{(x_b \mp x_a)^T u_k} = \frac{x_b r_k \mp C}{x_b u_k \mp A} = \alpha^*$$

$$\Rightarrow \alpha^* = \min_{b \notin A_k} \left\{ \frac{C + x_b^T r_k}{A + x_b^T u_k}, \frac{C - x_b^T r_k}{A - x_b^T u_k} \right\}$$

$b = \arg\min$ $\nearrow$

3.26 From 3.9 For fwd stepwise, at each step we choose k with

$$r_k = x_k - \mathrm{Proj}(x_k), \quad q_k = \frac{r_k}{|r_k|}$$

$$\hat{y} \to \hat{y} + (q_k^T y) q_k = \hat{y} + (q_k^T \underset{y - \hat{y}}{r}) q_k$$

$$y - \hat{y} = r \quad \Rightarrow r \to (\mathbb{1} - q_k q_k^T) r$$

$$RSS \to (q_k^T r)^2 \text{ reduction}$$

$\Rightarrow j$ for which $\mathrm{Corr}(x_{j,\mathrm{vA}}, r)$ is largest in magnitude

3.27 a) $\mathcal{L} = \mathcal{L}(\beta) + \lambda \sum_j |\beta_j|$

$$\beta_j = \beta_j^+ - \beta_j^- \qquad \beta_j^+, \beta_j^- \geq 0$$

$$\Rightarrow L = \mathcal{L}(\beta) + \lambda \sum_j (\beta_j^+ + \beta_j^-) - \sum_j (\lambda_j^+ \beta_j^+ + \lambda_j^- \beta_j^-)$$

$$\nabla \mathcal{L}_j(\beta) + \lambda - \lambda^+ = 0$$
$$-\nabla \mathcal{L}_j(\beta) + \lambda - \lambda^- = 0$$

b) $\lambda^+ + \lambda^- = 2\lambda \geq 0$

$\nabla \mathcal{L} = \frac{1}{2}(\lambda^- - \lambda^+)$

$\Rightarrow |\nabla \mathcal{L}| = \frac{1}{2}|\lambda^- - \lambda^+| \leq \frac{1}{2}(\lambda^- + \lambda^+) = \lambda$

$\Rightarrow \lambda = 0 \Rightarrow \nabla \mathcal{L}_j = 0$

KKT

$\beta_j^+ > 0,\ \lambda > 0 \Rightarrow \lambda^+ = 0 \quad \nabla \mathcal{L} = -\lambda < 0 \quad \beta_j^- = 0$

$\beta_j^- > 0,\ \lambda > 0 \Rightarrow \lambda^- = 0 \quad \nabla \mathcal{L} = \lambda > 0 \quad \beta_j^+ = 0$

$\Rightarrow \nabla \mathcal{L} = -\lambda \operatorname{sign}(\beta_j)$

$\Rightarrow \underbrace{\lambda = x_j^T (y - X\beta)}_{\text{all active preds } (\beta_j \neq 0) \text{ have same corr} = \lambda} = x_j^T r$

c) $X^T(y - X\hat{\beta}(\lambda)) = \theta(\lambda)$

$$\theta(\lambda)_j = \begin{cases} -\lambda \operatorname{sgn}(\beta_j) & \beta_j \neq 0 \quad \text{i.e. } j \in S \\ x_j^T y & \beta_j = 0 \quad \text{i.e. } j \notin S \end{cases}$$

$\hat{\beta}_j(\lambda) = (X^T X)^{-1}[X^T y - \theta(\lambda)]$

$$\underbrace{\hat{\beta}(\lambda) - \hat{\beta}(\lambda_0) = (X^T X)^{-1}(\underbrace{\theta(\lambda_0) - \theta(\lambda)}_{\begin{cases} (\lambda - \lambda_0) \operatorname{sgn} \beta_j(\lambda_0) & j \in S \\ 0 & j \notin S \end{cases}})}_{\text{linear in } \lambda}$$

3.28 $\hat{\beta}^{new} = \underset{\beta}{\operatorname{argmin}} \|y - X\beta - x_j^T \beta_j^*\|$

s.t. $|\beta|_1 + |\beta_j^*| \leq t$

$\tilde{\beta}_j = \beta_j + \beta_j^*$, $\tilde{\beta}$ has $\beta_j \to \tilde{\beta}_j$

$$\Rightarrow \quad \underset{\beta}{\operatorname{argmin}} \ \|y - X\tilde{\beta}\|$$

$$\text{s.t.} \quad |\tilde{\beta}|_1 + \underbrace{(|\beta_j| + |\beta_j^*| - |\tilde{\beta}_j|)}_{\geq 0} \leq t$$

$\Rightarrow$ More stringent

By symmetry, set $\beta_j = \beta_j^* = \frac{\beta^{orig}}{2} \Rightarrow |\tilde{\beta}|_1 = |\beta^{orig}|_1$

& objective is the same

$\Rightarrow$ sol'n

$\Rightarrow$ Final effect reduces β_j, β_j^* by $1/2$

3.29 $\quad \beta = \frac{X^T y}{X^T X + \lambda} \qquad X \in \mathbb{R}^{N,1}$

$$\tilde{X} = (X, X) \in \mathbb{R}^{N,2} \Rightarrow \beta = (\tilde{X}^T \tilde{X} + \lambda \mathbb{1})^{-1} \tilde{X}^T y$$

$$\tilde{\lambda} = \lambda \mathbb{1}_{2\times 2} \Rightarrow \beta_1 = \beta_2$$

$$\begin{pmatrix} X^T \\ X^T \end{pmatrix} (X \ X) = \begin{pmatrix} X^T X & X^T X \\ X^T X & X^T X \end{pmatrix} = \tilde{X}^T \tilde{X}$$

$$\|y - 2X\beta\|^2 + 2\lambda\|\beta\|^2 \Rightarrow \beta = \frac{2X^T y}{4X^T X + 2\lambda}$$

$$= \frac{X^T y}{2X^T X + \lambda}$$

for m copies

$$\frac{X^T y}{m X^T X + \lambda}$$

3.30 $\tilde{X} = \begin{pmatrix} X \\ \gamma \mathbf{1}_{p+1} \end{pmatrix} \in \mathbb{R}^{N+p+1,\, p+1}$ $\tilde{y} = \begin{pmatrix} y \\ 0 \end{pmatrix} \in \mathbb{R}^{N+p+1,\, 1}$

$\gamma = \sqrt{\lambda\alpha}$

$\Rightarrow$ Lasso for

$\|\tilde{y} - X\beta\|^2 + \lambda(1-\alpha)\|\beta\|_1$

Chapter 4

4.1 $\max\limits_{a^T W a = 1} \; a^T B a$

$\mathcal{L} = a^T B a - \lambda a^T W a$

$\Rightarrow \; Ba = \lambda W a \;\Rightarrow\; W^{-1} B a = \lambda a$

4.2 a) likelihood of data from 2 Gaussians w $\gamma_1 = \frac{N_1}{N}$
$\gamma_2 = \frac{N_2}{N}$

$(x-\mu_1)^T \Sigma (x-\mu_1) - (x-\mu_2)^T \Sigma (x-\mu_2) + \log \frac{N_1}{N_2}$

$x^T \Sigma (\mu_2 - \mu_1) - (\mu_2 + \mu_1) \Sigma^{-1} (\mu_2 - \mu_1) + \log \frac{N_1}{N_2}$

$x^T \Sigma^{-1} (\mu_2 - \mu_1) > (\mu_2 + \mu_1) \Sigma^{-1} (\mu_2 - \mu_1) - \log \frac{N_2}{N_1}$

b) $X = \begin{pmatrix} 1 & \bar{x}_i \end{pmatrix} \Rightarrow X^T X = \begin{pmatrix} N & N\bar{x} \\ N\bar{x} & \sum_i x_i x_i^T \end{pmatrix}$

Review

$X^T X \begin{pmatrix} \beta_0 \\ \beta \end{pmatrix} = X^T Y$ $\bar{x} \in \mathbb{R}^p$
$x_i \in \mathbb{R}^p$

$\Rightarrow N\beta_0 + N\bar{x}\cdot\beta = N\bar{y} \Rightarrow \beta_0 = \bar{y} - \bar{x}\cdot\beta$

$N\beta_0 \bar{x} + \sum_i x_i x_i^T \cdot \beta = \sum y_i x_i$

$* \left(\frac{1}{N}\sum x_i x_i^T - \bar{x}\bar{x}^T\right)\cdot\beta = \frac{1}{N}\sum y_i x_i - \bar{y}\bar{x}$

Take $\quad \mu_1 = \frac{1}{N_1} \sum_{G_1} x_i \quad \mu_2 = \frac{1}{N_2} \sum_{G_2} x_i$

$$y \in \left\{ \underset{G_1}{\frac{-N}{N_1}}, \underset{G_2}{\frac{N}{N_2}} \right\} \qquad \Rightarrow \bar{y} = 0 \qquad \bar{x} = \mu_1 + \mu_2$$

$$\sum x_i y_i = -N\mu_1 + N\mu_2$$

$$\Rightarrow \beta_0 = \cancel{(-N+N)}^{0} - \left(\frac{N_1\mu_1 + N_2\mu_2}{2}\right) \cdot \beta$$

$$* \Rightarrow \left(\sum x_i x_i^T - N\bar{x}\bar{x}^T\right) \cdot \beta = N(\mu_2 - \mu_1)$$

$\hat{\Sigma}$ is estimate of $Var(X|k)$

$$(N-2)\hat{\Sigma} = \left(\sum_{G_1} (x_i - \mu_1)(x_i - \mu_1) + \sum_{G_2} (x_i - \mu_2)(x_i - \mu_2)\right)$$

$$= \sum x_i x_i^T - N_1 \mu_1 \mu_1^T - N_2 \mu_2 \mu_2^T$$

$$\Rightarrow \sum x_i x_i^T - N\bar{x}\bar{x}^T = (N-2)\hat{\Sigma} + N_1\mu_1\mu_1^T + N_2\mu_2\mu_2^T - \frac{1}{N}(N_1\mu_1 + N_2\mu_2)(N_1\mu_1 + N_2\mu_2)^T$$

$$= (N-2)\hat{\Sigma} + \underbrace{\frac{N_1 N_2}{N}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T}_{N\hat{\Sigma}_B}$$

$N\hat{\Sigma}_B$, $\hat{\Sigma}_B = \frac{N_1}{N}\frac{N_2}{N}(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$

$*$ is as desired

c) $\quad \hat{\Sigma}_B \beta \propto \mu_2 - \mu_1 \quad$ by its structure

$$\Rightarrow \quad \hat{\Sigma}\beta \propto \mu_2 - \mu_1$$

$$\Rightarrow \beta \propto \underbrace{\hat{\Sigma}^{-1}(\mu_2 - \mu_1)}_{\text{LDA coeff}}$$

d) Replacing $\quad y \in \{t_1, t_2\}$

$$\beta_0 = \frac{1}{N}(N_1 t_1 + N_2 t_2) - \mu \cdot \beta \qquad \mu = \mu_1 + \mu_2$$

$$\Delta = \mu_1 - \mu_2$$

$$\left[(N-2)\hat{\Sigma} + N\hat{\Sigma}_B\right] \cdot \beta = (N_1 t_1 \mu_1 + N_2 t_2 \mu_2) - \frac{1}{N}(N_1 t_1 + N_2 t_2)(N_1\mu_1 + N_2\mu_2)$$

$$= \frac{N_1 N_2}{N}(t_1 - t_2)(\mu_1 - \mu_2)$$

e) $\beta_0 = -\mu \cdot \beta$

$\Rightarrow \hat{f}(x) = (x-\mu)\cdot\beta$

$(x-\mu)\Sigma^{-1}(\mu_2-\mu_1) = 0$ is decision bdy

$\cong$ to LDA when $N_1 = N_2$

else, we have extra $\log\frac{N_1}{N_2}$ const in LDA

4.3 $\pi^{new} = \pi^{old}$

$\mu^{new} = B^T\mu^{old}$

$\hat{\Sigma}^{new} = B^T\hat{\Sigma}^{old}B \Rightarrow (\hat{\Sigma}^{new})^{-1} = B^{-1}(\hat{\Sigma}^{old})^{-1}(B^T)^{-1}$

$$(x-\mu^{new})^T(\Sigma^{new})^{-1}(\mu_2^{new}-\mu_1^{new})$$
$$= (x-\mu^{old})^T B B^{-1}(\Sigma^{-1})^{old}(B^T)^{-1}B^T(\mu_2^{old}-\mu_1^{old})$$
$$= (x-\mu^{old})(\Sigma^{-1})^{old}(\mu_2^{old}-\mu_1^{old})$$

4.4 $\beta \in \mathbb{R}^{(p+1),(K-1)}$ $\beta \in \binom{\beta_0}{\vec{\beta}},\ x^T \in \binom{1}{x^T}$

$$P(G=k \mid X=x) = \frac{\exp(\beta_k^T x)}{1+\exp(\beta_k^T x)} \qquad k = 1,\dots,K-1$$

$$P(G=K \mid X=x) = \frac{1}{1+\exp(\beta_k^T x)}$$

$$l(\beta) = \sum_{i=1}^N \log P(g_i \mid x_i;\beta)$$

$$= \sum_{i=1}^N \sum_{k=1}^{K-1} \mathbb{1}(y_i=k)\beta_k^T x_{ik} - \log\left(1+\sum_i \exp\beta_k^T x_i\right)$$

$$\frac{\partial l}{\partial\vec{\beta}} = \sum_i \left[\mathbb{1}(y_i=k) - \frac{e^{\beta_k x_i}}{1+\Sigma\exp(\beta_k x_i)}\right]\vec{x_i}$$

$$\frac{\partial^2 \ell}{\partial \beta_k \partial \beta_\ell} = \sum_{i=1}^{N} p_k(x_i;\beta)\, p_\ell(x_i;\beta) \quad x_{ik} \; x_{i\ell} \qquad\qquad k \neq \ell$$

$$= -\sum_{i} p_k \,(1-p_k) \quad x_{ik} \; x_{ik} \qquad\qquad k=\ell$$

$$= -\sum_{i} \left[\mathrm{diag}(p_k(x_i;\beta)) - p_k(x_i;\beta)(1-p_\ell(x_i;\beta)) \right] \vec{x}_i \vec{x}_i^{\,T}$$

$$\beta^{new} = \beta^{old} - H^{-1} g$$

$$g = X^T (y_k - p_k) \quad \in \mathbb{R}^{p+1,\, k-1}$$

$$= \begin{pmatrix} X^T & & \\ & X^T & \\ & & \ddots \end{pmatrix} \begin{pmatrix} y_1 - p_1 \\ \vdots \\ y_{k-1} - p_{k-1} \end{pmatrix}$$

$$H = -X^T W X$$

$$W = \begin{pmatrix} P_1 & R_1 R_2 \cdots R_1 R_{k-1} \\ R_2 R_1 & \ddots & \vdots \\ \vdots & & P_{k-1} \\ R_{k-1} R_1 & \cdots & P_{k-1} \end{pmatrix} \qquad \begin{aligned} [R_k]_{ij} &= [\mathrm{diag}^{N \times N}\; p_k(x_i;\beta)]_{ij} \\ [P_k]_{ij} &= [\mathrm{diag}^{N \times N}\; p_k(x_i;\beta)(1-p_k(x_i;\beta))] \end{aligned}$$

$$\beta^{new} = \beta^{old} + (X^T W X)^{-1} X^T (y-p)$$

$$= (X^T W X)^{-1} (X^T W) [\underbrace{X\beta^{old} + W^{-1}(y-p)}_{z}]$$

z is our "target"

4.5 log likelihood:

$$\ell(\beta) = \sum_i \left[\beta_i x_i y_i - \log\left(1+e^{\beta_i x_i}\right)\right]$$

$$= \sum_i \left[(\beta_0 + \beta_1 x_i) y_i - \log Z \right]$$

$$= \sum_i \left[(\underbrace{\beta_0 + \beta_1 x_0}_{\text{choose } \beta_0 \text{ s.t. this is } 0} + \beta_1 (x_i - x_0)) y_i - \log Z\right]$$

$$\Rightarrow \sum_i \left[y_i \beta (x_i - x_0) - \log\left(1+e^{\beta(x_i - x_0)}\right)\right]$$

$$\Rightarrow \underbrace{\sum_{i \in N_1} \beta(x_i - x_0) - \log\left(1+e^{\beta(x_i - x_0)}\right)}_{\text{const} \to 0} - \underbrace{\sum_{i \in N_2} \log\left(1+e^{\beta(x_i - x_0)}\right)}_{\text{arb neg}}$$

a) Same story but w/ plane

$$\ell(\beta) \to \infty$$

b)
$$\ell(\beta) = \sum_{i=1}^{N} \left[\sum_{k=1}^{K-1} \mathbb{1}(y_i = k)\, \beta_k \cdot x_i - \log\left(1 + \sum_{\ell=1}^{K-1} e^{\beta_\ell \cdot x_i}\right)\right]$$

$$= \sum_k \sum_{i \in S_k} \left[\beta_k \cdot x_i - \log\left(1 + \sum_{\ell=1}^{K-1} e^{\beta_\ell \cdot x_i}\right)\right]$$

$$+ \sum_{i \in S_K} \left[-\log\left(1 + \sum_{\ell=1}^{K-1} e^{\beta_\ell \cdot x_i}\right)\right]$$

$\exists \beta_k$ for $k = 1, \dots, K-1$

s.t. $\beta_k \cdot x > 0 \quad \forall x \in S_k$

$$\Rightarrow \ell(\beta) \to \infty$$

4.6 a) $\exists \beta$ s.t. $\begin{matrix} \beta^T x_i > 0 & \text{if } y_i = 1 \\ \beta^T x_i < 0 & \text{if } y_i = -1 \end{matrix} \Rightarrow y_i \beta^T x_i > 0$

$\Rightarrow y_i \beta^T z_i > 0 \quad z_i = \frac{x_i}{|x_i|}$

if $m = \min y_i \beta^T z_i$ then $\frac{1}{m} y_i \beta^T z_i \geq 1 \quad \forall i$

set $\beta \to \frac{\beta}{m}$

b) $\|\beta_{new} - \beta_{sep}\|^2 = \|\beta_{old} - \beta_{sep} + y_i z_i\|^2$

$= \|\beta_{old} - \beta_{sep}\|^2 + \|y_i z_i\|^2 + 2 y_i (\beta_{old} - \beta_{sep})^T z_i$

$= \|\beta_{old} - \beta_{sep}\|^2 + 1 + \underbrace{2 y_i \beta_{old} z_i}_{<0} - \underbrace{2 y_i \beta_{sep} z_i}_{-2}$

$\leq \|\beta_{old} - \beta_{sep}\|^2 - 1$

since z_i was misclassified before

$\Rightarrow \leq \|\beta_{start} - \beta_{sep}\|^2$ steps

4.7 $D(\beta, \beta_0) = -\sum_i y_i \underbrace{(x_i^T \beta + \beta_0)}_{\text{signed dist to hyperplane}} \qquad \|\beta\| = 1$

No, because no need for optimal sep in case of class imbalance

4.8 $l(\mu, \Sigma) = \frac{-1}{2} \sum_{l=1}^{K} \sum_{g(i)=l} (x_i - \mu_l)^T \Sigma^{-1} (x_i - \mu_l) - N \log |\Sigma|$

4.9 See colab

Chapter 7

Goal ← in fact doesn't seem possible to estimate this using only one T

Test error / generalization error $= Err_T = \mathbb{E}_{X,Y \sim D^{test}}[L(Y, \hat{f}(X)) \mid T]$

Expected prediction error / ie Expected test error $Err = \mathbb{E}_T Err_T = \mathbb{E}_{X,Y} L(Y, \hat{f}(X))$

easier

$$\overline{err} = \frac{1}{N}\sum_i L(y_i, \hat{f}(x_i))$$

bad estimate of test

take $T = \{(x_1, y_1), \dots, (x_N, y_N)\}$

$$Err_T = \mathbb{E}_{X^0, Y^0}[L(Y^0, \hat{f}(X^0)) \mid T]$$

if $X^0 \neq x_1 \dots x_n \Rightarrow$ "extra sample"

in-sample takes $x_1 \cdots x_N$ & new responses $Y_1^0 \dots Y_N^0$

$$Err_{in} = \frac{1}{N}\sum_i \mathbb{E}_{Y^0}[L(Y_i^0, \hat{f}(x_i)) \mid T]$$

$$op = Err_{in} - \overline{err}$$

$$\mathbb{E}_Y\, op =: \omega$$

take $L = |\cdot|^2 \Rightarrow Err_{in} - \overline{err} = \frac{1}{N}\sum_i \cancel{\mathbb{E}_{Y_i^0}[(Y_i^0)^2]} - \cancel{\mathbb{E}_Y(Y^2)}$

$$- 2\,\mathbb{E}_{Y_i^0}[Y_i^0]\,\mathbb{E}[\hat{y}_i] \quad = \mathbb{E}_Y[Y]$$

$$+ 2\,\mathbb{E}_{Y_i}\, Y_i \hat{y}_i$$

$$= \frac{2}{N}\sum_i \mathrm{cov}(y_i, \hat{y}_i)$$

7.1 For lin reg

$$\sum_i \mathrm{Cov}(\hat{y}_i, y_i) = \mathbb{E}_Y\left[y^T X^T (X^T X)^{-1} X y^T\right]$$

$$= \mathrm{Tr}\left[H \underbrace{\mathrm{Cov}(y,y)}_{\sigma_\varepsilon^2 \delta_{ij}}\right]$$

$$= \mathrm{Tr}\, H \, \sigma_\varepsilon^2$$

$$= d\sigma_\varepsilon^2$$

$$\Rightarrow \frac{2}{N}\sum_i \mathrm{Cov}(\hat{y}_i, y_i) = \frac{2d}{N}\sigma_\varepsilon^2$$

$$\Rightarrow E_Y \, \mathrm{Err}_{in} = E_Y \, \overline{\mathrm{err}} + \frac{2d}{N}\sigma_\varepsilon^2 \leftarrow \mathrm{AIC}$$

AIC: $-2\mathbb{E}\log p_\theta(Y) = -2\mathbb{E}\log\mathrm{lik} + \frac{2d}{N}$

7.2 $\Pr(Y=1 \mid x_0) = f(x_0)$ $\qquad \hat{G} = \mathbb{1}[f(x_0) > 1/2]$

First let $G=1 \Rightarrow f(x_0) > 1/2$

$$\mathrm{Err}(x_0) = \Pr(Y \neq \hat{G}(x_0) \mid X = x_0)$$

$$= \Pr(Y=1 \mid X_0 = x_0)\Pr(\hat{G}=0 \mid X=x_0) + \ldots$$

$$= f(x_0)\Pr(\hat{G}=0 \mid X=x_0) + (1-f(x_0))(1-\Pr(\hat{G}=0 \mid X=x_0))$$

$$= 1 - f(x_0) + (2f_0 - 1)\Pr(\hat{G}=0 \mid X=x_0)$$

general form $\rightarrow$

$$= \mathrm{Err}_{Bayes} + |2f_0 - 1|\Pr(\hat{G} \neq G \mid X = x_0)$$

take again $f > 1/2$

$$\Pr(G \neq \hat{G} \mid X=x_0) = P(\hat{G}=0 \mid X=x_0)$$

$$= P(\hat{f}(x_0) < 1/2)$$

$$= Pr\left[\frac{\hat{f}(x_0) - E\hat{f}(x_0)}{\sqrt{Var(x_0)}} < \frac{\frac{1}{2} - E\hat{f}(x_0)}{\sqrt{Var\,\hat{f}(x_0)}}\right]$$

$$= \Phi\left[\frac{|\frac{1}{2} - E\hat{f}(x_0)|}{\sqrt{Var\,\hat{f}(x_0)}} \cdot sgn(\tfrac{1}{2} - EF)\right]$$

only bias

true pred

if $\hat{f}$ is on the wrong side its better to increase the var

7.3 $\hat{f} = Sy$

a) $S_{jj} = x_j^T(X^TX + \lambda\Omega)^{-1}x_j^T$

$\hat{f}(x_j) = x_j^T(X^TX + \lambda\Omega)^{-1}X^Ty$

$\Rightarrow \hat{f}^{-j}(x_j) = x_j(X_{-j}^T X_{-j} + \lambda\Omega)^{-1}X_{-j}^T y_{-j}$

$= x_j(X^TX - x_jx_j^T + \lambda\Omega)^{-1}(X^Ty - x_jy_j)$

$\underbrace{(X^TX + \lambda\Omega)^{-1}}_{A} + \frac{A^{-1}x_jx_j^TA^{-1}}{1 - x_j^TA^{-1}x_j}$

$\hat{f}^{-j}(x_j) = \hat{f}(x_j) - y_jS_{jj} - \frac{x_jA^{-1}x_jx_j^TA^{-1}X^Ty}{1 - x_j^TA^{-1}x_j} + \frac{x_j^TA^{-1}x_jx_j^TA^{-1}x_jy_j}{1 - x_j^TA^{-1}x_j}$ (S_{jj}, S_{jj}, S_{jj})

$= \hat{f}(x_j) - y_jS_{jj} + \frac{S_{jj}\hat{f}(x_j)}{1 - S_{jj}} - \frac{S_{jj}^2y_j}{1 - S_{jj}}$

$= \frac{\hat{f}(x_j)}{1 - S_{jj}} - \frac{y_jS_{jj}}{1 - S_{jj}}$

$= \frac{\hat{f} - y_jS_{jj}}{1 - S_{jj}}$

$\Rightarrow y^j - f^{-j}(x_j) = \frac{y_j - \hat{f}(x_j)}{1 - S_{jj}}$

b) $|y^j - f^{-j}(x_j)| = \frac{|y_j - \hat{f}(x_j)|}{1 - S_{jj}}$

$> |y_j - \hat{f}(x_j)|$

$S = (X^TX + \lambda\Omega)^{-1}$

$S^2 \le S \Rightarrow S_{jj}^2 \le \sum_{k \ne j} S_{jk}^2 + S_{jj}^2$

$\le S_{jj}$

$\Rightarrow 0 \le S_{jj} \le 1$

c) Replace y_i with $f^{-i}(x) =: y'$

$\Rightarrow \hat{f}^{-i} = Sy'$

$= \sum_{j \neq i} S_{ij} y_j + S_{ii} \hat{f}^{-i}$

$= \hat{f}(x_i) - S_{ii} y_i + S_{ii} f^{-i}$

$\Rightarrow y^i - f^{-i}(x_i) = \dfrac{y - f(x_i)}{1 - S_{ii}}$

7.4

take $L = |\cdot|^2 \Rightarrow Err_{in} - \overline{err} = \frac{1}{N} \sum_i \cancel{\mathbb{E}_{Y_i^0}[(Y_i^0)^2] - \mathbb{E}_{Y}(Y^2)}$

maybe wrong

$- 2 \mathbb{E}_{Y_i^0}[Y_i^0] \mathbb{E}_{Y_i}[\hat{y}_i]$ $= \mathbb{E}_Y[Y]$

$+ 2 \mathbb{E}_{Y_i} Y_i \hat{y}_i$

$= \frac{2}{N} \sum_i \mathrm{cov}(Y_i, \hat{y}_i)$

7.5

$\sum_i \mathrm{Cov}(y_i, S_{ij} y_j)$

$= \mathrm{Tr}[S_{ij} \mathrm{Cov}\, y_i, y_j]$

$= \mathrm{Tr}[S_{ij}] \sigma^2$

7.6

$\hat{y} = \frac{1}{k} \sum_{i: x_i \in N_k(x)} y_i = \frac{1}{k} \sum \eta_i y_i$

$\Rightarrow \hat{y} = Sy$

$\Rightarrow$ d.o.f. $= \mathrm{Tr}\, S = \frac{1}{k} \mathrm{Tr}\, \mathbb{1} + \text{off diag} = \frac{N}{k}$

7.7 $$\mathrm{GCV}(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{f}(x_i)}{1 - \frac{\mathrm{Tr}\,S}{N}}\right)^2$$

GCV Approx to S_{ii}

$$\approx \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{f}(x_i)|^2 \left(1 + 2\frac{\mathrm{Tr}\,S}{N}\right)$$

$$= \underbrace{\overline{err} + 2\frac{\mathrm{Tr}\,S}{N}\hat{\sigma}_\varepsilon^2}_{C_p}$$

using MLE rather than an unbiased one