Anurag Dhaipule

H/W 4

Q 1. Orthogonal matrix, not symmetric.

$$A = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

Note $AA^T = I$

2. From the defn + spectral theorem,

$$M = Q \Sigma Q^T$$

where Σ is a diagonal matrix.

$$\Rightarrow \Sigma = Q^T M Q.$$

From calculation of $A^T M A$,

$$\Sigma = Q^T M Q = \begin{pmatrix} q_1^T M q_1 & q_1^T M q_2 & \cdots & q_1^T M q_n \\ q_2^T M q_1 & q_2^T M q_2 & & \vdots \\ \vdots & & \ddots & \\ & & & q_n^T M q_n \end{pmatrix}$$

Here $Q = \begin{pmatrix} | & & | \\ q_1 & \cdots & q_n \\ | & & | \end{pmatrix}$

Since M is psd, $x_1^T M x_1 \geq 0 \Rightarrow$ all diagonal elements (eigen values) are ≥ 0 since diagonals are of the form $x_1^T M x_1$.

3. From psd definition,

$$x^T M x \geq 0$$

Assume $M = BB^T$

$$x^T (BB^T) x$$

$$(B^T x)^T (B^T x)$$

This is of the form $y^T y \geq 0$.

$$\Rightarrow (B^T x)^T (B^T x) \geq 0.$$

$\therefore$ substituting M as BB^T doesn't violate the def^n of psd.

I was also thinking along the lines of

By spectral thm

$$M = \cancel{Q}\, Q \Sigma Q^T$$

If Σ can be write as $x^T x$ somehow,

then $M = \cancel{Qxx^TQ^T} \quad Q x^T x Q^T$

$$= \cancel{Qx}\ (xQ^T)^T (xQ)$$

$$\cancel{= y^T y \geq 0}$$

$$= BB^T$$

$\Rightarrow$ M can be written as BB^T.

3. 1) From method in 2.2

$$\Sigma = \begin{pmatrix} q_1^T M q_1 & \cdots & q_1^T M q_n \\ q_2^T M q_2 & & \vdots \\ \vdots & & q_n^T M q_n \end{pmatrix}$$

Here all terms of the form $x_1^T M x_1 > 0$. This can be seen in diagonal matrix values which are eigen values.

2) $M = Q \Sigma Q^T$; By spectral theorem.

Taking inverse

$$M^{-1} = (Q \Sigma Q^T)^{-1}$$
$$= Q^{T^{-1}} \Sigma^{-1} Q^{-1}$$

But Q is a orthogonal matrix $Q^{-1} = Q^T$

$$M^{-1} = Q \Sigma^{-1} Q^T$$

3) M is psd

$\Rightarrow \Sigma = Q^T M Q$ all eigen values ≥ 0

When $M' = (M + \lambda I)$

$$\Sigma = Q^T (M + \lambda I) Q$$
$$= Q^T M Q + Q^T \lambda Q$$
$$= Q^T M Q + \lambda Q^T Q$$
$$= \geq 0 + > 0$$

$\left(Q^T Q > 0 \right.$ since Q^{-1}...

$\Rightarrow \Sigma > 0$

$$(M + \lambda I)^{-1} = M^{-1} + (\lambda I)^{-1} = Q \Sigma^{-1} Q^T + \frac{I}{\lambda}$$

4) we need to show

$$x^T(M+N)x > 0 \quad \text{for any } x \neq 0$$

$$x^T(M+N)x = x^TMx + x^TNx$$

Since $x^TMx \geq 0$ and $x^TNx > 0$

$$= \geq 0 + > 0$$

$$= > 0.$$

$$\Rightarrow x^T(M+N)x > 0$$

4) $$X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{bmatrix} \qquad XX^T = \begin{bmatrix} \langle x_1, x_1 \rangle & \langle x_1, x_2 \rangle & \cdots & \langle x_1, x_m \rangle \\ \vdots & & & \vdots \\ \langle x_m, x_1 \rangle & \cdots & \cdots & \langle x_m, x_m \rangle \end{bmatrix}$$

Note that diagonal values are of the form $\langle x_n, x_n \rangle$. $\Rightarrow$ diagonal values gives vector lengths.

$$d(x,y) = \|x-y\| = \sqrt{\cancel{\langle x, y \rangle}, \langle x-} = \sqrt{\langle (x-y), (x-y) \rangle}$$

$$= \sqrt{\langle x, x \rangle + \langle y, y \rangle + 2 \langle x, y \rangle}$$

Observe that we have ~~pco~~ every possible pair in XX^T and also vector lengths

hence $d(x,y)$ can be obtained.

5) 1) $J(w) = (Xw-y)^T(Xw-y) + \lambda w^T w$

$$\frac{\partial J(w)}{\partial w} = \frac{\partial}{\partial w}\left[(w^T X^T - y^T)(Xw - y) + \lambda w^T w\right]$$

$$= \frac{\partial}{\partial w}\left[w^T X^T X w - w^T X^T y - y^T X w - y^T y + \lambda w^T w \right]$$

$$= 2X^T X w - 2X^T y + 2\lambda w = 0$$

$$\cancel{(X^T X \quad X^T y} \quad (X^T X + \lambda I) w = X^T y$$

$$\Rightarrow \quad w = (X^T X + \lambda I)^{-1} X^T y.$$

Since $\cancel{X^T X} \rightarrow \underbrace{X^T X}_{\geq 0 \text{ psd}} + \lambda I$

psd $+ \lambda I$ is invertible for $\lambda > 0$.

from previous exercise.

2) $X^T X w + \lambda I w = X^T y$

$$\lambda I w = X^T y - X^T X w$$

$$w = \frac{1}{\lambda}(X^T y - X^T X w)$$

$$w = X^T \left(\frac{y - Xw}{\lambda}\right)$$

$$= X^T \alpha \quad \text{where} \quad \alpha = \frac{y - Xw}{\lambda}$$

3) Since $\alpha = \lambda^{-1}(y - Xw) \in \mathbb{R}^n$

we

$$w = \begin{pmatrix} | & \cdots & | \\ x_1 & & x_n \\ | & \cdots & | \end{pmatrix} \begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_n \end{pmatrix} = \alpha_1 x_1 + \cdots \alpha_n x_n$$

Hence w is in span of data.

4) We have

$$\alpha = \lambda^{-1}(y - Xw)$$
$$\lambda \alpha = y - Xw$$

But $w = X^T \alpha$

$$\lambda \alpha = y - XX^T \alpha$$
$$\lambda \alpha + XX^T \alpha = y$$
$$(\lambda I + XX^T)\alpha = y$$
$$\alpha = (\lambda I + XX^T)^{-1} y$$

5) $Xw = X(X^T \alpha)$
$$= (XX^T)(\lambda I + XX^T)^{-1} y$$

6) $f(x) = x^T w^*$
$$= x^T (X^T \alpha)$$
$$= \cancel{x^T (X} \; x^T X (\lambda I + XX^T)^{-1} y$$

For convenience writing,

$$K_x = \begin{pmatrix} x^T x_1 \\ \vdots \\ x^T x_n \end{pmatrix} = x^T X$$

$$f(x) = K_x (\lambda I + XX^T)^{-1} y$$

6.1

① Splitting trees based on different splitting variables and values

Colour :

White (5)

3N, 2Y

$P_1 = 2/5$

$q_1 = 12/25$

Brown (6)

3N, 3Y

$P_2 = 1/2$

$q_2 = 1/2$

$I = 5 * 12/25 + 6 * 1/2 = 27/5$

Spots :

N (4)

4N, 0Y

$P_1 = 0$

$q_1 = 0$

Y (7)

2N, 5Y

$P_2 = 5/7$

$q_2 = 20/49$

$I = \frac{20}{49} * 7 = 20/7$

Size : $<1.5\ (3)$

$1N,\ 2Y$

$P_1 = 2/3$

$Q_1 = 4/9$

$>1.5\ (8)$

$5N,\ 3Y$

$P_2 = 3/8$

$Q_2 = 8 * 3/8 * 5/8 = 15/32$

$I = 4/3 + \frac{15}{4} = \frac{61}{12}$

Size : $<2.5\ (5)$

$3N,\ 2Y$

$P_1 = 2/5$

$Q_1 = 12/25$

$>2.5\ (6)$

$3N,\ 3Y$

$P_2 = 1/2$

$Q_2 = 1/2$

$I = 12/5 + 3 = 27/5$

Size : $<3.5\ (6)$

$4\cancel{8}N,\ \cancel{5Y}\ 2Y$

$P_1 = 2/6 = 1/3$

$Q_1 = 4/9 \times 6$

$I = \frac{8}{3} + \frac{12}{5} = \frac{76}{15}$

$>3.5\ (5)$

$2N,\ 3Y$

$P_2 = 3/5$

$Q_2 = 12/25$

Size : $<4.5\ (9)$

~~4N~~, $5N,\ 4Y$

$P_1 = \frac{4}{9}$

$Q_1 = \frac{40}{81}$

$I = \frac{40}{9} + 1 = \frac{49}{9}$

$>4.5\ (2)$

$1N,\ 1Y$

$P_2 = 1/2$

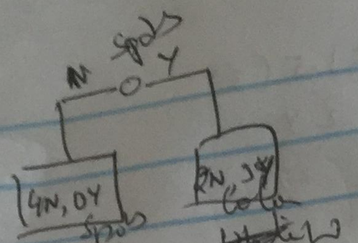
$Q_2 = 1/2$

Hence, the best possible split is based on spots which has lowest I of $\frac{20}{7}$.

(2) Same way, we determine the second split.

91.76

(2)



P	Size	Spots	~~~~
N	2	Y	B
N	3	Y	W
Y	5	Y	B
Y	4	Y	B
Y	4	Y	W
Y	1	Y	B
Y	1	Y	

color :

B (4)
N : 1 Y : 3
$P_1 = 3/4$
$q = \frac{6}{16}$

W (3)
N : 1 Y : 2
$P_2 = 2/3$
$q = 4/9$

$I = \frac{6}{4} + 4/3 = \frac{39}{12}$

Size

< 1.5 (2)
N : 0 Y : 2
$p = 1$
$q = 0$

> 1.5 (5)
N : 2 Y : 3
$P : 3/5$
$q = 12/25$

$I = 12/5$

Size

< 2.5 (3)
N : 1 Y : 2
$P_1 = 2/3$
$q_1 = 4/9$

> 2.5 (4)
N : 1 Y : 3
$P_2 = 3/4$
$q_2 = 6/16$

$I = 4/3 + \frac{6}{4} = \frac{34}{12}$

$\boxed{\text{Size}}$ Right split

< 3.5 (4)

N : 2 Y : 2

$p_1 = 1/2$

$q_1 = 1/2$

$8 = 2$

> 3.5

N : 0 Y : 3

$p_2 = 1$

$q_1 = 0$

Size

< 4.5 (6)

N : 2 Y : 4

$p_1 = 2/3$

$q_1 = 4/9$

$p_1 = 4/9 \times 6 = 8/3$

> 4.5

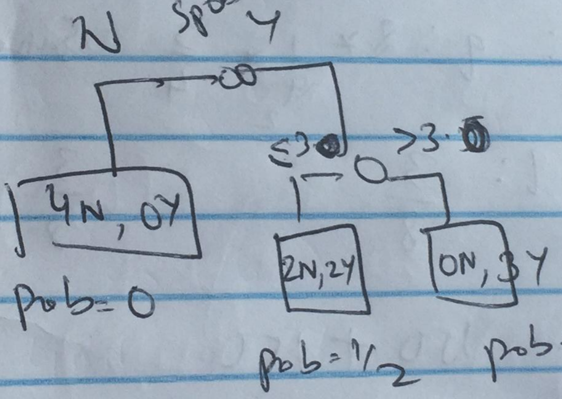
N : 0 Y : 1

$p_2 = 1$

$q_2 = 0$

So, the second best split is based on size > 3

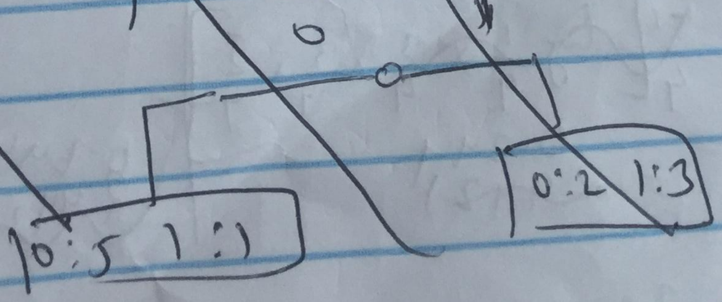
(2)



probability of poisonous =

(3) A : 0(6) | 1(5)

0 : 5 1 : 1 | 0 : 2 1 : 3



Training error : $\frac{2}{11}$

second split.

③ Note that there are two input values for which two different o/p are in training data.

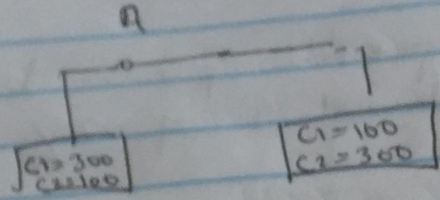
A	B	C	Y
0	0	1	0 and 1
1	1	1	0 and 1

There is no way, ~~the~~ 001 (Y=0) and 001 (Y=1) be split based on input, and hence both these value will come in the same leaf node. Since there are two such values in ~~a~~ training data

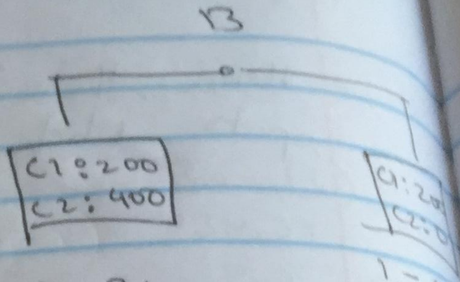
$$\text{Error} = \frac{2}{11} \times 100 = \frac{200}{11}\,\%$$

6.2 ① (7)

A



B



Misclassification:

A: $1 - 3/4$ $1 - 3/4$

$q = 1/4$ $q = 1/4$

$I = 400 * 1/4 + 400 * 1/4$

$I = 200$

B: $1 - 2/3$ $1 - \frac{1}{1}$

$q = 1/3$ $q = 0$

$I = 600 \times 1/3 + 0$

$= 200$

Gini index

A: $q = 2 * 1/4 * 3/4$ $2 * 1/4 * 3/4$

$q = 3/8$ $q = 3/8$

$I = 3/8 * 400 + 3/8 * 400$

$= 300$

B: $q = 2 * 1/3 * 2/3$ $q = 0$

$q = 4/9$ $q = 0$

$I = 4/9 * 600 + 0$

$= 800/3 = 266.17$

Entropy

A: $q = -\left[\frac{1}{4}\log\frac{1}{4} + \frac{3}{4}\log\frac{3}{4}\right]$ $q = \left[\frac{1}{4}\log\frac{1}{4} + \frac{3}{4}\log\frac{3}{4}\right]$

$q = 0.244$ $q = 0.244$

$I = 0.244 * 400 + 0.244 * 400$

$I = 195.2$

B: $q = -\left[\frac{1}{3}\log\frac{1}{3} + \frac{2}{3}\log\frac{2}{3}\right]$ $q = -[1 \log 1]$

$q = 0.2764$ $q = 0$

$I = 0.2764 * 600$

$I = 165.86$

Note: Gini index and Entropy impurity gives less value for tree B.

(7) (1) From pythagorean theorem

$$\|x\|^2 = \|m_0\|^2 + \|x - m_0\|^2$$

$$\langle x, x\rangle = \langle m_0, m_0\rangle + \langle x - m_0, x - m_0\rangle$$

Note that all these terms are of the form,

$\langle x, x\rangle \geq 0$ and $\langle x, x\rangle = 0$ if $x - m_0 = 0$

$\Rightarrow \langle x - m_0, x - m_0\rangle \geq 0$ and

$\langle x - m_0, x - m_0\rangle = 0$ when $x - m_0 = 0$

$\Rightarrow x = m_0$

$\therefore$ When $x = m_0$

$\langle x, x\rangle = \langle m_0, m_0\rangle$

$\|x\| = \|m_0\|$

(2) In continuation to the proof of representor theorem,

$\|w\| \leq \|w^*\|$

~~that~~ If $\|w\| < \|w^*\| \Rightarrow R(\|w\|) < R(\|w^*\|)$

This implies we are reaching the minimizer in the span of data and hence

$$w = \sum_{i=1}^{n} \alpha\, \varphi(x_i)$$

But when $\|w\| = \|w^*\|$

Then $R(\|w\|) \not< R(\|w^*\|) \Rightarrow R(\|w\|) = R(w^*)$

This is not being considered since given that R is increasing.

(8.2) (1) Given $\min \phi(w)$
subject to $\Omega(w) \leq r$

$$L(w,\lambda) = \phi(w) + \lambda(\Omega(w) - r)$$

(2) $$p^* = \min_{w} \sup_{\lambda \geq 0} L(w,\lambda)$$

$$g^* = \sup_{\lambda \geq 0} \min_{w} L(w,\lambda) \qquad g^* = \sup_{\lambda \geq 0} g(\lambda)$$

Dual objective function $g(\lambda) = \min_{w} L(w,\lambda)$

(3) ~~by~~ strong duality exists and complementary slackness

$$g(\lambda^*) = \min_{w} L(w, \lambda^*)$$

$$= \min_{w} \left(\phi(w) + \lambda^* \left[\Omega(w) - \lambda^* r \right] \right.$$

Note that λr ~~and~~ is constant

$$\Rightarrow \min_{w} \arg \left(\phi(w) + \lambda^*(\Omega(w) - \lambda^* r) \right) = \min_{w} \arg \left[\phi(w) + \lambda\Omega \right]$$

$$\Rightarrow w^* = \arg\min \left[\phi(w) + \lambda \Omega(w) \right]$$

(8.3)

Q3 ① $J(\theta) = (Xw-y)^T (Xw-y)$

$\|w\|^2 \leq r$

If ~~$\|w\|$~~ $w=0$, then $\|w\| < r$ if $r>0$

We assume r is greater than 0 because it doesn't make sense to have $r=0$

$\Rightarrow$ for $w=0$, it is strictly feasible.