154 Midterm Two study guide solution

1)

1. F
2. F (Counter example: regression trees)
3. T
4. F (Issues: might overfit the data)
5. F
6. F (Expit instead of Gaussian)
7. F
8. F

2)

1.
Similarities: (a).$X|Y = k \sim$ Normal (b).require prior prob (generative modelling) (c).predictions are based on posterior prob.
Differences: (a).LDA assumes common within group covariance matrix, while QDA assumes different covariance. (b).LDA $\sim$ linear decision boundary, while QDA $\sim$ quadratic decision boundary.

2. Linear Regression yields value outside [0,1], while logistic regression does not.

3.
Advantages: (a).easy to explain (b).closer to human decision making (c).can be displayed graphically (d).can easily handle qualitative variables.
Disadvantages: (a).generally weaker productive power (b).easy to overfit.

3)

(a).

$$
\begin{aligned}
TSS &= \sum_{k=1}^{K}\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_k + \bar{x}_k - \bar{x})^2 \\
&= \sum_{k=1}^{K}\sum_{i=1}^{n_k}[(x_{ik} - \bar{x}_k)^2 + (\bar{x}_k - \bar{x})^2 + 2 * (x_{ik} - \bar{x}_k)(\bar{x}_k - \bar{x})] \\
&= WSS + BSS + \sum_{k=1}^{K}\sum_{i=1}^{n_k} 2 * (x_{ik} - \bar{x}_k)(\bar{x}_k - \bar{x}) \\
&= WSS + BSS + 2\sum_{k=1}^{K}(\bar{x}_k - \bar{x})\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_k) \\
&= WSS + BSS
\end{aligned}
$$

where $\sum_{i=1}^{n_k}(x_{ik} - \bar{x}_k) = 0$

(b).

$$\delta^2 = \frac{BSS}{TSS} = \frac{BSS}{BSS + WSS} \in [0, 1]$$

If $\delta^2 = 0$, then BSS = 0, which interprets no BSS.
If $\delta^2 = 1$, them WSS = 0, which interprets no WSS :)

(c).

$$F = \frac{n-K}{K-1}\frac{BSS}{TSS - BSS} = \frac{n-K}{K-1}\frac{1}{\frac{1}{n^2} - 1}$$

which shows one-to-one relationship.

4)

performance metrics $< -$ function(cm)
a $< -$ cm[1,1]
b $< -$ cm[1,2]
c $< -$ cm[2,1]
d $< -$ cm[2,2]
tpr $< -$ a/(a+c)
fpr $< -$ b/(b+d)
tnr $< -$ d/(b+d)
fnr $< -$ c/(a+c)
specifying $< -$ 1 - fpr
sensitivity $< -$ tpr
return(c(.......))