2.6 We wish to find a function $f(x)$ such that f minimizes $\sum_{i=1}^{N}(y_i - f(x_i))^2$, and we are told that some or all of the observations have identical x values. Let $(x_i^*, \ldots, x_k^*)$ be the unique values of x, indexed from 1 to K, and $N_i = \{j : x_j = x_i^*\}$ be the set of indices of x which map to x_i^*. This gives:

$$\begin{aligned}
\sum_{i=1}^{K}\sum_{j\in N_i}(y_j - f(x_i^*))^2 &= \sum_{i=1}^{K}\sum_{j\in N_i}(y_j^2 - 2f(x_i^*)y_j + f(x_i^*)^2) \\
&= \sum_{i=1}^{K}\left[\sum_{j\in N_i} y_j^2 - 2f(x_i^*)\sum_{j\in N_i} y_j + |N_i|f(x_i^*)^2\right] \\
&= \sum_{i=1}^{K}\left[\sum_{j\in N_i} y_j^2 - 2f(x_i^*)|N_i|\bar{y}_i + |N_i|f(x_i^*)^2\right] \\
&= \sum_{i=1}^{K}\sum_{j\in N_i} y_i^2 - \sum_{i=1}^{K}\left[|N_i|f(x_i^*)^2 - 2f(x_i^*)|N_i|\bar{y}_i + |N_i|\bar{y_i}^2 - |N_i|\bar{y_i}^2\right] \\
&= \sum_{i=1}^{K}\sum_{j\in N_i} y_i^2 - \sum_{i=1}^{K}\left[|N_i|(f(x_i^*) - \bar{y}_i)^2\right] - \sum_{i=1}^{K}|N_i|\bar{y_i}^2
\end{aligned}$$

Leaving us to minimize $\sum_{i=1}^{K}\left[|N_i|(f(x_i^*) - \bar{y}_i)^2\right]$ which is a weighted least squares problem on the reduced data set $(x_1, \ldots, x_k)$

2.7 (a) If $\hat{f}(x_i)$ is a linear function, then for arbitrary $\mathbf{x_0} = (x_1, \ldots, x_n) \in \chi$ we have

$$\begin{aligned}
\hat{f}(\mathbf{x_0}) &= \hat{\beta}_0 + \hat{\beta}_1\mathbf{x_0} \\
&= \hat{\beta}\mathbf{x_0} \\
&= [\mathbf{1}\ \mathbf{x_0^T}](\mathbf{X^TX})^{-\mathbf{1}}\mathbf{X^Ty} \\
&= \sum_{i=1}^{N}[\mathbf{1}\ \mathbf{x_0^T}](\mathbf{X^TX})^{-\mathbf{1}}\mathbf{X^T}y_i
\end{aligned}$$

Which is a linear estimator in y_i with $l_i(x;\chi) = [\mathbf{1}\ \mathbf{x_0^T}](\mathbf{X^TX})^{-\mathbf{1}}\mathbf{X^T}$.
On the other hand, if $\hat{f}(x_i)$ is a K-nearest neighbor function,
then for arbitrary $\mathbf{x_0} = (x_1, \ldots, x_n) \in \chi$ we have

$$\begin{aligned}
\hat{f}(\mathbf{x_0}) &= \sum_{i\in N_K(\mathbf{x_0})}\frac{y_i}{K} \\
&= \sum_{i=1}^{N}\frac{1}{K}I(y_i \in N_K(\mathbf{x_0}))
\end{aligned}$$

Which is a linear estimator in y_i with $l_i(x;\chi) = \frac{1}{K}$.

(b) $E_{Y|X}\left[(f(\mathbf{x_0}) - \hat{f}(\mathbf{x_0}))^2\right] = Var_{Y|X}\left[\hat{f}(\mathbf{x_0}))\right] + \left(E_{y|x}\hat{f}(\mathbf{x_0}) - f(x_0)\right)^2$

In the case where the estimator is a weighted sum of the y_i's we have

$E_{Y|X}\left[\hat{f}(\mathbf{x_0})\right] = \sum_{i=1}^{n} l_i E[y_i] = \sum_{i=1}^{n} l_i \hat{f}(\mathbf{x_i})$ and $Var_{Y|X}\left[\hat{f}(\mathbf{x_0})\right] = \sum_{i=1}^{n} l_i^2 \sigma^2$

So we have:
$Var_{Y|X} = \sum_{i=1}^{n} l_i^2 \sigma^2$
$Bias_{Y|X} = \sum_{i=1}^{n} l_i \hat{f}(\mathbf{x_i}) - f(\mathbf{x_0})$

2.8 The training and testing error of the K-nearest neighbor classifier on digits 2 and 3 from the zip code data is

4 (a) The training and testing error of the K-nearest neighbor classifier on digits 1,2 and 3 from the zip code data is

(b) The training error for the LDA classifier on digits 1,2,and 3 is signif(lda.train2.error,2) and the testing error is signif(lda.test2.error,2)

5 (a) The decision rule for a binary classification is a function $\hat{f}(\mathbf{x})$ such that $\hat{f}(\mathbf{x}) = 1$ if the ratio of the posterior probability masses for class one vs. class two at $\mathbf{x}$ is greater than 1, and 2 otherwise. Here we know the posterior masses exactly, so we can calculate the Bayes decision rule.

$$\begin{aligned}
\frac{(2\pi|\mathbf{\Sigma_1}|)^{-\frac{1}{2}} exp\left[-\frac{1}{2}(\mathbf{x} - \mu_\mathbf{1})^T \mathbf{\Sigma_1^{-1}}(\mathbf{x} - \mu_\mathbf{1})\right]}{(2\pi|\mathbf{\Sigma_2}|)^{-\frac{1}{2}} exp\left[-\frac{1}{2}(\mathbf{x} - \mu_\mathbf{2})^T \mathbf{\Sigma_2^{-1}}(\mathbf{x} - \mu_\mathbf{2})\right]} &> 1 \\
\frac{exp\left[-\frac{1}{2}(\mathbf{x} - \mu_\mathbf{1})^T \mathbf{\Sigma_1^{-1}}(\mathbf{x} - \mu_\mathbf{1})\right]}{exp\left[-\frac{1}{2}(\mathbf{x} - \mu_\mathbf{2})^T \mathbf{\Sigma_2^{-1}}(\mathbf{x} - \mu_\mathbf{2})\right]} &> \left(\frac{|\mathbf{\Sigma_2}|}{|\mathbf{\Sigma_1}|}\right)^{-\frac{1}{2}} \\
(\mathbf{x} - \mu_\mathbf{1})^T \mathbf{\Sigma_1^{-1}}(\mathbf{x} - \mu_\mathbf{1}) - (\mathbf{x} - \mu_\mathbf{2})^T \mathbf{\Sigma_2^{-1}}(\mathbf{x} - \mu_\mathbf{2}) &< log\left(\frac{|\mathbf{\Sigma_2}|}{|\mathbf{\Sigma_1}|}\right)
\end{aligned}$$

So $\hat{f}(\mathbf{x}) = 1$ if $(\mathbf{x} - \mu_\mathbf{1})^T \mathbf{\Sigma_1^{-1}}(\mathbf{x} - \mu_\mathbf{1}) - (\mathbf{x} - \mu_\mathbf{2})^T \mathbf{\Sigma_2^{-1}}(\mathbf{x} - \mu_\mathbf{2}) < 1.386$ and 2 otherwise.

(b) The Bayes, LDA, and QDA training and testing error percentages for 200 training and 2000 testing points from the scenario in part (a), with equal class priors.

4.2 (a) LDA assumes that each class comes from a multivariate gaussian distribution, and that the classes have equal covariance. Let $\pi_1 = P(Y = 1)$ and $\pi_2 = P(Y = 2)$, and $\mathbf{X}|\mathbf{Y}\ N(\mu_i, \mathbf{\Sigma})$. The LDA classifies to 2 when we have

$$\begin{aligned}
\frac{P(Y=2|\mathbf{X}=\mathbf{x})}{P(Y=1|\mathbf{X}=\mathbf{x})} &> 1 \\
\frac{\pi_2(2\pi\mathbf{\Sigma})^{-\frac{1}{2}} exp\left[-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{2})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_\mathbf{2})\right]}{\pi_1(2\pi\mathbf{\Sigma})^{-\frac{1}{2}} exp\left[-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{1})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_\mathbf{1})\right]} &> 1 \\
\frac{exp\left[-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{2})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_\mathbf{2})\right]}{exp\left[-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{1})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_\mathbf{1})\right]} &> \frac{\pi_1}{\pi_2} \\
\left[-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{2})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_\mathbf{2})\right] - \left[-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{1})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\mu_\mathbf{1})\right] &> log(\pi_1) - log(\pi_2)
\end{aligned}$$

We can multiply this out, cancelling the $\mathbf{x}^T\mathbf{\Sigma}\mathbf{x}$ terms, using the symmetry of the covariance matrix to combine the $\mathbf{x}^T\mathbf{\Sigma}\mu_\mathbf{2}$ and $\mu_\mathbf{2}{}^T\mathbf{\Sigma}\mathbf{x}$ terms, and we get

$$\begin{aligned}
\left(\mathbf{x}^T\mathbf{\Sigma}\mu_\mathbf{2} - \mathbf{x}^T\mathbf{\Sigma}\mu_\mathbf{1}\right) - \frac{1}{2}\left(\mu_\mathbf{2}{}^T\mathbf{\Sigma}\mu_\mathbf{2} - \mu_\mathbf{1}{}^T\mathbf{\Sigma}\mu_\mathbf{1}\right) &> log(\pi_1) - log(\pi_2) \\
\mathbf{x}^T\mathbf{\Sigma}\left(\mu_\mathbf{2} - \mu_\mathbf{1}\right) &> \frac{1}{2}\mu_\mathbf{2}{}^T\mathbf{\Sigma}\mu_\mathbf{2} - \frac{1}{2}\mu_\mathbf{1}{}^T\mathbf{\Sigma}\mu_\mathbf{1} + log\left(\frac{N_1}{N}\right) - log\left(\frac{N_2}{N}\right)
\end{aligned}$$

As required.

(b) We wish to minimize $\sum_{i=1}^{N}(y_i - \beta_0 - \beta_1 x_i)^2 = \sum_{i=1}^{N}(y_i - \beta x_i)^2$ in terms of β. This is just an OLS minimization and we know that β will satisfy

$$\mathbf{X^T X}\beta = \mathbf{X^T y}$$

which is in the same form as the equation which is our target. The matrices $\mathbf{X}$ and $\mathbf{Y}$ are

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & \dots & 1 \\ x_1 & x_2 & \dots & x_n \end{bmatrix}$$

$$\mathbf{Y} = \begin{bmatrix} -\frac{N}{N_1} & -\frac{N}{N_1} & \dots & \frac{N}{N_2} & \frac{N}{N_2} \end{bmatrix}$$