Elements of Statistical Learning by Hastie, Tibshirani, and Friedman

Worked Exercises: Chapter 2

Kevin O'Connor

July 1, 2018

Ex. 2.1

Suppose each of K-classes has an associated target t_k, which is a vector of all zeroes, except a one in the kth position. Show that classifying to the largest element of $\hat{y}$ amounts to choosing the closest target, $\min_k \|t_k - \hat{y}\|$, if the elements of $\hat{y}$ sum to one.

Proof. Here I will use $\hat{y}_k$ to denote the kth component of $\hat{y}$. To rephrase the question, we want to prove that

$$\operatorname{argmax}_k \hat{y}_k = \operatorname{argmin}_k \|t_k - \hat{y}\|$$

Note first that we can write the squared distance from t_k to $\hat{y}$ as

$$\|t_k - \hat{y}\|^2 = \hat{y}_1^2 + ... + (1 - \hat{y}_k)^2 + ... + \hat{y}_K^2$$

Now fix i such that $\max_k \hat{y}_k = \hat{y}_i$ and $1 \leqslant j \leqslant K$, $j \neq i$. Then

$$\begin{aligned}
\|t_i - \hat{y}\|^2 - \|t_j - \hat{y}\|^2 &= (1 - \hat{y}_i)^2 + \hat{y}_j^2 - \hat{y}_i^2 - (1 - \hat{y}_j)^2 \\
&= 1 - 2\hat{y}_i + \hat{y}_i^2 + \hat{y}_j^2 - \hat{y}_i^2 - 1 + 2\hat{y}_j - \hat{y}_j^2 \\
&= 2(\hat{y}_j - \hat{y}_i) \\
&\leqslant 0
\end{aligned}$$

So $\|t_i - \hat{y}\|^2 \leqslant \|t_j - \hat{y}\|^2$, $\forall j \neq i$, which gives $\|t_i - \hat{y}\| \leqslant \|t_j - \hat{y}\|$, $\forall j \neq i$. Thus we have

$$\operatorname{argmax}_k \hat{y}_k = \hat{y}_i = \operatorname{argmin}_k \|t_k - \hat{y}\|$$

Assuming the maximum element of $\hat{y}$ is unique, this gives us the result. ■

Ex. 2.2

Show how to compute the Bayes decision boundary for the simulation example in Figure 2.5.

Solution. First I'll explain the model the simulation uses. 10 means, $m_1, ..., m_{10}$, are generated for each class (*Blue* and *Orange*) from distributions,

$$\mathcal{N}\left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, I_2\right) \quad \text{and} \quad \mathcal{N}\left(\begin{bmatrix} 0 \\ 1 \end{bmatrix}, I_2\right)$$

respectively. Then 100 observations are drawn from each class by choosing a mean, m_k, from the 10 corresponding to its class, uniformly at random and then generating a point from the distribution, $\mathcal{N}(m_k, I_2/5)$. Now fix an arbitrary data point, x. Using Bayes theorem, we have

$$\mathbb{P}(Orange|x) = \frac{\mathbb{P}(x|Orange)\mathbb{P}(Orange)}{\mathbb{P}(x)}$$

The classes *Blue* and *Orange* occur with equal frequency so $\mathbb{P}(Blue) = \mathbb{P}(Orange) = 1/2$. At this point, $\mathbb{P}(x|Orange)$ is not directly computable due to the unobserved mean on which they depend. But we can further expand these probabilities to make this dependence explicit. Note that we have three latent variables at work here: k, m_k, and color.

$$\begin{aligned}
\mathbb{P}(x|Orange) &= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\sum_{k=1}^{10}\mathbb{P}(x, m_k, k|Orange)dm_k \\
&= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\sum_{k=1}^{10}\mathbb{P}(x|m_k, k, Orange)\mathbb{P}(m_k|k, Orange)\mathbb{P}(k|Orange)dm_k \\
&= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\sum_{k=1}^{10}\Bigg[\left(\frac{5}{2\pi}\exp\left\{-\frac{5}{2}(x-m_k)^T(x-m_k)\right\}\right) \times \\
&\qquad \left(\tfrac{1}{2\pi}\exp\left\{-\tfrac{1}{2}\left(m_k - \begin{bmatrix}0\\1\end{bmatrix}\right)^T\left(m_k - \begin{bmatrix}0\\1\end{bmatrix}\right)\right\}\right)\Bigg]\left(\tfrac{1}{10}\right)dm_k \\
&= \int\limits_{-\infty}^{\infty}\int\limits_{-\infty}^{\infty}\Bigg[\left(\frac{5}{2\pi}\exp\left\{-\frac{5}{2}(x-m_k)^T(x-m_k)\right\}\right) \times \\
&\qquad \left(\tfrac{1}{2\pi}\exp\left\{-\tfrac{1}{2}\left(m_k - \begin{bmatrix}0\\1\end{bmatrix}\right)^T\left(m_k - \begin{bmatrix}0\\1\end{bmatrix}\right)\right\}\right)\Bigg]dm_k
\end{aligned}$$

Then the boundary is found by setting $\mathbb{P}(x|Orange) = 1/2$ and solving for x. ■

Ex. 2.3

Derive equation (2.24).

Proof. To remind ourselves of what equation (2.24) is, it says that for N data points uniformly distributed over a p-dimensional unit ball, the median distance to the closest data point is given by

$$d(p, N) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p}$$

First suppose that the volume contained in a p-dimensional ball of radius r is given by

$$V(r) = C_p r^p$$

for some constant C_p. Now since the data points are uniformly distributed, their cdf depends on the volume of the ball. What do I mean by this? For a single point, x,

$$\mathbb{P}(\|x\| \leqslant r) = \frac{V(r)}{V(1)} \mathbb{1}_{(0,1)}(r) + \mathbb{1}_{[1,\infty)}(r) = r^p \mathbb{1}_{(0,1)}(r) + \mathbb{1}_{[1,\infty)}(r)$$

Now let's restrict our consideration to the interesting case when $r \in (0, 1)$. Then we have

$$\mathbb{P}(\|x\| \geqslant r) = 1 - r^p$$

Assuming the data points are drawn independently, this gives us the joint probability,

$$\mathbb{P}\left(\min_i \|x_i\| \geqslant r\right) = \mathbb{P}\left(\|x_i\| \geqslant r,\, \forall 1 \leqslant i \leqslant N\right) = \prod_{i=1}^{N} \mathbb{P}(\|x_i\| \geqslant r) = (1 - r^p)^N$$

This gives us the distribution of distances from the origin to the closest data point. We are asked for the median of this distribution so all we need to do is set it equal to 1/2 and solve for r.

$$(1 - r^p)^N = 1/2 \Longrightarrow r = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p}$$

This gives us the solution. ■

Ex. 2.4

The edge effect problem discussed on page 23 is not peculiar to uniform sampling from bounded domains. Consider inputs drawn from a spherical multinormal distribution, $X \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p)$. The squared distance from any sample point to the origin has a χ^2_p distribution with mean p. Consider a prediction point x_0 drawn from this distribution, and let $a = x_0/\|x_0\|$ be an associated unit vector. Let $z_i = a^T x_i$ be the projection of each of the training points on this direction.

Show that the z_i are distributed $\mathcal{N}(0, 1)$ with expected squared distance from the origin 1, while the target point has expected squared distance p from the origin.

Hence for $p = 10$, a randomly drawn test point is about 3.1 standard deviations from the origin, while all the training points are on average one standard deviation along direction a. So most prediction points see themselves as lying on the edge of the training set.

Proof. The projection characterization of the multinormal distribution says that a random variable X is multinormal if and only if all of its 1-dimensional projections have a univariate normal distribution. Thus $z_i = a^T x_i$ is normally distributed. Since the normal distribution is characterized by its mean and variance, we need only compute the mean and variance of z_i.

$$\mathbb{E}z_i = \mathbb{E}[a^T x_i] = a^T \mathbb{E}x_i = a^t \mathbf{0}_p = 0$$

$$\operatorname{Var}(z_i) = \operatorname{Var}(a^T x_i) = a^T \operatorname{Var}(x_i) a = a^T \mathbf{I}_p a = a^T a = 1$$

Thus $z_i = a^T x_i \sim \mathcal{N}(0,1)$. It follows from this that $\mathbb{E}z_i^2 = \operatorname{Var}(z_i) + (\mathbb{E}z_i)^2 = 1$. On the other hand, as mentioned $x_i^T x_i \sim \chi_p^2$ so $\mathbb{E}[\|x_i\|^2] = \mathbb{E}[x_i^T x_i] = p$. ■

Ex. 2.5

(a) Derive equation (2.27). The last line makes use of (3.8) through a conditioning argument.

Proof. To remind ourselves of the situation in question here, we have a linear model,

$$Y = X^T \beta + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

and are interested in the expected prediction error (EPE) at a point, x_0. Equation (2.27) states that

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

where τ is the training set. The authors give us an outline of the derivation in the chapter but we will walk through each line of the derivation and explain where things come from. First we have

$$EPE(x_0) = \mathbb{E}\left[\mathbb{E}_\tau \left[(y_0 - \hat{y}_0)^2 \right] | x_0 \right] = \mathbb{E}_{y_0|x_0} \left[\mathbb{E}_\tau \left[(y_0 - \hat{y}_0)^2 \right] \right]$$

which is just the definition of the expected prediction error at x_0 and some alternate notation. Now we will add and subtract $\mathbb{E}_\tau \hat{y}_0$ in the inner expectation.

$$\begin{aligned}
\mathbb{E}_\tau [(y_0 - \hat{y}_0)^2] &= \mathbb{E}_\tau [(y_0 - \mathbb{E}_\tau \hat{y}_0 + \mathbb{E}_\tau \hat{y}_0 - \hat{y}_0)^2] \\
&= \mathbb{E}_\tau [(y_0 - \mathbb{E}_\tau \hat{y}_0)^2] + 2\mathbb{E}_\tau [(y_0 - \mathbb{E}_\tau \hat{y}_0)(\mathbb{E}_\tau \hat{y}_0 - \hat{y}_0)] \\
&\quad + \mathbb{E}_\tau [(\mathbb{E}_\tau \hat{y}_0 - \hat{y}_0)^2] \\
&= \mathbb{E}_\tau [(y_0 - \mathbb{E}_\tau \hat{y}_0)^2] + 2(y_0 - \mathbb{E}_\tau \hat{y}_0)\mathbb{E}_\tau [(\mathbb{E}_\tau \hat{y}_0 - \hat{y}_0)] \\
&\quad + \mathbb{E}_\tau [(\mathbb{E}_\tau \hat{y}_0 - \hat{y}_0)^2] \\
&= (y_0 - \mathbb{E}_\tau \hat{y}_0)^2 + \mathbb{E}_\tau [(\mathbb{E}_\tau \hat{y}_0 - \hat{y}_0)^2]
\end{aligned}$$

where we used the fact above that $y_0 \perp\!\!\!\perp \tau$. Now add and subtract $\mathbb{E}y_0$ within the first term.

$$
\begin{aligned}
(y_0 - \mathbb{E}_\tau \hat{y}_0)^2 &= (y_0 - \mathbb{E}y_0 + \mathbb{E}y_0 - \mathbb{E}_\tau \hat{y}_0)^2 \\
&= (y_0 - \mathbb{E}y_0)^2 + 2(y_0 - \mathbb{E}y_0)(\mathbb{E}y_0 - \mathbb{E}_\tau \hat{y}_0) + (\mathbb{E}y_0 - \mathbb{E}_\tau \hat{y}_0)^2
\end{aligned}
$$

Which gives,

$$
\begin{aligned}
\mathbb{E}\left[(y_0 - \mathbb{E}_\tau \hat{y}_0)^2 | x_0\right] &= \mathbb{E}\left[(y_0 - \mathbb{E}y_0)^2 | x_0\right] + 2(\mathbb{E}y_0 - \mathbb{E}_\tau \hat{y}_0)\mathbb{E}\left[y_0 - \mathbb{E}y_0 | x_0\right] \\
&\quad + (\mathbb{E}y_0 - \mathbb{E}_\tau \hat{y}_0)^2 \\
&= \mathrm{Var}(y_0|x_0) + (\mathbb{E}y_0 - \mathbb{E}_\tau \hat{y}_0)^2
\end{aligned}
$$

We can combine these steps to get

$$
\begin{aligned}
EPE(x_0) &= \mathbb{E}\left[(y_0 - \mathbb{E}_\tau \hat{y}_0)^2 + \mathbb{E}_\tau\left[(\mathbb{E}_\tau \hat{y}_0 - \hat{y}_0)^2\right] \middle| x_0\right] \\
&= \mathrm{Var}(y_0|x_0) + (\mathbb{E}y_0 - \mathbb{E}_\tau \hat{y}_0)^2 + \mathbb{E}\left[\mathbb{E}_\tau\left[(\mathbb{E}_\tau \hat{y}_0 - \hat{y}_0)^2\right] | x_0\right] \\
&= \mathrm{Var}(y_0|x_0) + (\mathbb{E}y_0 - \mathbb{E}_\tau \hat{y}_0)^2 + \mathbb{E}_\tau\left[(\mathbb{E}_\tau \hat{y}_0 - \hat{y}_0)^2\right]
\end{aligned}
$$

which is the second line in the derivation. The third line follows by substituting using definitions of variance and bias,

$$
EPE(x_0) = \mathrm{Var}(y_0|x_0) + \mathrm{Bias}(\hat{y}_0)^2 + \mathrm{Var}_\tau(\hat{y}_0)
$$

To find the last line, we consider each of these terms individually. It is easy to see from our model that

$$
\mathrm{Var}(y_0|x_0) = \mathrm{Var}(x_0^T\beta + \epsilon_0|x_0) = \mathrm{Var}(\epsilon_0) = \sigma^2
$$

and

$$
\mathbb{E}\hat{y}_0 = \mathbb{E}[x_0^T\hat{\beta}] = \mathbb{E}x_0^T\mathbb{E}\hat{\beta} = \mathbb{E}x_0^T\beta = \mathbb{E}y_0
$$

so $\mathrm{Bias}(\hat{y}_0)^2 = 0$. Finally, the last term can be found using the decomposition of $\hat{y}_0$ given in the book,

$$
\hat{y}_0 = x_0^T\beta + \sum_{i=1}^{N} l_i(x_0)\epsilon_i = x_0^T\beta + x_0^T(X^TX)^{-1}X^T\epsilon
$$

Then the variance is computed as

$$\begin{aligned}
\mathrm{Var}_\tau(\hat{y}_0) &= \mathrm{Var}_\tau\left(x_0^T\beta + x_0^T(X^TX)^{-1}X^T\epsilon\right) \\
&= \mathrm{Var}_\tau\left(x_0^T(X^TX)^{-1}X^T\epsilon\right) \\
&= x_0^T\,\mathrm{Var}_\tau\left((X^TX)^{-1}X^T\epsilon\right)x_0 \\
&= x_0^T\mathbb{E}_\tau\left[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}\right]x_0 \\
&= x_0^T\mathbb{E}_\tau\left[\mathbb{E}_\tau\left[(X^TX)^{-1}X^T\epsilon\epsilon^TX(X^TX)^{-1}\middle|X\right]\right]x_0 \\
&= x_0^T\mathbb{E}_\tau\left[(X^TX)^{-1}X^T(\mathbb{E}_\tau\epsilon\epsilon^T)X(X^TX)^{-1}\right]x_0 \\
&= x_0^T\mathbb{E}_\tau\left[(X^TX)^{-1}X^T\sigma^2IX(X^TX)^{-1}\right]x_0 \\
&= x_0^T\mathbb{E}_\tau(X^TX)^{-1}x_0\sigma^2 \\
&= \mathbb{E}_\tau\left[x_0^T(X^TX)^{-1}x_0\sigma^2\right]
\end{aligned}$$

Thus we have the desired result,

$$EPE(x_0) = \sigma^2 + \mathbb{E}_\tau\left[x_0^T(X^TX)^{-1}x_0\sigma^2\right]$$

■

(b) Derive equation (2.28), making use of the *cyclic* property of the trace operator [trace(AB) = trace(BA)], and its linearity (which allows us to interchange the order of trace and expectation).

Proof. Taking expectation of the result from part (a), we have

$$\mathbb{E}_{x_0}\left[EPE(x_0)\right] = \mathbb{E}_{x_0}\left[\mathbb{E}_\tau\left[x_0^T(X^TX)^{-1}x_0\right]\right]\sigma^2 + \sigma^2$$

Now use the fact that $x_0 \perp\!\!\!\perp \tau$ and $(X^TX)^{-1} \approx \mathrm{Cov}(X)^{-1}/N$, to get

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

Then writing this as an expectation of a trace (of a 1 × 1 matrix) and using the cyclic property, we get the second line of the derivation.

$$\begin{aligned}
\mathbb{E}_{x_0}\left[EPE(x_0)\right] &= \mathbb{E}_{x_0}\left[tr\left(x_0^T\mathrm{Cov}(X)^{-1}x_0\right)\right]\sigma^2/N + \sigma^2 \\
&= \mathbb{E}_{x_0}\left[tr\left(\mathrm{Cov}(X)^{-1}x_0x_0^T\right)\right]\sigma^2/N + \sigma^2 \\
&= tr\left(\mathbb{E}_{x_0}\left[\mathrm{Cov}(X)^{-1}x_0x_0^T\right]\right)\sigma^2/N + \sigma^2 \\
&= tr\left(\mathrm{Cov}(X)^{-1}\mathbb{E}_{x_0}\left[x_0x_0^T\right]\right)\sigma^2/N + \sigma^2 \\
&= tr\left(\mathrm{Cov}(X)^{-1}\mathrm{Cov}(x_0)\right)\sigma^2/N + \sigma^2
\end{aligned}$$

Now for the last line, we need only prove that

$$tr\left(\mathrm{Cov}(X)^{-1}\mathrm{Cov}(x_0)\right) = p$$

But assuming $X \overset{\mathrm{d}}{=} x_0$, we have

$$tr\left(\mathrm{Cov}(X)^{-1}\mathrm{Cov}(x_0)\right) = tr(I_p) = p$$

This gives us

$$\mathbb{E}_{x_0}\left[EPE(x_0)\right] = \sigma^2(p/N) + \sigma^2$$

which completes the proof. ■

Ex. 2.6

Consider a regression problem with inputs x_i and outputs y_i, and a parametrized model $f_\theta(x)$ to be fit by least squares. Show that if there are observations with *tied* or *identical* values of x, then the fit can be obtained from a reduced weighted least squares problem.

Proof. For notation's sake, let us think about the data falling into K groups where the inputs are equal for all data in a given group. We will denote the kth group mean of the outputs as $\overline{y}_{k\cdot}$ and size of the kth group as n_k. We begin by writing the residual sum of squares for a test model, $f_{\hat{\theta}}(x)$ and add and subtract the groups means.

$$\begin{aligned}
RSS &= \sum_{i=1}^{N}(y_i - f_{\hat{\theta}}(x_i))^2 \\
&= \sum_{i=1}^{N}(y_i - \overline{y}_{i\cdot} + \overline{y}_{i\cdot} - f_{\hat{\theta}}(x_i))^2 \\
&= \sum_{i=1}^{N}(y_i - \overline{y}_{i\cdot})^2 + 2\sum_{i=1}^{N}(y_i - \overline{y}_{i\cdot})(\overline{y}_{i\cdot} - f_{\hat{\theta}}(x_i)) + \sum_{i=1}^{N}(\overline{y}_{i\cdot} - f_{\hat{\theta}}(x_i))^2 \\
&= \sum_{i=1}^{N}(y_i - \overline{y}_{i\cdot})^2 + 2\sum_{k=1}^{K}(\overline{y}_{k\cdot} - f_{\hat{\theta}}(x_k))\sum_{i=1}^{n_k}(y_i - \overline{y}_{k\cdot}) + \sum_{k=1}^{K} n_k(\overline{y}_{k\cdot} - f_{\hat{\theta}}(x_k))^2
\end{aligned}$$

At this point, we recognize that the second term is zero since $\sum_{i=1}^{n_k}(y_i - \overline{y}_{k\cdot}) = 0,\ \forall k$. Furthermore, the first term is a constant and does not depend on $\hat{\theta}$. So in minimizing the RSS, we may ignore this term. Thus we may rewrite this as a reduced weighted least squares problem with RSS given by,

$$RSS = \sum_{k=1}^{K} n_k(\overline{y}_{k\cdot} - f_{\hat{\theta}}(x_k))^2$$

■

Ex. 2.7

Suppose we have a sample of N pairs x_i, y_i drawn i.i.d. from the distribution characterized as follows:

$x_i \sim h(x)$, the design density

$y_i = f(x_i) + \epsilon_i$, f is the regression function

$\epsilon_i \sim (0, \sigma^2)$ (mean zero, variance σ^2)

We construct an estimator for f *linear* in the y_i,

$$\hat{f}(x_0) = \sum_{i=1}^{N} l_i(x_0; \mathcal{X}) y_i,$$

where the weights $l_i(x_0; \mathcal{X})$ do not depend on the y_i, but do depend on the entire training sequence of x_i, denoted here by $\mathcal{X}$.

(a) Show that linear regression and k-nearest-neighbor regression are members of this class of estimators. Describe explicitly the weights $l_i(x_0; \mathcal{X})$ in each of these cases.

Proof. In linear regression, $f(x) = x^T\beta$ and we use the estimator,

$$\hat{f}(x_0) = x_0^T X(X^TX)^{-1}X^TY = \sum_{i=1}^{n} (X(X^TX)^{-1}X^Tx_0)_i y_i$$

so $l_i(x_0, \mathcal{X}) = (\mathcal{X}(\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^Tx_0)_i$. In k-nearest neighbor regression we use,

$$\hat{f}(x_0) = \frac{1}{k}\sum_{i=1}^{n} \mathbb{1}\left(\|x_0 - X_i\| \leqslant \|x_0 - X_{(k)}\|\right)$$

so $l_i(x_0, \mathcal{X}) = \mathbb{1}\left(\|x_0 - X_i\| \leqslant \|x_0 - X_{(k)}\|\right)$. ■

(b) Decompose the conditional mean-squared error

$$E_{\mathcal{Y}|\mathcal{X}}(f(x_0) - \hat{f}(x_0))^2$$

into a conditional squared bias and a conditional variance component. Like $\mathcal{X}$, $\mathcal{Y}$ represents the entire training sequence of y_i.

Proof. The usual trick of adding and subtracting the (conditional) expectation gives us the result.

$$\begin{aligned}
\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[(f(x_0) - \hat{f}(x_0))^2\right] &= \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\Bigg[\bigg(f(x_0) - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[f(x_0)\right] \\
&\qquad + \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[f(x_0)\right] - \hat{f}(x_0)\bigg)^2\Bigg] \\
&= \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[\left(f(x_0) - \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[f(x_0)\right]\right)^2\right] \\
&\qquad + \left(\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[f(x_0)\right] - \hat{f}(x_0)\right)^2 \\
&= \mathrm{Var}\left(f(x_0)|\mathcal{X}\right) + \mathrm{Bias}(\hat{f}(x_0)|\mathcal{X})^2
\end{aligned}$$

■

(c) Decompose the (unconditional) mean-squared error

$$E_{\mathcal{Y},\mathcal{X}}(f(x_0) - \hat{f}(x_0))^2$$

into a squared bias and a variance component.

Proof. Repeating the steps above with the unconditional expectation gives us

$$\begin{aligned}\mathbb{E}_{\mathcal{Y},\mathcal{X}}\left[(f(x_0) - \hat{f}(x_0))^2\right] &= \mathbb{E}_{\mathcal{Y},\mathcal{X}}\left[(f(x_0) - \hat{f}(x_0))^2\right] + \Big(\mathbb{E}_{\mathcal{Y},\mathcal{X}}\left[f(x_0)\right] - \hat{f}(x_0)\Big)^2 \\ &= \mathrm{Var}(f(x_0)) + \mathrm{Bias}(\hat{f}(x_0))^2\end{aligned}$$

■

(d) Establish a relationship between the squared biases and variance in the above two cases.

Proof. Suppose we have a conditional density, $f_{\mathcal{Y}|\mathcal{X}}(x, y)$ and thus a joint density, $f_{\mathcal{Y}|\mathcal{X}}(x, y)h(x)$. Now writing $\hat{f}(x_0; \mathcal{X}, \mathcal{Y})$ to make the dependence on the training data explicit, we can write the conditional and unconditional MSE as integrals.

$$\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[(f(x_0) - \hat{f}(x_0))^2\right] = \int (f(x_0) - \hat{f}(x_0; \mathcal{X}, \mathcal{Y}))^2 f_{\mathcal{Y}|\mathcal{X}}(\mathcal{X}, \mathcal{Y})d\mathcal{Y}$$

$$\mathbb{E}_{\mathcal{Y},\mathcal{X}}\left[(f(x_0) - \hat{f}(x_0))^2\right] = \int\int (f(x_0) - \hat{f}(x_0; \mathcal{X}, \mathcal{Y}))^2 f_{\mathcal{Y}|\mathcal{X}}(\mathcal{X}, \mathcal{Y})h(\mathcal{X})d\mathcal{Y}d\mathcal{X}$$

From this we can see that

$$\begin{aligned}\mathbb{E}_{\mathcal{Y},\mathcal{X}}\left[(f(x_0) - \hat{f}(x_0))^2\right] &= \int \mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[(f(x_0) - \hat{f}(x_0))^2\right] h(\mathcal{X})d\mathcal{X} \\ &= \mathbb{E}_{\mathcal{X}}\left[\mathbb{E}_{\mathcal{Y}|\mathcal{X}}\left[(f(x_0) - \hat{f}(x_0))^2\right]\right]\end{aligned}$$

Using the results from parts (b) and (c), we have

$$\mathrm{Var}(f(x_0)) + \mathrm{Bias}(\hat{f}(x_0))^2 = \mathbb{E}_{\mathcal{X}}\left[\mathrm{Var}(f(x_0)|\mathcal{X}) + \mathrm{Bias}(\hat{f}(x_0)|\mathcal{X})^2\right]$$

■

Ex. 2.8

Compare the classification performance of linear regression and k-nearest neighbor classification on the `zipcode` data. In particular, consider only the 2's and 3's, and k = 1,3,5,7 and 15. Show both the training and test error for each choice. The `zipcode` data are available from the book website `www-stat.stanford.edu/ElemStatLearn`.

Solution: First we read in the data and subset to only 2's and 3's, then subset to just the desired variables.

```
# Reading data.
train <- read.table(file.path(getwd(), "zipcode_train"))
test  <- read.table(file.path(getwd(), "zipcode_test"))
## Filtering to 2’s and 3’s and desired variables.
train <- train[train[,1] %in% c(2, 3),]
test <- test[test[,1] %in% c(2, 3),]
pixels <- c("V1", "V3", "V5", "V7", "V15")
train <- train[,pixels]
test <- test[,pixels]
```

Now we fit the two models. For comparison's sake, we will use $K = 5$. Though, we could hypothetically improve the error rate by choosing an optimal K through cross-validation.

```
# Running linear regression.
lin.mod <- lm(train[,1] ˜ ., data=train[,-1])
weighted.ave <- predict(lin.mod, test[,2:5])
pred.vals.lin <- ifelse(weighted.ave>2.5, 3, 2)
error.rate.lin <- mean(pred.vals.lin!=test[,1])

# Running K-nearest neighbors.
require(class)
pred.vals.knn <- knn(train[,2:5], test[,2:5], train[,1], k=5)
error.rate.knn <- mean(pred.vals.knn!=test[,1])
```

Comparing the two error rates, we observe

```
> print(error.rate.lin)
[1] 0.3956044
> print(error.rate.knn)
[1] 0.4038462
```

So the performances of the two are comparable, with the linear model performing slightly better than K-nearest neighbors here.

Ex. 2.9

Consider a linear regression model with p parameters, fit by least squares to a set of training data $(x_1, y_1), ..., (x_N, y_N)$ drawn at random from a population. Let $\hat{\beta}$ be the least squares estimate. Suppose we have some test data $(\tilde{x}_1, \tilde{y}_1), ..., (\tilde{x}_M, \tilde{y}_M)$ drawn at random from the same population as the training data. If $R_{tr}(\beta) = \frac{1}{N}\sum_1^N (y_i - \beta^T x_i)^2$ and $R_{te}(\beta) = \frac{1}{M}\sum_1^M (\tilde{y}_i - \beta^T \tilde{x}_i)^2$, prove that

$$E[R_{tr}(\hat{\beta})] \leqslant E[R_{te}(\hat{\beta})]$$

where the expectations are over all that is random in each expression.

Proof. To make notation simpler, we start by writing the model in matrix notation,

$$y = X\beta + \epsilon$$

which allows us to rewrite R_{tr} and R_{te} as

$$R_{tr}(\hat{\beta}) = \frac{1}{N}(y - X\hat{\beta})^T(y - X\hat{\beta}) = \frac{1}{N}\left(y^Ty - 2y^TX\hat{\beta} + \hat{\beta}^TX^TX\hat{\beta}\right)$$

$$R_{te}(\hat{\beta}) = \frac{1}{M}(\tilde{y} - \tilde{X}\hat{\beta})^T(\tilde{y} - \tilde{X}\hat{\beta}) = \frac{1}{M}\left(\tilde{y}^T\tilde{y} - 2\tilde{y}^T\tilde{X}\hat{\beta} + \hat{\beta}^T\tilde{X}^T\tilde{X}\hat{\beta}\right)$$

Now using the fact that the data are drawn independently from the same populations, we have

(i) $\mathbb{E}\left[\frac{1}{N}y^Ty\right] = \mathbb{E}\left[\frac{1}{M}\tilde{y}^T\tilde{y}\right]$

(ii) $\mathbb{E}\left[\frac{1}{N}y^TX\right] = \mathbb{E}\left[\frac{1}{M}\tilde{y}^T\tilde{X}\right]$

(iii) $\mathbb{E}\left[\frac{1}{N}X^TX\right] = \mathbb{E}\left[\frac{1}{M}\tilde{X}^T\tilde{X}\right]$

Now consider $\mathbb{E}R_{te}(\hat{\beta}) - \mathbb{E}R_{tr}(\hat{\beta})$. To simplify this, lets take this difference term by term. We can see right away that (i) shows that the first terms will cancel. Now let's think about the second terms. Using the independence of $\hat{\beta}$ and $(\tilde{X}, \tilde{y})$ and applying (ii), we have

$$-\frac{2}{M}\mathbb{E}\left[\tilde{y}^T\tilde{X}\hat{\beta}\right] = -\frac{2}{M}\mathbb{E}\left[\tilde{y}^T\tilde{X}\right]\mathbb{E}\hat{\beta} = -\frac{2}{M}\left(\frac{M}{N}\mathbb{E}[y^TX]\right)\beta = -\frac{2}{N}\beta^T\mathbb{E}[X^TX]\beta$$

Furthermore,

$$\begin{aligned}
-\tfrac{2}{N}\mathbb{E}[y^TX\hat{\beta}] &= -\tfrac{2}{N}\mathbb{E}\left[y^TX(X^TX)^{-1}X^Ty\right] \\
&= -\tfrac{2}{N}\mathbb{E}\left[(X\beta + \epsilon)^TX(X^TX)^{-1}X^T(X\beta + \epsilon)\right] \\
&= -\tfrac{2}{N}\left(\beta^T\mathbb{E}[X^TX]\beta + \mathbb{E}[\epsilon^TX(X^TX)^{-1}X^T\epsilon]\right)
\end{aligned}$$

Thus the difference between the second terms is

$$-\frac{2}{M}\mathbb{E}\left[\tilde{y}^T\tilde{X}\hat{\beta}\right] + \frac{2}{N}\mathbb{E}\left[y^TX\hat{\beta}\right] = \frac{2}{N}\mathbb{E}\left[\epsilon^TX(X^TX)^{-1}X^T\epsilon\right]$$

Now we consider the third terms. Again using independence and applying (iii),

$$\frac{1}{M}\mathbb{E}\left[\hat{\beta}^T\tilde{X}^T\tilde{X}\hat{\beta}\right] = \beta^T\mathbb{E}\left[\frac{1}{M}\tilde{X}^T\tilde{X}\right]\beta = \beta^T\mathbb{E}\left[\frac{1}{N}X^TX\right]\beta = \frac{1}{N}\beta^T\mathbb{E}[X^TX]\beta$$

Furthermore, we have

$$\begin{aligned}
\tfrac{1}{N}\mathbb{E}\left[\hat{\beta}^TX^TX\hat{\beta}\right] &= \tfrac{1}{N}\mathbb{E}\left[(X\beta + \epsilon)^TX(X^TX)^{-1}X^TX(X^TX)^{-1}X^T(X\beta + \epsilon)\right] \\
&= \tfrac{1}{N}\mathbb{E}\left[(X\beta + \epsilon)^TX(X^TX)^{-1}X^T(X\beta + \epsilon)\right] \\
&= \tfrac{1}{N}\left(\mathbb{E}\left[\beta^TX^TX\beta\right] + \mathbb{E}\left[\epsilon^TX(X^TX)^{-1}X^T\epsilon\right]\right) \\
&= \tfrac{1}{N}\left(\beta^T\mathbb{E}[X^TX]\beta + \mathbb{E}\left[\epsilon^TX(X^TX)^{-1}X^T\epsilon\right]\right)
\end{aligned}$$

Then the difference between the third terms is

$$\frac{1}{M}\mathbb{E}\left[\hat{\beta}^T\tilde{X}^T\tilde{X}\hat{\beta}\right]-\frac{1}{N}\mathbb{E}\left[\hat{\beta}^T X^T X\hat{\beta}\right]=-\frac{1}{N}\mathbb{E}\left[\epsilon^T X(X^TX)^{-1}X^T\epsilon\right]$$

Now we can combine these terms to see that

$$\begin{aligned}\mathbb{E}R_{te}(\hat{\beta})-\mathbb{E}R_{tr}(\hat{\beta})&=\tfrac{2}{N}\mathbb{E}\left[\epsilon^T X(X^TX)^{-1}X^T\epsilon\right]-\tfrac{1}{N}\mathbb{E}\left[\epsilon^T X(X^TX)^{-1}X^T\epsilon\right]\\&=\tfrac{1}{N}\mathbb{E}\left[\epsilon^T X(X^TX)^{-1}X^T\epsilon\right]\\&=\tfrac{1}{N}\mathbb{E}\left[\left((X^TX)^{-1/2}X^T\epsilon\right)^T\left((X^TX)^{-1/2}X^T\epsilon\right)\right]\\&\geqslant 0\end{aligned}$$

So we have the final result,

$$\mathbb{E}R_{te}(\hat{\beta})\geqslant\mathbb{E}R_{tr}(\hat{\beta})$$

$$\blacksquare$$