EECS 800: Homework 2 March 14, 2017

Name: Elise McEllhiney Collaborators: Bekah Coggin, David Young, David Menager

1. 4.18 Naive Bayes with mixed features
   Consider a 3-class naive Bayes classifier with one binary feature and one Gaussian feature:

$$y \sim Mu(y|\pi, 1), x_1|y = c \sim Ber(x_1|\theta_c), x_2|y = c \sim \mathcal{N}(x_2|\mu_c, \sigma_c^2)$$

   Let the parameter vectors be as follows:

$$\boldsymbol{\pi} = (0.5, 0.25, 0.25), \quad \boldsymbol{\theta} = (0.5, 0.5, 0.5), \quad \boldsymbol{\mu} = (-1, 0, 1), \quad \boldsymbol{\sigma}^2 = (1, 1, 1)$$

Solution:
$\circ \triangleq$ Hadamard Product

a. Compute $p(y|x_1 = 0, x_2 = 0)$(the result should be a vector of 3 numbers that sums to 1)

$$p(y|x_1 = 0, x_2 = 0) = \frac{\boldsymbol{\pi} \circ \boldsymbol{\theta} \circ N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2)}{\sum_j \pi_j \theta_j N(x_j|\mu_j, \sigma_j^2)} = \frac{(0.06, 0.05, 0.03)}{0.14} = (0.43, 0.36, 0.21)$$

b. Compute $p(y|x_1 = 0)$

$$p(y|x_1 = 0) = \frac{\boldsymbol{\pi} \circ \boldsymbol{\theta}}{\sum_j \pi_j \theta_j} = \frac{(0.25, 0.125, 0.125)}{0.5} = (0.5, 0.25, 0.25)$$

c. Compute $p(y|x_2 = 0)$

$$N(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

$$N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2) = (0.24, 0.40, 0.24)$$

$$p(y|x_2 = 0) = \frac{\boldsymbol{\pi} \circ N(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\sigma}^2)}{\sum_j \pi_j N(x_j|\mu_j, \sigma_j^2)} = (0.43, 0.36, 0.21)$$

d. Explain any interesting patterns you see in your results. Hint: look at the parameter vector θ

Since x_1 is independent of y we can see that it doesn't increase the amount of information we have about y. This shows in the θ vector since $(0, 1)$ are equally likely for any class. You always have a 50% chance of getting either a 0 or 1 regardless of the class.

TEACHER SOLUTION

$$p(y = c, x_1, x_2) = \pi(c)\text{Ber}(x_1|\theta_c)\mathcal{N}(x_2|\mu_c, \sigma_c^2)$$

where

$$\text{Ber}(x_1|\theta_c) = \theta_c^{I(x_1=1)}(1 - \theta_c)^{I(x_1=1)} = 0.5^{I(x_1=1)}0.5^{I(x_1=1)} = 0.5$$

So we can see that feature 1 is irrelevant, so we have $p(y|x_1, x_2) = p(y|x_2)$ and $p(y|x_1) = p(y)$. So we have

$$p(y = c, x_2 = 0) = \pi(c)\frac{1}{\sqrt{2\pi\sigma_c}} \exp\left(-\frac{1}{2\sigma_c^2}(x_2 - \mu_c)^2\right) \quad = \pi(c)\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\mu_c^2\right) \propto \pi(c) \exp\left(-\frac{1}{2}\mu_c^2\right) \tag{1}$$

We get the vector

$$p(y, x_2 = 0) = [0.5, 0.25, 0.25]. * \exp([-0.5, 0, -0.5]) \qquad = [0.3033, 0.2500, 0.1516] \tag{2}$$

and

$$p(y|x_1, x_2) = [0.3033, 0.2500, 0.1516]./0.7049 = [0.4302, 0.3547, 0.2151]$$

We find that $p(y|x_1, x_2) = p(y|x_2)$ since x_1 is uninformative. We also find $p(y|x_1) = p(y)$ for the same reason. We don't need code to solve this problem as we have shown the analytical solution above but coding can be used to confirm these solutions.

2. 7.2 Multi-output linear regression

Solution:

$$\hat{W} = (\Phi(x)^T \Phi(x))^{-1} \Phi(x)^T y$$

$$\Phi(x) = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}$$

$$\Phi(x)^T \Phi(x) = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$$

$$(\Phi(x)^T \Phi(x))^{-1} = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{bmatrix} = (1/3)\boldsymbol{I}_2$$

$$\boldsymbol{\hat{W}} = (\Phi(x)^T \Phi(x))^{-1} \Phi(x)^T y = \begin{bmatrix} \frac{1}{3} & 0 \\ 0 & \frac{1}{3} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} -1 & -1 \\ -1 & -2 \\ -2 & -1 \\ 1 & 1 \\ 1 & 2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} -\frac{4}{3} & -\frac{4}{3} \\ \frac{4}{3} & \frac{4}{3} \end{bmatrix}$$

3. 7.7 Sufficient statistics for online linear regression

$$\bar{x}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} x_i \qquad \bar{y}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$C_{xx}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \qquad C_{xy}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \qquad C_{yy}^{(n)} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2$$

Solution:

(a) What are the minimal set of statistics that we need to estimate w_0?

$C_{xx}^{(}n)$ and $C_{xy}^{(n)}$ are sufficient statistics for w_0

From equation 7.99 we can see we need covariance of X to Y and the variance of X. Thus we can show that $w_1 = C_{xy}/C_{xx}$, so C_{xy} and C_{xx} are sufficient statistics for computing w_1.

(b) What are the minimal set of statistics that we need to estimate w_1?

$\bar{x}^{(n)}$, $\bar{y}^{(n)}$, $C_{xx}^{(}n)$ and $C_{xy}^{(n)}$ are sufficient statistics w_1

From the problem statement and the rules of linear regression, we know $w_0 = \bar{y} - w_1\bar{x}$ where w_1 needs C_{xy} and $C_{xx}^{(n)}$. So we need all the statistics of the data.

(c) Suppose a new data point, x_{n+1}, y_{n+1} arrives and we want to update our sufficient statistics without looking at the old data, which we have not stored. Show that we can do this for $\bar{x}$ as follows.

$$\begin{aligned}
\bar{x}^{(n+1)} &= \frac{1}{n+1}\sum_{i=1}^{n+1} x_i \\
&= \frac{1}{n+1}(n\bar{x}^n + x_{n+1}) \qquad \text{the new sample summed with the current average current average then divided by the new total of samples. This gives equal weight to each sample} \\
&= \bar{x}^n + \frac{1}{n+1}(x_{n+1} - \bar{x}^n) \\
\bar{y}^{(n+1)} &= \frac{1}{n+1}\sum_{i=1}^{n+1} y_i \\
&= \frac{1}{n+1}(n\bar{y}^{(n)} + x_{n+1}) \\
&= \frac{1}{n+1}(y_{n+1} - \bar{y}^{(n)})
\end{aligned}$$

TEACHER SOLUTION

$$\begin{aligned}
\bar{x}^{(n+1)} &= \frac{1}{n+1}\big(n\bar{x}^{(n)} + x_{n+1}\big) \\
&= \frac{(n+1)\bar{x}^{(n)} - \bar{x}^{(n)}}{n+1} + \frac{1}{n+1}x_{n+1} \\
&= \bar{x}^{(n)} - \frac{1}{n+1}\bar{x}^{(n)} + \frac{1}{n+1}x_{n+1} \\
&= \bar{x}^{(n)} + \frac{1}{n+1}\big(x_{n+1} - \bar{x}^{(n)}\big)
\end{aligned}$$

This has the form: new estimate is old estimate plus correction. We see that the size of the correction diminishes over time (i.e., as we get more samples). Derive a similar expression to update $\bar{y}$

(d) Show that one can update C_{xy}^{n+1} recursively using

$$C_{xy}^{(n+1)} = \frac{1}{n+1}[x_{n+1}y_{n+1} + nC_{xy}^{(n)} + n\bar{x}^{(n)}\bar{y}^{(n)} - (n+1)\bar{x}^{n+1}\bar{y}^{n+1}]$$

Let $n = n + 1$:

$$
\begin{aligned}
C_{xy}^{(n+1)} &= \frac{1}{n+1}\sum_{i=1}^{n+1}(x_i - \bar{x}^{(n+1)})(y_i - \bar{y}^{(n+1)}) \\
&= \frac{1}{n+1}\Big(\sum_{i=1}^{n+1} x_i y_i - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}\Big) \\
&= \frac{1}{n+1}\Big(x_{(n+1)}y_{(n+1)} + \sum_{i=1}^{n} x_i y_i - n\bar{x}^{(n)}\bar{y}^{(n)} + n\bar{x}^{(n)}\bar{y}^{(n)} - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}\Big) \\
&= \frac{1}{n+1}\Big(x_{(n+1)}y_{(n+1)} + C_{xy}^{(n)} + n\bar{x}^{(n)}\bar{y}^{(n)} - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}\Big)
\end{aligned}
$$

Derive a similar expression to update C_{xx}

$$
\begin{aligned}
C_{xx}^{(n+1)} &= \frac{1}{n+1}\sum_{i=1}^{n+1}(x_i - \bar{x}^{(n+1)})^2 \\
&= \frac{1}{n+1}\Big(\sum_{i=1}^{n+1} x_i x_i - (n+1)\bar{x}^{(n+1)}\bar{x}^{(n+1)}\Big) \\
&= \frac{1}{n+1}\Big(x_{(n+1)}x_{(n+1)} + \sum_{i=1}^{n} x_i^2 - n\bar{x}^{(n)}\bar{x}^{(n)} + n\bar{x}^{(n)}\bar{x}^{(n)} - (n+1)\bar{x}^{(n+1)}\bar{x}^{(n+1)}\Big) \\
&= \frac{1}{n+1}\Big(x_{(n+1)}x_{(n+1)} + C_{xx}^{(n)} + n\bar{x}^{(n)}\bar{x}^{(n)} - (n+1)\bar{x}^{(n+1)}\bar{x}^{(n+1)}\Big)
\end{aligned}
$$

TEACHER SOLUTION Let $n = n + 1$:

$$
\begin{aligned}
C_{xy}^{(n)} &= \frac{1}{n}\left[\Big(\sum_{i=1}^{n} x_i y_i\Big) + \Big(\sum_{i=1}^{n} \bar{x}^{(n)}\bar{y}^{(n)}\Big) - \bar{x}^{(n)}\Big(\sum_{i=1}^{n} y_i\Big) - \bar{y}^{(n)}\Big(\sum_{i=1}^{n} x_i\Big)\right] \\
&= \frac{1}{n}\left[\Big(\sum_{i=1}^{n} x_i y_i\Big) + n\bar{x}^{(n)}\bar{y}^{(n)} - \bar{x}^{(n)}n\bar{y}^{(n)} - \bar{y}^{(n)}n\bar{x}^{(n)}\right] \\
&= \frac{1}{n}\left[\Big(\sum_{i=1}^{n} x_i y_i\Big) - n\bar{x}^{(n)}\bar{y}^{(n)}\right]
\end{aligned}
$$

Hence

$$\sum_{i=1}^{n} x_i y_i = nC_{xy}^{(n)} + n\bar{x}^{(n)}\bar{y}^{(n)}$$

and

$$C_{xy}^{(n+1)} = \frac{1}{n+1}\big[x_{n+1}y_{n+1} + nC_{xy}^{(n)} + n\bar{x}^{(n)}\bar{y}^{(n)} - (n+1)\bar{x}^{(n+1)}\bar{y}^{(n+1)}$$

4. Additional Question 1.

Compute the numbers of parameters that need to be estimated for the models listed below.

Solution:
(a) Naive Bayes (categorical input variables)

$$pc + c - 1 \quad \text{parameters are needed for this model}$$

$$MLE = \theta_{jc}\pi_c$$

We also know that $\sum \pi_c = 1$
Since we don't know that about the θ value as the theta values do not necessarily sum to one because they are dependent on the class that they are associated with, we know that we need $pc + c - 1$ parameters for this model

(b) Quadratic Discriminant Analysis (QDA)

$$\Big(\frac{p(p+1)}{2} + p\Big)c \quad \text{parameters are needed for this model}$$

We need the covariance matrix for sigma which is symmetric so we need the variance values and the half of the covariance values (since the matrix is symmetric we only need half the covariance values) then we need μ. We need these values for each class, and hence the above conclusion.

(c) Linear Discriminant Analysis (LDA)

$$\frac{p(p+1)}{2} + pc \quad \text{parameters are needed for this model}$$

We need the same covariance matrix for this model, but we assume the Σ parameter is the same for all the classes so we only need one covariance p x p matrix. We still need

(d) Diagonalized QDA (e.g. covariance(Xi, Xj) = 0 given i 6= j)

$$(p + p)c \quad \text{parameters are needed for this model}$$

Again, a more limited number of parameters than the QDA since we only need the diagonal of the covariance matrix for the parameters. Again we still need a μ value for each feature.

(e) Diagonalized LDA (e.g. covariance(Xi, Xj) = 0 given i 6= j)

$$p + pc \quad \text{parameters are needed for this model}$$

Again, we share the sigma so we need a weight for each feature and a μ for each class

(f) Linear Regression

$$p + 1 \quad \text{parameters are needed for this model}$$

Which is the size of β and this is equal to the number of features, the +1 is for β_0

(g) Ridge Regression

$$p + 2 \quad \text{parameters are needed for this problem}$$

P is the size of β and this is equal to the number of features. One +1 is for the β_0 parameter and the other +1 is the parameter estimating λ

TEACHER SOLUTION

- Naive Bayes with binary input features. We need to know the prior probabilities of each class $\pi_c = p(y = c)$ as well as each $\theta_{jc} = p(x_j = 1|y = c)$. We also know that to fully specify the probability of each class, we need only know $c - 1$ probabilities as the class probabilities are constrained to sum to 1. So we know that there are $c - 1$ parameters coming from $\boldsymbol{\pi}$ and $\boldsymbol{\theta}$ includes $p * c$ parameters. In total this model requires

  $$pc + c - 1$$

  parameters.

- For QDA we again need to know π_c for each class, requiring $(c - 1)$ parameters (as they sum to 1). We further note that we assume each class is a multivariate normal that is characterized by $\boldsymbol{\mu}$ which is of length p and a covariance matrix $\boldsymbol{\Sigma}$ which is of size $p \times p$. But this covariance matrix is symmetric so we need only to estimate the diagonal and the upper triangle. In total, for each class, we must estimate $\frac{p^2+p}{2}$ to fully specify $\boldsymbol{\Sigma}$. Now we need to estimate both $\boldsymbol{\mu}_c$ and $\boldsymbol{\Sigma}_c$ for each class c. These are fully independent given class. Thus we can conclude we need pc for all $\boldsymbol{\mu}_c$s and $\frac{p^2+p}{2}c$ for all $\boldsymbol{\Sigma}_c$s. Then we also need $c - 1$ parameters for each class. In total we need

  $$pc + \frac{p^2 + p}{2}c + c - 1$$

- LDA: This is identical to QDA, except we assume that the covariances are shared across classes. So we need

  $$pc + \frac{p^2 + p}{2} + c - 1$$

- Diagonalized QDA: If we have a diagonalized QDA we will need to know a total of p values in our covariance matrix for each class (e.g. the diagonal), leading to a total of cp estimates. We additionally have our means to estimate as above and our π_c estimates of each class. Our total is thus:

  $$pc + pc + c - 1$$

- Diagonalized LDA: If we have a diagonalized LDA then we have the same model as above but instead of a different estimate of variance for each class, we assume that a single diagonalized covariance matrix is used for all classes so we go from cp estimates related to variance to just p. Note this is similar to Naive Bayes but X can now have continuous values and thus we need to estimate the mean and variance for each class and each feature.

  $$pc + p + c - 1$$

- Linear regression: We need only know the β estimate for each parameter plus the intercept term (β_0), so we need a total of:

  $$p + 1$$

- Ridge regression: We certainly need to fit the λ value so that adds an extra parameter above our linear regression model, so

  $$p + 1 + 1$$

  . Note that λ is regularization term but we still need to estimate it to fully specify our model.

  Note, we have to normalize the input and center the output but this requires us to compute statistics, not infer parameters and thus does not increase the parameter requirement.

5. Additional Question 2.
(bonus) Show that the ridge regression estimate is the mean (and mode) of the posterior distribution, under a Gaussian prior $\beta \sim N(0, \tau \boldsymbol{I})$ where I is the identity matrix, and Gaussian sampling model $y \sim N(\boldsymbol{X}\beta, \sigma^2 \boldsymbol{I})$. Find the relationship between the regularization parameter λ in the ridge formula and the variances τ and σ^2.

Solution:

6. Additional Question 3.

Solution:
(a) Derive the OLS linear regression estimate.

$$\min_{\hat{\beta}_0, \hat{\beta}_1} \sum_{i=1}^{N} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2$$

$$\frac{\partial W}{\partial \hat{\beta}_0} = \sum_{i=1}^{N} -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial W}{\partial \hat{\beta}_1} = \sum_{i=1}^{N} -2x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\frac{\partial W}{\partial \hat{\beta}_0} = \sum_{i=1}^{N} -2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1) = 0$$

$$\sum_{i=1}^{N} y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1 = 0$$

$$N\hat{\beta}_0 = N\bar{y} - N\hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\frac{\partial W}{\partial \hat{\beta}_1} = \sum_{i=1}^{N} -2x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_1) = 0$$

$$\sum_{i=1}^{N} x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2) = 0$$

$$\sum_{i=1}^{N} x_i y_i - (\bar{y} - \hat{\beta}_1 \bar{x}) x_i - \hat{\beta}_1 x_i^2) = 0$$

$$\sum_{i=1}^{N} x_i y_i - \bar{y} \sum_{i=1}^{N} x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^{N} x_i - \hat{\beta}_1 \sum_{i=1}^{N} x_i^2 = 0$$

$$\sum_{i=1}^{N} y_i = N\bar{y}$$

$$\sum_{i=1}^{N} x_i = N\bar{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i\bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

TEACHER SOLUTION
First, note that the trace of a matrix is the sum of the main diagonal entries. This measure has the property such that if AB is a square matrix, then $\mathrm{tr}AB = \mathrm{tr}BA$. Corollaries of this are that $\mathrm{tr}ABC = \mathrm{tr}BCA = \mathrm{tr}CAB$. With a bit more linear algebra, we can show that

$$\bigtriangledown_A \mathrm{tr}AB = B^T \bigtriangledown_{A^T} f(A) \qquad = (\bigtriangledown_A f(A))^T \bigtriangledown_A ABA^TC = CAB + C^TAB^T \tag{3}$$

We can now show that

$$\bigtriangledown_{A^T} \mathrm{tr}ABA^TC = B^TA^TC^T + BA^TC$$

Now define our loss function as

$$J(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{X\theta} - \vec{y})^T(\boldsymbol{X\theta} - \vec{y}) \tag{4}$$

$$\bigtriangledown_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \bigtriangledown_{\boldsymbol{\theta}} \frac{1}{2}(\boldsymbol{X\theta} - \vec{y})^T(\boldsymbol{X\theta} - \vec{y}) \tag{5}$$

$$= \frac{1}{2} \bigtriangledown_{\boldsymbol{\theta}} (\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X\theta} - \boldsymbol{\theta}^T\boldsymbol{X}^T\vec{y} - \vec{y}^T\boldsymbol{X\theta} + \vec{y}^T\vec{y}) \tag{6}$$

$$= \frac{1}{2} \bigtriangledown_{\boldsymbol{\theta}} \mathrm{tr}(\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X\theta} - \boldsymbol{\theta}^T\boldsymbol{X}^T\vec{y} - \vec{y}^T\boldsymbol{X\theta} + \vec{y}^T\vec{y}) \tag{7}$$

$$= \frac{1}{2} \bigtriangledown_{\boldsymbol{\theta}} (\mathrm{tr}\boldsymbol{\theta}^T\boldsymbol{X}^T\boldsymbol{X\theta} - 2\mathrm{tr}\vec{y}^T\boldsymbol{X\theta}) \tag{8}$$

$$= \frac{1}{2}(\boldsymbol{X}^T\boldsymbol{X\theta} + \boldsymbol{X}^T\boldsymbol{X\theta} - 2\boldsymbol{X}^T\vec{y}) \tag{9}$$

$$= \boldsymbol{X}^T\boldsymbol{X\theta} - \boldsymbol{X}^T\vec{y} \tag{10}$$

setting its derivatives to zero we obtain

$$\boldsymbol{X}^T\boldsymbol{X\theta} = \boldsymbol{X}^T\vec{y}$$

The value of $\boldsymbol{\theta}$ that minimizes $J(\boldsymbol{\theta})$ is given by

$$\boldsymbol{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\vec{y}$$

(b) Derive the Ridge regression estimate.

$$
\begin{aligned}
RSS &= ||y - \hat{y}||_2^2 + \lambda||\beta||_2^2 \\
&= (Y - X\beta)^2 + \lambda(\beta)^2 \\
&= (Y - X\beta)^T(Y - X\beta) + \lambda(\beta)^T\beta \\
&= (Y^T - \beta^T X^T)(Y - X\beta) + \lambda(\beta^T\beta) \\
&= Y^TY - \beta^T X^T Y + Y^T X\beta + \beta^T X^T X\beta + \lambda(\beta^T\beta)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial\beta^T} &= 0 - X^TY + 0 + X^TX\beta + \lambda\beta = 0 \\
&= X^TY + X^TX\beta + \lambda\beta = 0
\end{aligned}
$$

$$
\begin{aligned}
X^TY &= X^TX\beta + \lambda\beta \\
&= \beta(X^TX + \lambda I)
\end{aligned}
$$

$$
\hat{\beta} = (X^TX + \lambda I)^{-1}(X^TY)
$$

TEACHER SOLUTION

$$
\begin{aligned}
RSS &= ||y - \hat{y}||_2^2 + \lambda||\beta||_2^2 \\
&= (Y - X\beta)^2 + \lambda(\beta)^2 \\
&= (Y - X\beta)^T(Y - X\beta) + \lambda(\beta)^T\beta \\
&= (Y^T - \beta^T X^T)(Y - X\beta) + \lambda(\beta^T\beta) \\
&= Y^TY - \beta^T X^T Y + Y^T X\beta + \beta^T X^T X\beta + \lambda(\beta^T\beta)
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial}{\partial\beta^T} &= 0 - X^TY + 0 + X^TX\beta + \lambda\beta = 0 \\
&= X^TY + X^TX\beta + \lambda\beta = 0
\end{aligned}
$$

$$
\begin{aligned}
X^TY &= X^TX\beta + \lambda\beta \\
&= \beta(X^TX + \lambda I)
\end{aligned}
$$

$$
\hat{\beta} = (X^TX + \lambda I)^{-1}(X^TY)
$$

(c) Show that the estimation of β_λ^{ridge} is a biased estimator. [Hint: compute the expected value of

the ridge estimate and show that it is not equal to the expected value of the OLS estimate.]

$$\begin{aligned}\hat{\beta}_{\lambda}^{ridge} &= (X^T X + \lambda I)^{-1}(X^T y) \\ &= (X^T X + \lambda I_p)^{-1}(X^T X)(X^T X)^{-1} X^T y \\ &= (X^T X(I + (X^T X)^{-1}))^{-1}(X^T X)\hat{\beta}^{OLS} \\ &= (I + (X^T X)^{-1}))^{-1}\hat{\beta}^{OLS}\end{aligned}$$

$$(I + (X^T X)^{-1}))^{-1}\hat{\beta}^{OLS} \neq \hat{\beta}^{OLS} \quad \text{unless } \lambda = 0$$

TEACHER SOLUTION

$$\begin{aligned}\hat{\beta}_{\lambda}^{ridge} &= (X^T X + \lambda I)^{-1}(X^T y) \\ &= (X^T X + \lambda I_p)^{-1}(X^T X)(X^T X)^{-1} X^T y \\ &= (X^T X(I + \lambda(X^T X)^{-1}))^{-1}(X^T X)\hat{\beta}^{OLS} \\ &= (I + \lambda(X^T X)^{-1}))^{-1}\hat{\beta}^{OLS}\end{aligned}$$

$$(I + \lambda(X^T X)^{-1}))^{-1}\hat{\beta}^{OLS} \neq \hat{\beta}^{OLS} \quad \text{unless } \lambda = 0$$