COMS 4721 – HW3

Daniel Kost-Stephenson – dpk2124

April 5th, 2016

Question 1

a)

In the first part of this question, we are asked to minimize the loss function of a binary random variable Y, that can take on values $\{-1, +1\}$. The loss function is exponential and is given as:

$$f(z) = E\left[l_{exp}(Yz)\right] = E[e^{Yz}]$$

Since $P(Y = 1) = \mu$ and $P(Y = -1) = 1 - \mu$, we can write $f(z)$ as

$$f(z) = e^{-z}\mu + e^{z}(1 - \mu)$$

By taking the derivative and setting it equal to zero, we can obtain an expression that minimizes $f(z)$:

$$\frac{\partial f(z)}{\partial z} = -e^{-z}\mu + e^{z}(1 - \mu) = 0$$

$$-\mu + e^{2z}(1 - \mu) = 0$$

$$2z = \log\left(\frac{\mu}{1 - \mu}\right)$$

$$z = \frac{1}{2}\log\left(\frac{\mu}{1 - \mu}\right)$$

b)

For the second part of the question, we consider Y as a continuous random variable that can take on real values instead of the binary case in part a). The loss function is given as:

$$f(z) = E[|Y - z|]$$

Since this is continuous random variable, we can express $E[|Y - z|]$ as

$$f(z) = \int_{-\infty}^{\infty} |Y - z| \cdot f(y)\, dy$$

By splitting the absolute value, we obtain:

$$f(z) = \int_{-\infty}^{z} -(Y - z) \cdot f(y)\, dy + \int_{z}^{\infty} (Y - z) \cdot f(y)\, dy$$

$$f(z) = -\int_{-\infty}^{z} Y \cdot f(y)\, dy + \int_{-\infty}^{z} z \cdot f(y)\, dy + \int_{z}^{\infty} Y \cdot f(y)\, dy - \int_{z}^{\infty} z \cdot f(y)\, dy$$

$$f(z) = -\int_{-\infty}^{z} Y \cdot f(y)\, dy + \int_{z}^{\infty} Y \cdot f(y)\, dy + z\Pr(Y < z) - z\Pr(Y > z)$$

Again taking the derivative and setting it equal to zero, we obtain:

$$\frac{\partial f(z)}{\partial z} = 0 + 0 + \Pr(Y < z) - \Pr(Y > z) = 0$$

$$\Pr(Y < z) - (1 - \Pr(Y < z))$$

$$2\Pr(Y < z) = 1$$

$$\Pr(Y < z) = \frac{1}{2}$$

Similarly, $\Pr(Y > z) = \frac{1}{2}$ and thus the minimizer of the absolute loss function is the *median*.

Question 2

a)

For part a) we are asked to find the relation between Q_j and P. Since our method of reduction is to define $z_{i,j} = \mathbb{1}\left\{y_i \geq \frac{j}{m}\right\}$, we are essentially verifying if the input, y_i, is greater than a certain fraction $\frac{j}{m}$ which will always be increasing by factors of $\frac{1}{16}$. The procedure is to map Q_j as 1 if y_i is greater than $\frac{j}{m}$ and 0 otherwise. Thus, we are essentially setting a cut off for the value of y_i and the points at which 1s switch to 0s are the points that bound y_i. For example, if $y_i = \frac{7}{32}$ then the Q_j values would look like :

$$\{1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0\}$$

Because $\frac{3}{16} < y_i < \frac{4}{16}$. Thus, we are essentially creating "bins" $\left(Q_j\right)$ which bound the real values of P.

b)

We are now given that $\Pr(X = a) = \Pr(X = b) = \frac{1}{2}$ and that the conditional distributions of a and b respectively are $\left[\frac{1}{8}, \frac{3}{8}\right]$ and $\left[\frac{5}{8}, \frac{7}{8}\right]$. Since we are only given if the label is a or b without knowing the exact numeric value, we should minimize our loss for each class by always assuming that an observation with label a or b will behave like its class conditional expectation: $E[Y|X = a] = \frac{2}{8} = \frac{4}{16}$ and $E[Y|X = b] = \frac{6}{8} = \frac{12}{16}$. Since we can only predict a numeric value, it makes sense to predict the mean of each class. The actual optimal classifier can be specified as follows:

$$if\ \frac{j}{m} \leq \frac{4}{16}, set\ Q_j = 1\ for\ both\ a\ and\ b$$

$$if\ \frac{5}{16} \leq \frac{j}{m} \leq \frac{12}{16}, set\ Q_j = 0\ for\ a\ and\ Q_j = 1\ for\ b$$

$$if\ \frac{13}{16} \leq \frac{j}{m} \leq \frac{16}{16}, set\ Q_j = 0\ for\ both\ a\ and\ b$$

c)

As mentioned above, the resulting real-valued predictor is simply the conditional expectation for each class, so $E[Y|X=a]=\frac{2}{8}=\frac{4}{16}$ and $E[Y|X=b]=\frac{6}{8}=\frac{12}{16}$. The expected absolute loss is simply $\frac{1}{16}$ because on average, that's how much the actual value of a or b deviates from the expectation. Since a or b come from uniform distributions, the range above and below the means $\left(\frac{4}{16}\text{and}\frac{12}{16}\right)$ are uniformly distributed along $U\sim\left[\frac{1}{8},\frac{2}{8}\right]$ and $U\sim\left[\frac{2}{8},\frac{3}{8}\right]$ for a and $U\sim\left[\frac{4}{8},\frac{5}{8}\right]$ and $U\sim\left[\frac{5}{8},\frac{6}{8}\right]$ for b. Thus, the expected loss is: $E[loss]=E[|X-E[X]|]=\frac{1}{16}$.

Question 3

a)

The OLS estimator was computed after removing the first column from the training set and test set. The intercept term, ω_0, is given by $\omega_0=\bar{y}-\vec{\boldsymbol{\omega}}\cdot\bar{x}$ and since the data has already been centred and scaled, we should expect the intercept term to be the mean of the y values in the training set. Indeed, when the linear model was fit to the Boston housing data set, the intercept term was 22.575 and the mean of the y values was also 22.575.

b)

The average squared loss (or mean squared error) is given by:

$$E[(Y-\langle\boldsymbol{\omega},\boldsymbol{X}\rangle)^2]$$

Which resulted in an average squared loss of 22.104 on the training set, and 24.407 on the test set.

c)

To calculate the sparse weight vector, the Lasso technique was used. The Lasso is a technique that adds a regularization term to the OLS formulation which forces some coefficients to converge to zero. The Lasso is given by:

$$\omega\rightarrow argmin\ \|\boldsymbol{Ax}-\boldsymbol{b}\|_2^2+\lambda\|\boldsymbol{\omega}\|_1$$

In this example, a penalty (λ) of 2.5 was used to keep only three non-zero terms in the weight vector. Here, the variables RM, PTRATIO and LSTAT were the three that were kept. These variables make sense: the average number of rooms per dwelling definitely influences how valuable a house is, as does the pupil-teacher ratio and % lower status of the population. By simply knowing what these three variables, we would intuitively expect them to increase or decrease with the median value of occupied homes. However, I am a bit surprised that some of the other variables such as crime rate and property tax rate, weren't part of the three.