COMS 4772 Fall 2015: Homework #1

Name Surname - uni@columbia.edu
Discussants: none

April 3, 2016

Problem 1

Problem 2

(a) Q_j is a Bernoulli random variable with parameter p equal to $P(y|x \geq j/n)$. That is, the jth Q is a Bernoulli distribution whose probability depends on the random variable P, as P is a joint random variable of X and Y. The two are connected because Q_j is a function of X and Z and Z is a function of Y, J, and M.

(b) The optimal binary classifier for each j is function of j and the random variables X and Y. Up to (but not including) $j = 5$, Y (regardless of X) is greater than or equal to j/m, so the optimal classifier should return 1. From $j = 5$ to 12 optimality depends on X. For $X = a$, Y is less than j/m and therefore the optimal predictor is 0. All values of $Y = b$ in $j = 5to12$ are greater than j/m so the optimal predictor is 1. For $j \geq 13$ again Y values for both X values are less than j/m so the optimal classifier is 0.

More tersely:
For $j_{1,4}$, f_j^* is:

$$f_j^*(x) = 1$$

For $j_{5,12}$ it is:

$$\begin{cases} f_j^*(x = a) = 0 \\ f_j^*(x = b) = 1 \end{cases}$$

For $j_{13,16}$ it is:

$$f_j^*(x) = 0$$

(c) The optimal classifier minimizes $\mathbb{E}[|Y - g(X)|]$. To find this, we need to determine the expected value of $g(x)$. Because $g(X)$ depends on the the value of X, the overall expected value will be the sum of the two cases weighted by their probabilities. We'd have:

$$\mathbb{E}[|Y - g(X)|] = \frac{1}{2}\mathbb{E}[|Yg(X)||x = a] + \frac{1}{2}\mathbb{E}[|Yg(X)||x = b]$$

The expected values of $g(x)$ for $X = a$ and $X = b$ are .25 and .75 respectively. This means the expected absolute loss is:

$$\mathbb{E}[|Y - g(X)|] = \frac{1}{2}\mathbb{E}[|Y.25||x = a] + \frac{1}{2}\mathbb{E}[|Y.75||x = b] = \frac{1}{16}$$

Problem 3

(a) Using the LinearRegression class from the scikit-learn package I got an intercept of 25.57470356. This number is equal to the mean of the training outcomes, which makes sense given the data has been standardized. In effect, the standardization process has shifted the data so that mean of the output is centered around zero. Without considering other factors you'd expect the value of a house to be the expected value of the dataset.

(b) The training MSE was 22.1037987797. The testing MSE 24.4065641284

(c) I used three models from the scikit-learn linear models package: Lasso, Lars, and Orthogonal Matching Pursuit. The selected columns were as follows: Lasso: 'RM', 'PTRATIO', 'LSTAT' Lars: 'RM', 'PTRATIO', 'LSTAT' OMP: 'ZN', 'INDUS', 'PTRATIO'