COM 525000 Statistical Learning
Homework #1

(Due October 24, 2019 at the beginning of class)

Note: Detailed derivations are required to obtain a full score for each problem. (Total 100%)

1. (8%+4%) Recall that, for data points $\mathbf{y} = (y_1, \ldots, y_n)^T$ and $\mathbf{X} = (x_1, \ldots, x_n)^T$ where $\mathbf{X}^T\mathbf{X}$ is nonsingular, the least squares solution for linear regression with p predictors is given by

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

The fitted value for $\mathbf{y}$ is thus given by $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$. The standard error for the j-th coefficient estimate is

$$\mathrm{SE}(\hat{\beta}_j) = \sqrt{\{\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\}_{jj}}.$$

(a) Show that the standard error expression reduces to Eq. (3.4) of the textbook in the single predictor case (i.e., when $p = 1$).

(b) Show that $\mathbf{H} \triangleq \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is a projection matrix that projects a vector onto the subspace spanned by the columns of $\mathbf{X}$. That is, for any vector $\mathbf{z}$ that is a linear combination of the columns of $\mathbf{X}$ (i.e., $\mathbf{z} = \mathbf{X}\mathbf{c}$, for some $\mathbf{c} = (c_0, \ldots, c_p)^T$), $\mathbf{H}\mathbf{z} = \mathbf{z}$. (Comment: Consequently, $\hat{\mathbf{y}}$ is the projection of $\mathbf{y}$ onto the subspace spanned by the columns of $\mathbf{X}$.)

2. (8%+8%+4%) In linear regression, we adopt the linear model

$$Y = X^T\beta + \epsilon,$$

where $X = (1, X_1, \ldots, X_p)^T$, $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Let

$$\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\} \Big(\text{or concisely written as } \{\mathbf{X}, \mathbf{y}\}\Big)$$

be the set of available data points that are generated independently by the above model.

(a) Find β that maximizes the likelihood function $p(\mathbf{y}|\mathbf{X}, \beta)$, assuming that σ^2 is known.

(b) Now suppose that the entries of β are i.i.d. $\mathcal{N}(0, \gamma^2)$. Find β that maximizes the posterior probability

$$p(\beta|\mathbf{y}, \mathbf{X}).$$

(c) Comment on the similarities and differences between least squares linear regression and the above schemes.

3. (10%+10%+8%) Suppose that there were 200 coupons for each of the discount percentages 5%, 10%, 15%, 20%, and 30% (i.e., the input values x_i, for $i = 1, \ldots, 1000$), and that the number of coupons redeemed for the above cases are 33, 48, 62, 99, and 132, respectively (i.e., the number of responses in each case that yields $y_i = 1$).

(a) Fit a simple linear regression to the observed proportions 33/200, 48/200, 62/200, 99/200, and 132/200. List the fitted values for the above discount values. According to this regression, at what redemption rate will you get for a 25% price reduction?

(b) Find the 95%-confidence interval of the estimated coefficients (by treating estimates of the standard error as the true standard errors).

(c) Repeat (a) using logistic regression. (Hint: Just describe the procedure and derivations. No need to compute the actual coefficient values.)

4. (10%) Show that the F-statistic in eq. (3.24) for dropping a single coefficient from a model is equal to the square of the t-score defined as

$$t = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2\{(\mathbf{X}^T\mathbf{X})^{-1}\}_{jj}}}.$$

5. (2%+2%+2%+6%) Problem 4 (a),(b),(c), (e) in Chapter 4 of the textbook.

6. (2%+8%+4%+4%) Let us consider the simple linear regression problem, where the data points $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ are fit to a linear model $f(X) = \beta_0 + \beta_1 X$.

(a) Show that the least squares solution yields estimates

$$\hat{\beta}_1^{(n)} = \frac{C_{xy}^{(n)}}{C_{xx}^{(n)}} \quad \text{and} \quad \hat{\beta}_0^{(n)} = \bar{y}^{(n)} - \hat{\beta}_1^{(n)}\bar{x}^{(n)},$$

where $\bar{x}^{(n)} = \frac{1}{n}\sum_{i=1}^n x_i$, $\bar{y}^{(n)} = \frac{1}{n}\sum_{i=1}^n y_i$, $C_{xx}^{(n)} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}^{(n)})^2$, $C_{xy}^{(n)} = \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x}^{(n)})(y_i - \bar{y}^{(n)})$. (Hint: You can use results from Problem 1.)

(b) Suppose that a new data point (x_{n+1}, y_{n+1}) arrives. Show that the sample mean and sample covariance matrices can be updated as

$$\bar{x}^{(n+1)} = \bar{x}^{(n)} + \frac{1}{n+1}(x_{n+1} - \bar{x}^{(n)})$$

and

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

(c) Show that a similar update can be derived for $\bar{y}^{(n+1)}$ and $C_{xx}^{(n+1)}$.

(d) Describe the implications of the above derivations in terms of online computation of linear regression.