COM 599200 Statistical Learning
Homework #1

(Due April 2, 2018 at the beginning of class)

Note: Detailed derivations are required to obtain a full score for each problem. (Total 80%)

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 31, 52, 68, 101, and 144, 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 31/200, 52/200, 68/200, 101/200, and 144/200. 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) Repeat (a) using logistic regression. (Hint: Just describe the procedure and derivations. No need to compute the actual coefficient values.)

(c) Repeat (a) using linear discriminant analysis. (Hint: You will need to use the estimated normal distribution to compute $\Pr(Y = 1|X = x)$.)

4. (8%) Problem 7 in Chapter 3 of the textbook.

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