COM 525000 Statistical Learning
Homework #2

(Due November 12, 2019 noon to the TA at EECS 613)

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

1. (8%+14%)

(a) To find the coefficients for the logistic regression in Problem 3(c) of HW #1, we adopt the gradient descent approach using a fixed step-size $\eta = 0.1$. Find the update $\beta(k+1)$ when the values in the current iteration is $\beta(k) = (\beta_0(k), \beta_1(k)) = (10, 4)$ and $(1, 1)$, respectively.

(b) Solve Problem 3(a) of HW #1 using LDA and QDA, respectively.

2. (14%) Consider the data set $\mathcal{D} = \{((x_{11}, x_{12}), y_1), \cdots, ((x_{61}, x_{62}), y_6)\} = \{((1.5, 2), 1), ((1, 1), 1), ((2, 0.5), 1), ((-2, 0), 0), ((-1, 0), 0), ((-2, -1), 0)\}$. Find the classification rule using QDA, and determine the prediction for new input $(0, 0)$.

3. (10%) Solve Problem 7 of Chapter 4, but with the observed variance being $\hat{\sigma}^2 = 25$ for those that issued a dividend and $\hat{\sigma}^2 = 36$ for those that didn't.

4. (14%) Let

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i; \mathcal{D} \setminus \{(x_i, y_i)\})\right)^2$$

be the leave-one-out cross-validation (LOOCV) error. Show that

$$\mathrm{CV}_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

where $\hat{y}_i = \hat{f}(x_i; \mathcal{D})$ is the i-th fitted value from the original least square fit (using the entire data set $\mathcal{D}$), and h_i is the leverage statistic.
(Hint: Fill in the details of the sketch proof shown in class.)

5. (8%+10%+14%+8%) Suppose that the available data set is $\mathcal{D} = \{(x_1, y_1), (x_2, y_2), (x_3, y_3)\} = \{(3, 7), (5, 8), (10, 15)\}$. Linear regression is to be performed on the above data.

(a) Find the training error (defined by the squared loss) when using the entire set to perform the model fit.

(b) Find the error estimate using leave-one-out cross-validation.

(c) Suppose that $B = 3$ bootstrap datasets are obtained as $\mathcal{D}^{*1} = \{(3,7),(5,8),(5,8)\}$, $\mathcal{D}^{*2} = \{(5,8),(10,15),(10,15)\}$, and $\mathcal{D}^{*3} = \{(3,7),(3,7),(10,15)\}$. Find the coefficient estimates $\hat{\beta}^{*1}, \hat{\beta}^{*2}, \hat{\beta}^{*3}$ obtained from each dataset, and compute the leave-one-out bootstrap error estimate

$$\widehat{\mathrm{Err}}_{\mathrm{boot}} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{|\mathcal{C}^{(-i)}|}\sum_{b\in\mathcal{C}^{(-i)}}\|y_i - \hat{\beta}_0^{*b} - \hat{\beta}_1^{*b}x_i\|^2$$

where $\mathcal{C}^{(-i)}$ is the set of indices of the bootstrap datasets that do not contain sample i. (In our example, $n = 3$ and $|\mathcal{C}^{(-i)}|$ is only 2 for all i.)

(d) Following (c), find the standard errors of the coefficient estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, and compare with the standard error estimates

$$\widehat{\mathrm{SE}}(\hat{\beta}_0)^2 = \hat{\sigma}^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]$$

and

$$\widehat{\mathrm{SE}}(\hat{\beta}_1)^2 = \frac{\hat{\sigma}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

where $n = 3$ in this case.