COM 525000 Statistical Learning
Homework #2

(Due November 17, 2020 noon to the TA at EECS 613)

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

1. (8%+16%) Suppose that Elon Musk captures 7 aliens from outer space, of which 3 are from Mars ($Y = 1$) and 4 are from Krypton ($Y = 0$). The weight and height of the 3 aliens from Mars are $(50, 164)$, $(60, 140)$, and $(66, 152)$; and that of the 4 aliens from Krypton are $(46, 148)$, $(40, 160)$, $(48, 180)$, and $(65, 172)$.

(a) To obtain a classifier using logistic regression, 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 $\beta(k) = (\beta_0(k), \beta_1(k), \beta_2(k)) = (4, 2, 3)$, respectively.

(b) Find the decision boundaries for LDA and QDA, respectively.

2. (6%+8%) Consider the data set $\mathcal{D} = \{((x_{11}, x_{12}), y_1), \cdots, ((x_{61}, x_{62}), y_6)\} = \{((1.5, 2), 1), ((1, 1), 1), ((2, 0.5), 1), ((-2, 0), 2), ((-1, 0), 2), ((-2, -1), 2), ((-1, 2), 3), ((0, 1), 3), ((1, 2), 3)\}$.

(a) Find the classification rule using QDA.

(b) Determine the estimated posterior probabilities of the three classes given the input $(0, 0)$.

3. (6%+8%+2%) Consider the data set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{6} = \{(0, 1), (2, 3), (1, 2), (2, 2), (1, 1), (3, 3)\}$. We want to evaluate the performance of linear regression on the data set using k-fold cross validation. Find the test MSE estimate for $k = 2$ and $k = 3$, respectively. Assume that the data points are partitioned equally in the order given above (e.g., for $k = 2$, the first three points are one fold and the last three points are another fold). Explain your observations.

4. (14%) Let

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

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

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

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%+14%+10%) 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) 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.)

(c) Following (b), 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.