COM 525000 Statistical Learning
Homework #3

(Due December 30, 2019 before noon to the TA at EECS 613.)

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

1. (8%) Recall from Chapter 5 that the LOOCV MSE can be written as

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

where $\mathbf{S} \triangleq \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$. By the approximation $\mathbf{S}_{ii} \approx \frac{1}{n}\mathrm{tr}(\mathbf{S})$ and by the fact that $1/(1-x)^2 \approx 1 + 2x$, for x that is small, show the similarity between the above cross-validation error and C_p statistic.

2. (10%+6%+8%) Suppose that the input and output variables of the n training data points can be expressed as $\mathbf{X} = (x_1, \ldots, x_n)^T$ and $\mathbf{y} = (y_1, \ldots, y_n)^T$, respectively. In ridge regression, the intercept β_0 and the coefficient vector $\beta = (\beta_1, \ldots, \beta_p)^T$ are chosen such that $\|\mathbf{y} - \mathbf{X}\beta - \mathbf{1}\beta_0\|^2 + \lambda\|\beta\|^2$ is minimized for some $\lambda \geq 0$, where $\mathbf{1}$ is an n-dimensional all-one vector.

(a) Show that the resulting coefficient estimate is given by

$$\hat{\beta}_\lambda = (\mathbf{X}_c^T\mathbf{X}_c + \lambda\mathbf{I})^{-1}\mathbf{X}_c^T\mathbf{y}_c,$$

where $(\mathbf{X}_c, \mathbf{y}_c)$ is the centered data set.

(b) Show that $\|\beta_\lambda\|\big|_{\lambda>0} \leq \|\beta_\lambda\|\big|_{\lambda=0}$

(c) Show that the training error is

$$\overline{\mathrm{err}} = \frac{1}{n}\mathbf{y}_c\big[\mathbf{I} - \mathbf{X}_c(\mathbf{X}_c^T\mathbf{X}_c + \lambda\mathbf{I})^{-1}\mathbf{X}_c^T\big]^2\mathbf{y}_c$$

and that it is an increasing function of λ.

3. (10%+12%) Let us consider the data set

$$\mathcal{D} = \{(x_{i1}, x_{i2}, y_i)\}_{i=1}^{n} = \{(-2, 1, 1), (1, -3, -1), (1, 2, 0)\}.$$

(a) By performing subset selection, find the best linear regression model using C_p statistic for comparison among different model sizes.

(b) Find the solution for lasso with regularizer weight $\lambda = 10$.

4. (10%+6%+8%) Consider 3 data points $\{(x_{i1}, x_{i2}, x_{i3})\}_{i=1}^{n}$ given by $(-1,-1,-1)$, $(0,0,0)$, $(2,1,-1)$.

(a) What are the first and second principal components (write down the actual vector)?

(b) If we project the original data points into the 1-d subspace by the principal component you choose, what are their coordinates in the 1-d subspace? And what is the variance of the projected data?

(c) Suppose that $y_1 = -3$, $y_2 = 1$, and $y_3 = 5$. Find the first partial least squares (PLS) direction and compare with the first principal component direction.

5. (8%) Let us consider the basis representation for cubic splines with K interior knots, where

$$f(X) = \sum_{j=0}^{3} \beta_j X^j + \sum_{k=1}^{K} \beta_{k+3}(X - \xi_k)_+^3.$$

Prove that the linear boundary conditions for natural cubic splines imply that $\beta_2 = \beta_3 = 0$ and $\sum_{k=1}^{K} \beta_{k+3} = 0$, $\sum_{k=1}^{K} \xi_k \beta_{k+3} = 0$.

6. (4%+4%+2%+2%+2%) Problem 1 of Chapter 7.