Introduction to Machine Learning
Homework 3b: Cross Validation and Feature Selection

Prof. Yao Wang

Spring 2018

1. Suppose you are given a dataset with N samples (each with a feature vector $\mathbf{x}_n$ and an observed target value y_n). Based on this dataset, you are tasked to design a multi-linear regression function that can be used to predict the target value y for any new sample with a feature vector $\mathbf{x}$. Furthermore, you should report the expected prediction error (mean square error between the predicted value and the true (but unknown) target value for all possible new samples.) The following are several options:

   (a) Use all N samples to determine the optimal linear regressor that will minimize the mean square prediction error for these N samples. Furthermore, calculate the mean square error between the predicted values and the true values among these samples.

   (b) Divide the N samples to two halves, train your linear regressor on one half (training set), and then apply the trained regressor on the samples in the other half (validation set), and evaluate the mean square error for the validation set.

   (c) Run a K-fold cross validation, to generate K regressors, and determine the mean square error for the validation set in each fold. Finally determine the average of the mean square errors obtained by the K validation sets in the K-folds.

   What may be the problem with each approach? Which method would you use? Your answer should consider two cases: when N is very large, and when N is relatively small. Also, with the cross validation approach, how would you use the K different regressors developed to predict a new sample?

2. Suppose you used K fold cross validation to generate K linear regressors denoted by $\boldsymbol{\beta}^k, k = 1, 2, \ldots, K$, with $\boldsymbol{\beta}^k = [\beta_0^k, \beta_1^k, \ldots, \beta_J^k]$. To predict the target value for a new sample with feature $\mathbf{x} = [x_1, x_2, \ldots, x_J]$, you could apply each predictor $\boldsymbol{\beta}^k$ on $\mathbf{x}$ to generate K predictions, $y^k, k = 1, 2, \ldots, K$, and then average these predictions to obtain your final prediction. Show that this is equivalent to derive an average predictor $\bar{\boldsymbol{\beta}} = [\bar{\beta}_j, j = 0, 1, \ldots, J]$, with $\bar{\beta}_j = (\sum_{k=1}^{K} \beta_j^k)/K$ and apply this average predictor to the sample.

3. Consider the development of a linear regressor from a training data again. Suppose the feature vector contains many features and you know that only a subset of them are helpful for predicting the target variable, but you do not know how many features should be included and what they are. One way to do feature selection is by using LASSO regression. Describe how would you go about determine the optimal subset. Consider two cases: when N is very large, and when N is relatively small.

4. Continue with the previous problem. Instead of using the LASSO method, list some other methods that you may use for feature selection.

5. Given raw data samples $x_{i,j}^r, y_i^r$, we often perform data normalization so that normalized features $x_{i,j} = (x_{i,j}^r - \bar{x}_j)/\sigma_j$ and target $y_i = (y_i^r - \bar{y})/\sigma_y$. Here $\bar{x}_j$ and σ_j denote the mean and standard deviation for feature j, and $\bar{y}$ and σ_y denote the mean and the standard deviation for the target, all computed from the given data samples.

   (a) Show that the normalized features and target each have zero mean and unit variance.

   (b) Suppose you would like to predict the normalized target y from the normalized features $x_1, x_2, \ldots, x_J$ using a linear regressor $\hat{y} = \beta_0 + \beta_1 x_1 + \ldots + \beta_J x_J$. You will use the normalized training data to derive the regression coefficients $\beta_j, = 0, 1, \ldots, J$. Show that the optimal intercept term should be zero, i.e., $\beta_0 = 0$.

   (c) Let the regression coefficients determined for the normalized data be $\boldsymbol{\beta} = [\beta_1, \beta_2, \ldots, \beta_J]$. Describe how do you apply $\boldsymbol{\beta}$ to any given raw test sample with features $\mathbf{x}^r = [x_1^r, x_2^r, \ldots, x_J^r]$. What are the equivalent regression coefficients $\boldsymbol{\beta}^r = [\beta_0^r, \beta_1^r, \ldots, \beta_J^r]$ for the raw data?

6. Why is data normalization important with ridge regression and LASSO regression?

7. What are the difference between ridge regression and LASSO regression? What are their pros and cons?

8. Ridge Regression is a linear predictor that minimizes the following loss function

$$J(\boldsymbol{\beta}) = \|A\boldsymbol{\beta} - \mathbf{y}\|^2 + \alpha\|\boldsymbol{\beta}\|^2$$

   Show that the optimal solution is

$$\boldsymbol{\beta}_{\text{opt}} = (A^T A + \alpha I)^{-1} A^T \mathbf{y}$$

9. Instead of using either ridge or LASSO loss function, one can develop a linear regressor by minimizing the following loss function (known as elastic-net):

$$J(\boldsymbol{\beta}) = \|\mathbf{y} - A\boldsymbol{\beta}\|^2 + \alpha(\lambda\|\boldsymbol{\beta}\|^2 + (1-\lambda)\|\boldsymbol{\beta}\|_1)$$

   Show how can you turn this into a LASSO problem, using an augmented version of A and $\mathbf{y}$.