Homework IV

Due date: Oct. 7 (11:59pm)

Problem 1: Ridge Regression and Lasso (40 points)

In this problem, you are asked to learn regression models using ridge regression and lasso. The data set that we are going to use is the Boston Housing data. The task is to predict the house price in suburbs of Boston. More information about the data can be found here `https://archive.ics.uci.edu/ml/datasets/Housing`.

For this assignment, we use the preprocessed data available here `http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/regression.html`. There are two versions: the original version and the scaled version. The txt format and the Matlab format of the data can be downloaded following this google drive link [1]. For the txt format, the first column is the target output y, and the remaining columns are features in the form of (feature_index:feature_value). For the Matlab format, $data.X$ is the data matrix of $n \times d$, and $data.y$ is the target output of $n \times 1$. For question (1), (2), (3), please use the scaled version. For question 4, please use the original version. If we let $\mathbf{x} \in \mathbb{R}^d$ denote the feature vector, the prediction is given by $\mathbf{w}^\top \mathbf{x} + w_0$, where $\mathbf{w} \in \mathbb{R}^d$ contains the coefficients for all features, and w_0 is an intercept (aka bias) coefficient.

For running both ridge regression and Lasso, you can use the scikit-learn Pytohn library, where is available here `https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html` and here `http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Lasso.html`. If you use the Python module in sklearn, you can run lasso by

```
sklearn.linear_model.Lasso(alpha, fit_intercept=True)
```

or ridge regression by

```
sklearn.linear_model.Ridge(alpha*(2*n), fit_intercept=True)
```

where alpha is a (scaled) regularization parameter (similar to the λ in the lecture). Note that we multiple the alpha by $2 * n$ (2 times the number of training examples) when calling for Ridge regression, which is to make it consistent with Lasso.

(1). Solution of Ridge Regression and Lasso: Set the value of the regularization parameter $\alpha = 1$, compute the optimal solution for ridge regression and lasso. Report the optimal solutions for both ridge regression and lasso.

(2). Training and testing error with different values of α: (i) Take the first 400 examples as training data and remaining 106 examples as testing data. (ii) For each value of α in [0, 0.001, 0.01, 0.1, 1, 10, 100], run the ridge regression and lasso to obtain a model ($\mathbf{w}$) and then compute the root mean square error on both the training and the testing data of the obtained model.

[1] `https://drive.google.com/drive/folders/1oSIx8P4xQcdNOJHvzeODVXO5VjNcjTCU`

(iii) Plot the error curves for root mean square error on both the training data and the testing data vs different values of α. You need to show the curves, and discuss your observations of the error curves, and report the best value of λ and the corresponding testing error.

(3). Cross-validation: Use the selected 400 examples as training, follow the 5-fold cross-validation procedure to select the best value of α for both ridge regression and lasso. Then train the model on the 400 examples using the selected α and compute the root mean square error on the testing data. Report the best α and the testing error for both ridge regression and lasso.

(4). Repeat (3) using the original version of the data and compare the results with that obtained for (3).

Remark 1: for a set of examples $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, the root mean square error of a prediction function $f(\cdot)$ is computed by RMSE $= \sqrt{\sum_{i=1}^{n}(f(\mathbf{x}_i) - y_i)^2/n}$.

Remark 2: You are allowed to explore some other libraries for ridge regression and Lasso. To make sure consistency, you should map the regularization parameter in your used package to the same value of α in sklearn for Ridge and Lasso, respectively.

Problem 5: Ridge Regression and Lasso

Repeat the first three questions (1), (2), (3) as in Problem 1 on E2006-tfidf data using the provided training/testing split (the description of the data can be found here `http://www.cs.cmu.edu/~ark/10K/`). The data is from LibSVM website. Note that we use the original testing set for training and the original training set for testing.