COEN 129 Machine Learning and Data Mining

Homework 3 (100 points)

Due: 3:30pm, Wednesday, May 10, 2017

Implement Linear Regression and Ridge Regression. You can use Python, R or Matlab for this assignment. Please do not use any machine learning library.

Training data: http://www.cse.scu.edu/~yfang/coen129/crime-train.txt
Test data: http://www.cse.scu.edu/~yfang/coen129/crime-test.txt
A description of the variables: http://www.cse.scu.edu/~yfang/coen129/communities.names

The data consist of local crime statistics for 1,994 US communities. The response y is the crime rate. The name of the response variable is *ViolentCrimesPerPop*, and it is held in the first column of *df_train* and *df_test*. There are 95 features x_i. These features include possibly relevant variables such as the size of the police force or the percentage of children that graduate high school. The data have been split for you into a training and test set with 1,595 and 399 entries, respectively. The features have been standardized to have mean 0 and variance 1.

Exercises:

- Perform linear regression directly using the closed form solution. Compute the RMSE value on the training data and test data, respectively.

- Perform ridge regression directly using the closed form solution. Use k-fold cross validate (k=5) to select the optimal λ parameter. Compute the RMSE value on the test data.

  You can begin by running the solver with λ = 400. Then, cut λ down by a factor of 2 and run again. Continue the process of cutting λ by a factor of 2 until you have models for 10 values of λ in total.

- Perform linear regression using the gradient descent algorithm. Compute the RMSE value on the training data and test data, respectively.

  For the initial weights, you can just use Gaussian *N(0, 1)* random variables. Define "converging" as the change in any coefficient between one iteration and the next is no larger than 10^{-5}.

- Perform ridge regression using the gradient descent algorithm. Compute the RMSE value on the test data.