Lab 5

Implement cross-validation and gradient descent for regression. Please do not use any machine learning library. The dataset is the same with Lab 4.

Training data: http://www.cse.scu.edu/~yfang/coen140/crime-train.txt
Test data: http://www.cse.scu.edu/~yfang/coen140/crime-test.txt
A description of the variables: http://www.cse.scu.edu/~yfang/coen140/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:

1. 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.
    a. You can begin by running the solver with $\lambda = 400$. 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.
    b. Report the λ you chose, why you chose it, and the associated training error.

2. Perform linear regression using the gradient descent algorithm. Compute the RMSE value on the training data and test data, respectively. Compare the results with your Lab 4 and see if you can obtain the same results. 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 a small value (e.g., 10^{-5}).
    a. Report training and testing error.
    b. Define a function called problem2(samples): that, given an NxP array, returns predictions in the form of a Nx1 array. The dummy feature will be included in the last column of the NxP array.

3. Perform ridge regression with 5-fold cross valuation using the gradient descent algorithm. Compute the RMSE value on the test data and see if you can obtain the same results with those from the closed form solution in Exercise 1.
    a. Report training and testing error.
    b. Define a function called problem2(samples): that, given an NxP array, returns predictions in the form of a Nx1 array. The dummy feature will be included in the last column of the NxP array.