CS6220 Data Mining Techniques – Spring 2015
Assignment 1

Submission Instructions

- **Your program must run on CCIS machines in WVH 166 lab**
- **Create a README file, with simple, clear instructions on how to compile and run your code**
- **Zip all your files (code, README, written answers, etc.) in a zip file named** $\{firstname\}\_\{lastname\}\_CS6220\_HW1.zip$ **and upload it to Blackboard**

In this assignment, you are given 3 datasets. Each dataset has a training and test file. Specifically, these files are:

1000_100_train.csv,	1000_100_test.csv
100_100_train.csv,	100_100_test.csv
100_10_train.csv,	100_10_test.csv

Create 3 additional training files from the 1000_100_train.csv by taking the first 50, 100, and 150 instances respectively. Call them: 50(1000)_100_train.csv, 100(1000)_100_train.csv, 150(1000)_100_train.csv. The corresponding test file for these dataset would be 1000_100_test.csv and no modification is needed.

1. Implement $L2$ regularized linear regression algorithm with λ range from 0 to 150 (integers only). For each of the 6 dataset, plot both the training set MSE and the test set MSE as a function of λ (x-axis) in one graph.

   Discuss : How does λ affect the MSE in general? How does the choice of λ depend on the number of features vs. examples? How does λ change with number of examples when the number of features is fixed?

2. Fix $\lambda = 1$, 25, 150. For each of these values, plot a learning curve for the algorithm using the dataset 1000_100.csv.
   Note: a learning curve plots the performance as a function of the size of the training set. To produce the curve, you need to draw random subsets (of increasing sizes) and record performance (MSE) on the corresponding test set when training on these subsets. In order to get smooth curves, you should repeat the process at least 10 times and average the results.

3. From the plots in question 1, we can tell which value of λ is best for each dataset once we know the test data and its labels. This is not realistic in real world applications. In this part, we use cross validation to set the value for λ. Implement the CV technique given in the class slides. For each dataset, compared the values of λ and MSE with the values in question 1).

   Discuss: How do the values for λ and MSE obtained from CV compare to the ones in question 1? What are the drawbacks of CV? What are the factors affecting the performance of CV?