CS6930 Data Mining – Fall 2017
Assignment 1

Submission Instructions

- **Your program must run on machines in Leon Lowenstein Bldg. 812**
- **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** $\{firstname\}\_\{lastname\}\_CS6930\_HW1.zip$ **and upload it to Blackboard**

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

dataset 1:	train-100-10.csv	test-100-10.csv
dataset 2:	train-100-100.csv	test-100-100.csv
dataset 3:	train-1000-100.csv	test-1000-100.csv

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

1. Implement $L2$ regularized linear regression algorithm with λ ranging 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.

   (a) For each dataset, which λ value gives the least **test** set MSE?

   (b) For each of datasets 100-100, 50(1000)-100, 100(1000)-100, provide an additional graph with λ ranging from 1 to 150.

   (c) Explain why $\lambda = 0$ (i.e., no regularization) gives abnormally large MSEs for those three datasets in (b).

2. 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 (CV) to set the value for λ. Implement the 10-fold CV technique discussed in class (pseudo code given in Appendix A) to select the best λ value from the **training** set.

   (a) Using CV technique, what is the best choice of λ value and the corresponding test set MSE for each of the six datasets?

(b) How do the values for λ and MSE obtained from CV compare to the choice of λ and MSE in question 1(a)?

(c) What are the drawbacks of CV?

(d) What are the factors affecting the performance of CV?

3. 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 (i.e., test set MSE) 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.

Appendix A
10-Fold Cross Validation for Parameter Selection

Cross Validation is the standard method for evaluation in empirical machine learning. It can also be used for parameter selection if we make sure to use the training set only.

To select parameter λ of algorithm $A(\lambda)$ over an enumerated range $\lambda \in [\lambda_1, \ldots, \lambda_k]$ using dataset D, we do the following:

1. Split the data D into 10 disjoint folds.
2. For each value of $\lambda \in [\lambda_1, \ldots, \lambda_k]$:
   (a) For $i = 1$ to 10
   - Train $A(\lambda)$ on all folds but i^{th} fold
   - Test on i^{th} fold and record the error on fold i

   (b) Compute the average performance of λ on the 10 folds.
3. Pick the value of λ with the best average performance

Now, in the above, D only includes the training data and the parameter is chosen without knowledge of the test data. We then re-train on the entire train set D using the chosen parameter value and evaluate the result on the test set.