**CS 535/EE 514 – Machine Learning**

**Assignment 4**

Deadline: Nov. 12 (Sunday) at 11:59 PM

Question 1 (30 points) [using built-in functions is not allowed]

Perform k-NN classification on the Iris dataset. This dataset, which is available on UCI Machine Learning Repository, captures a 3-class classification problem defined over 4 attributes. Here is what you have to do:

1. Randomly separate 45 examples for testing (15 examples from each class)
2. From the remaining 105 examples, create training datasets of sizes (i) 30, (ii) 90, and (iii) all 105 examples. The (i) and (ii) sets will be randomly selected from the 105 examples left after removing the test set.
3. Report performance as accuracy on the test set when using each of the training set with k = 1, 3, and 7 (9 results). Plot the variation of accuracy with k and training data set size. Comment o the results.

Do the above in MATLAB or the language of your choice. Provide a readme file with instructions on running the experiments.

Question 2 (40 points) [using built-in functions for linear regression is not allowed]

This exercise explores linear models for regression on real data sets. You will have to code the program, ideally in MATLAB (using formulas studied in class and in PRML book).

Download and study the Wine Quality (Red) dataset from UCI Repository.

1. Partition the dataset into three: training (70%), validation (15%), and test (15%). Find the regularization coefficient that gives the best performance (root mean squared error, RMSE) on the validation set. Report the regularization coefficient value, the weight values, and corresponding models’ performances on the validation dataset. For the best performing regularization coefficient, train on the entire training +validation set and report results (rMSE) on the test set.
2. Choose four regularization coefficient values lying between zero and the best found in (a) above. For each of these values build 10 models on 10 randomly selected sets (80%) and evaluate it on the remaining 20% dataset. Report the average and standard deviation of these results for each of the four regularization parameters.
3. Comment on the results obtained above, especially regarding regularization coefficient, over-fitting, and weights and their interpretations.

Question 3 (30 points)

Experiment with Python’s scikit-learn implementation of OLS (ordinary least squares, standard linear model based on SSE), ridge regression, and lasso regression. For help yon check out:

<http://scikit-learn.org/stable/modules/linear_model.html>. Do these experiments on both Wine Quality (Red) dataset from UCI repository and Housing (Boston) dataset uploaded on LMS. You can also use Boston dataset already available with scikit (<http://scikit-learn.org/stable/modules/generated/sklearn.datasets.load_boston.html> ).

1. Build OLS model on the datasets. Report, evaluate, and comment on the results (i) RMSE/SSE/negSSE after 10-fold cross-validation, (ii) weights and their confidence intervals.
2. Build ridge regularized linear models. Use the automatic procedure to fix the regularization coefficient. Report (i) and (ii) as in (a) above.

**Clarifications:** Please note that for lambdas, use any value starting from zero. But your ultimate goal is to find best regularization coefficient so try to experiment with at least 10 different values to see which one gives you best results and explain your results in the report. Also, you have to use squared term for X as features instead of just using original X. (add squared term of every x to form more features in other words). This will show you the results of varying value of lambda.

For 10-fold cross validation, you will have to report RMSE/SSE/negSSE for every validation i.e 10 values.

**Instructions:**

1. **This is assignment is focused on understanding of basic concepts. So provide brief reasoning for your answers in the Report that you will be writing in word document.**
2. **All assignments must be submitted on LMS.**
3. **Plagiarism will be checked. Plagiarized assignments will either face points deduction or will be reported to disciplinary committee.**
4. **Late submissions will be penalized according to policy announced by the instructor. The entire assignment will be considered late even if parts of it are submitted before the deadline.**