COMP-551 : Applied Machine Learning — Given on: Jan 16, 11pm

Programming Assignment #1 — Due on : Jan 26, 5pm [1]

- This is an individual assignment. However, you are allowed to discuss the problems with other students in the class. But you should write your own code and report.
- If you have any discussion with others, you should acknowledge the discussion in your report by mentioning their name.
- Be precise with your explanations in the report. Unnecessary verbosity will be penalized.
- You have to hand in the report as a hard-copy in the assignments hand-in box opposite to MC318. You have to submit the code in myCourses. Note that you should do both these submissions before the deadline.
- After the deadline, you have one week to submit your assignment with 30% penalty.
- You are free to use libraries with general utilities, such as numpy and scipy for python. However, you should implement the algorithms yourself, which means **you should not use pre-existing implementations of the algorithms as found in SciKit learn, Tensorflow, etc.**!
- If you have questions regarding the assignment, you can ask for clarifications in the class discussion forum or go to the following office hours: Prasanna, Philip (section 1), Sanjay, Lucas (section 2).

1 Model Selection

You have to use Dataset-1 for this experiment. Dataset-1 consists of train, validation, and test files. The input is a real valued scalar and the output is also a real valued scalar. The dataset is generated from an n-degree polynomial and a small Gaussian noise is added to the target.

1. Fit a 20-degree polynomial to the data. Report the training and validation MSE (Mean-Square Error). Do not use any regularization. Visualize the fit. Comment about the quality of the fit.
2. Now add L2 regularization to your model. Vary the value of λ from 0 to 1. For different values of λ, plot the training MSE and the validation MSE. Pick the best value of λ and report the test performance for the corresponding model. Also visualize the fit for the chosen model. Comment about the quality of the fit.
3. What do you think is the degree of the source polynomial? Can you infer that from the visualization produced in the previous question?

[1] Because of a wrong date appearing on the slides in class, we'll accept submissions until Jan 29 at **noon**. We recommend aiming for the original deadline since the second assignment will start Jan 26.

2 Gradient Descent for Regression

You have to use Dataset-2 for this experiment. Dataset-2 consists of train, validation, and test files. The input is a real valued scalar and the output is also a real valued scalar.

1. Fit a linear regression model to this dataset by using stochastic gradient descent. You will do online-SGD (with one example at a time). Use the step size of 1e-6. Compute the MSE on validation set for every epoch. Plot the learning curve i.e. training and validation MSE for every epoch.

2. Try different step sizes and choose the best step size by using validation data. Report the test MSE of the chosen model.

3. Visualize the fit for every epoch and report 5 visualizations which shows how the regression fit evolves during the training process.

3 Real life dataset

For this question, you will use the Communities and Crime Data Set from the UCI repository (`http://archive.ics.uci.edu/ml/datasets/Communities+and+Crime`).

1. This is a real-life data set and as such would not have the nice properties that we expect. Your first job is to make this dataset usable, by filling in all the missing values. Use the sample mean of each column to fill in the missing attribute. Is this is a good choice? What else might you use? If you have a better method, describe it, and you may use it for filling in the missing data. Turn in the completed data set.

2. Fit the above data using linear regression. Report the 5-fold cross-validation error: The MSE of the best fit achieved on test data, averaged over 5 different 80-20 splits, along with the parameters learnt for each of the five models.

3. Use Ridge-regression on the above data. Repeat the experiment for different values of λ. Report the MSE for each value, on test data, averaged over 5 different 80-20 splits, along with the parameters learnt. Which value of λ gives the best fit? Is it possible to use the information you obtained during this experiment for feature selection? If so, what is the best fit you achieve with a reduced set of features?

Instructions on how to use 80-20 splits

1. Make 5 different 80-20 splits in the data and name them as $CandC-train\ \langle num \rangle\ .csv$ and $CandC-test\ \langle num \rangle\ .csv$.
2. For all 5 datasets that you have generated, learn a regression model using the 80% data and test it using 20% data.
3. Report the average MSE over these 5 different runs.

Instruction for code submission

1. Submit a single zipped folder with your McGill id as the name of the folder. For example if your McGill ID is 12345678, then the submission should be 12345678.zip.
2. If you are using python, you must submit your solution as a jupyter notebook.
3. Make sure all the data files needed to run your code is within the folder and loaded with relative path. We should be able to run your code without making any modifications.

Instruction for report submission

1. You report should be brief and to the point. When asked for comments, your comment should not be more than 3-4 lines.
2. Report all the visualizations (learning curves, regression fit).
3. Either print colored version of your plot or differentiate the plot items by using different symbols. Make sure that you are not submitting a black and white version of a colored plot which we cannot evaluate.