Machine Learning Report

In Assignment 1, two famous Machine learning diagnostics (techniques) were applied on our House price Dataset: Model selection and Regularization.

Note: comparison is done based on the least test cost, where the cost is reflection of error, specifically mean square error.

In Model Selection, I divided the dataset into three parts: training, cross validation, and test set. First, I normalized the training set using mean normalization, and the then i normalized the cross validation and test set using the mean and standard deviation of the training set in order for our model to converge faster. First, I looped over different degrees of polynomial hypotheses and I fit our training data on each degree in order to calculate the thetas using gradient descent. Then for each thetas obtained from different degrees, i calculated the cross validation error using the cost function. The degree that yields the least cross validation error was considered the optimal degree. Therefore, such thetas corresponding to the optimal degree was tried on the test set to obtain test data error using cost function. In my code, I used polynomial features function predefined in Sickit-learn. Then, I used Linear Regression function as well to ensure that our gradient descent approach has converged after correct number of iterations with the suitable learning rate for each degree of the hypothesis. This function was compared with gradient descent function implemented in previous assignment, and similar results for thetas were obtained. Moreover, it can be seen i tried three degrees: one, two, and three, yet i actually saw the results up to degree 5 in a previous long run, and it was found that

degrees 4 and 5 results in much higher cross validation errors (about 10e+10 higher than error in degree 2).As for the results, it was found that the optimal degree in our case is degree one with cross validation cost of 2.5249098726844902e+17, giving a test cost (error) of 2.7878747625683152e+17.

After that, another approach was implemented, which is K-fold cross-validation (sampling). I used 5-fold sampling, as if having 5 different training and test sets. I fit each of training set to a polynomial of degree 1, since by intuition degree 1 in previous approach gave a much less cost than that of degree 2. Then, the average test error of 5 test sets was obtained, giving a lower test error than model selection method above with only one test set. The calculated test cost is 2.5854714604325715e+17.

Then, Stratified 5-fold sampling was done in a similar fashion. This ensures that class proportions are maintained in each selected set, reducing average test cost to 2.5791884343706106e+17.

Finally, Regularization was applied using 6 values of lambdas with 2 different degrees only as after that python would stop responding. The idea is similar to first method except for the fact that cost and gradient functions are a bit different, due to the presence of lambda, which tends to keep all Features but reduce the parameters of some features. The combo that resulted in least cross validation cost was 0.04 for lambda and 2 for optimal degree for chosen lambdas and degrees. Regularization resulted in the least test cost of 1.4643704081597123e+17.

As for assignment 2, no Dataset was given for logistic regression as expected, so it was not implemented since old assignment dataset was so small, not suitable for our two methods. However, the idea is almost the same as above, with the cost function of logistic regression as the main difference. That cost function was already seen in previous assignment. Misclassification error (0/1 Error), as seen in lecture 5 in slide 7, can also be used as an alternative to cost function. Everything else is identical to what we have implemented above, for Model Selection and Regularization.