

LAB4

ELISE MCELLHINEY

1. PART 1

The solution to the OLS solution has betas: [1.02978479, 4.17324793, 2.74546011, 3.2875157, -3.43724678, -4.43006177, 1.30612728, 2.61200878, -4.36851666, 0.70934935, -3.48629435]

Ein OLS MSE: 1.02998036313

Eout OLS MSE: 1.03448674192

2. PART 2

I capped all my models maximum iterations at 1000 so I could get a sense of how fast they converged. From the results, listed and reported in testfile.csv, I found that batchsize has a significant impact on how fast the data converged to a solution. Depending on the alpha that I used, I found that within 1000 iterations that a batch size of 20-50 seemed to work the best with regards to both training and test data. The stop condition was interesting since it seemed to most impact the online learning since it then tested the "batch" of size 1 for MSE and would cut the models training off early since it's fairly likely that a single value will be guessed within the tolerance before training should be complete.

If I didn't use a stop-condition, I found that online took a long time since it didn't take advantage of vector calculations. If I used a low enough stop-condition, given a large enough batch size and unlimited by iterations, my solution converged towards the OLS solution. During other tests, no longer in the code, I used `train_test_split` to cross-validate and check that in-sample and validation splits followed expected patterns. Essentially I wanted to check that in-sample and out-of-sample were correlated. The bias of linear models is typically high, so modeling too much variance wasn't a huge concern in this lab. Given data that is this nice, the OLS solution is a reasonably good model. I'd say that stopping training within defined tolerances on batch sizes of 20-50 would be reasonable.

3. PART 3

The results of this experiment, recorded every 1000 iterations, are in `perceptron_testfile.csv`. I ran it for an entire day on my virtual machine, but my virtual machine is too slow to make it to the number of iterations required for it to converge within a reasonable amount of time. This is why I had to use a late day. After 2084000 iterations, my training and testing confusion matrices were:

Training confusion matrix:

$$\begin{bmatrix} 492 & 5 \\ 7 & 496 \end{bmatrix}$$

Date: March 27, 2017.

Testing confusion matrix:

$$\begin{bmatrix} 259 & 4 \\ 3 & 234 \end{bmatrix}$$