Shade Alabsa

HW 03

Mignon Kang

Machine Learning

This homework was a bit harder than I expected and I had some issues wrapping my head around some of the things. I’ll point them out as I talk about what I did. To start off with I watched Andrew’s Ng video on logistic regression and got most/all of my equations from there. This is where I defined my sigmoid function, predict, cost, and gradient descent formulas.

Alright to start off with I used Kfold from the sklearn library to do the k fold cross validation bit. From there I split up the data into y\_training, x\_training, x\_test, and y\_test. I then normalized the data and divided the entire x sets by 255. This wasn’t strictly necessary for this assignment I don’t think since they are already contained between 0 and 255. From there I ensured the y sets were between 0 and 1 by modding by 3 and if it was 0 then set to 0 otherwise set it to 1. From there I ran the logistic regression method using a learning rate of 5e-5 and 1000 iterations. My logistic regression functions returns the weights and costs. Technically I should check to make sure the costs are converging but I’m lazy and I just checked my accuracy. I then predicted on the test set. Then with the return value of predict I set everything to 0 or 1 pending on if it’s greater .5 or not. From there I calculate accuracy, the lazy part from earlier then it’s time check true positive rate and false positive rate. Linda had a genius idea of using a function from numpy called logical\_and. She used it slightly different but I wanted to make sure she got credit for showing me that function. Once I calculate TPR and FPR I add them to a list to take care of after. Now let’s go into logistic regression and gradient descent. My logistic regression method got n(1000) times calculating the descent using x and weights which gets updated each time. I update the costs and weights and keep going. This part wasn’t terribly difficult but the fact we don’t do anything with the cost bothers me. Afterwards I used the averaged true positive and false positive rates to plot the ROC curve. I first add 0 and 1 to both to give me my end points, sort them, and then plot them. The sorting was suggested to me by another student when it looked like a 5 year old drew my graph. What I don’t understand is why sorting them works since the FPR and TPR have two correlated values but when it gets sorted they are no longer correlated. But it appears to work. Here is my ROC curve.

