Siyuan Zou 111639762 CSE353: Homework 1 Algorithm Report

Empirical Risk/Performance

Naïve Perceptron:

- At iteration 16: Accuracy of 0.8172
- At iteration 100: Accuracy of 0.8031
- At iteration 500: Accuracy of 0.8594

Pocket Perceptron:

- At iteration 16: Accuracy of 0.8172
- At iteration 100: Accuracy of 0.8576
- At iteration 500: Accuracy of 0.8787
- At iteration 3000: Accuracy of 0.8787

Linear Regression:

- Accuracy of 0.3444

Discussion

Naïve Perceptron:

In our program we had a number that dictates the maximum amount of iteration hence, the program will also end once the outer loop hits that certain number. However, for the purpose of this discussion lets assume a few things. First the data is correct hence no error in mathematical computation (i.e. the dot products). Next the number of maximum iterations is not limited. The reason behind why a perceptron algorithm not stopping due to Non-Linear separability this is a phenomenon such that there doesn't not exist and "straight line" that can correctly separate the data points. Our Breast Cancer dataset consisted of 5 total features and over 500 data points and due to this it is hard to truly say if our data is linearly separable or not hence after 500 iterations, we still only reach an accuracy of 85%. Maybe with a smaller size dataset and simpler dimensions linear separability will be more apparent. A side note of my perceptron algorithm is that it will terminate if that dataset finds a weight that result in an accuracy of 1 to speed up termination. However, this ultimately might just slow down the program even more as during each iteration we would have to loop through all datapoint twice once during setting new weight and once during checking accuracy.

Pocket Perceptron:

The Pocket Perceptron algorithm is very similar to that of the naïve version, as I didn't change much of the algorithm only injecting a few if statements comparing the best weights whose accuracy was the highest among all previous weights. Hence this algorithm suffers a similar fate of unable to know if the dataset given is linearly separable. Thus, the parameter of max iterations remains as it is. Overall if I remove the accuracy check in naïve version, pocket version will run slower than that of the naïve version, but the time complexity or big O ultimately remain the same O(xnm), x the number of max iterations, n the size of dataset, and m the length of an instance.

Linear Regression:

In the Linear Regression portion of this homework, we used eigenvalue decomposition as a way of doing the linear regression for the benign dataset and the malignant dataset. This was done through mostly linear algebra using the numpy library. Obvious draw back of this method was

simply that it was a bad algorithm for binary classification. This is clear as the resulting accuracy was very low. In more detail reason the use of one of the feature-vector as the label vector was perhaps not a good way to fit the data. As this feature might be a strong factor that dictates if there is a good linear fit to the data. Aside from the drawback there is no good reason as to using this algorithm for binary classification, as other algorithm such as perceptron and logistic regression can be used for a better result. The biggest takeaway from this portion was understanding how to do a linear regression using eigenvalue decomposition.

README

Running the code

python3 perceptron.py --version [naïve, pocket] (<- mandatory parameter)

--dataset [/path/to/file] (default: ./data/Breast-cancer-data.csv)

python3 linear_regression.py --dataset [/path/to/file] (default: ./data/Breast-cancer-data.csv)

--dataset (can pass in path to other dataset)