I: Results

Perceptron for linearly-separable-dataset

ERM

Weighs :[−0.9999999999999999, −4.002891027159523, 2.0011281397947203]

Errors : 0

Mean error : 0

10-folds

Weights:

[[−0.5, −1.99973802564, 0.99905664602], [−0.5, −1.99800193175, 0.99804043537], ...]

Errors: [0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.008695652173912993]

Mean error: 0.087%

Perceptron for breast-cancer-data

ERM Weights: [[390.9000000000204, 3479.260800000512, −3862.098000001615, 3714.0840000002177, −462.17999999966224, −132.97643099999803]]

Errors: 0.14938488576449915

Mean Error: 14.938%

10-folds

Weights: [[347.80000000, 2916.1303000, −2581.7009999, 2585.6089999, −361.849999999, −134.141813000], ...]

Errors: [0.16071428571, 0.16071428571, 0.14285714285, 0.10714285714, 0.21428571428, 0.125, 0.17857142857, 0.05357142857, 0.08928571428, 0.21428571428]

Mean Error: 14.464%

Adaboost for breask-cancer-data

ERM

Weights: [[1.0690517271111761, 0.6014829819939954, 0.6044019474621684, 0.6047723374583879, 0.38683032423882596, 0.3423632242046578, 0.29615277096509846, 0.3788585081463763]]

Errors: 0.05623901581722324

Mean Error: 5.624%

10-folds

Weights: [[1.0705877129221277, 0.5998269206346589, 0.5971452189857973, 0.609035192681899, 0.37960075095246293, 0.3166058342508909, 0.2926110241039865, 0.36861995172827045], ...]

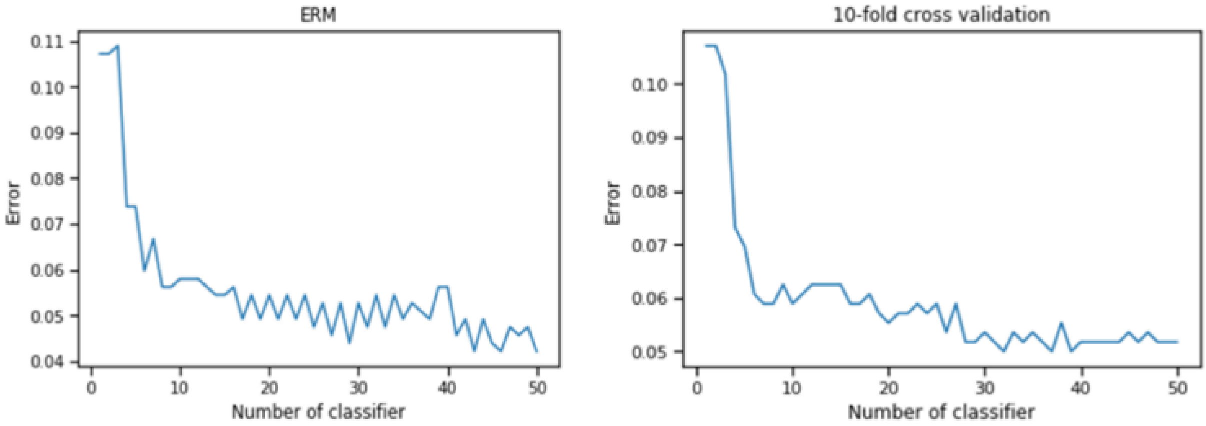
Errors: [0.03571428571, 0.05357142857, 0.03571428571, 0.08928571428, 0.10714285714, 0.03571428571, 0.0, 0.05357142857, 0.07142857142, 0.10714285714]

Mean Error:5.893%

II: Perceptron algorithm

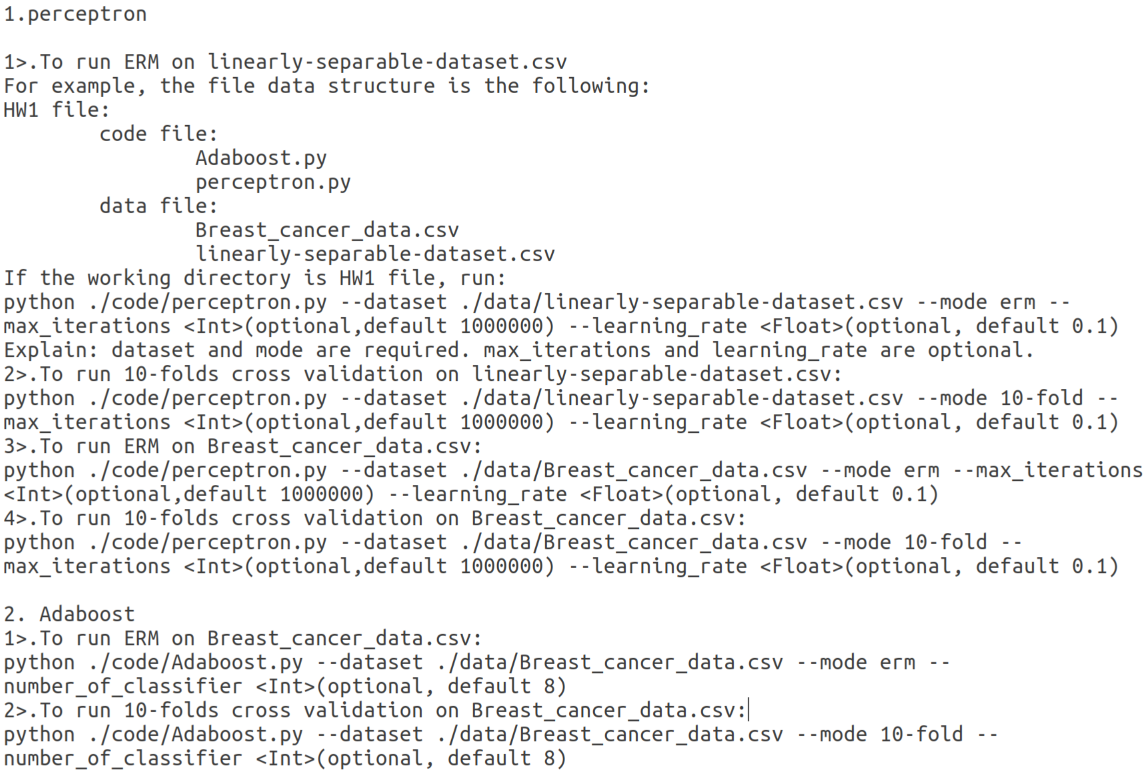
The running time of perceptron algorithm is bounded by O(T * m) where T is the total number of iterations and m is the number of data points. If the dataset is linear separable and T is bigger enough, then at certain iteration, it will loop through all the data points and there does not exist one data point such that $y_i * < w^{(t)}, x_i > \leq 0$. In this case the perceptron algorithm will terminate since the w it calculated successfully separated the dataset. After some reasonable amount of time, if the algorithm did not stop, the strategy I am using is to set a maximum number of iterations, then it will go into a while true loop and at any iteration, loop through all the data points. If there exists one data point such that $y_i * < w^{(t)}, x_i > \leq 0$, increment the number of iterations starts from 0. In this case we know that our algorithm can learn better w. After looping all the data point we want to know if the number of iterations we have already perform is bigger than maximum number of iterations or not. If the answer is yes, then we terminate the while true loop and we know that this dataset is non-linear separable. If it is not set, maximum number of iterations is by default 1000000. The reason why the algorithm did stop or did not stop is: at the while true loop, we initialize a variable flag to be 1. And as well as we loop through all the data points, at certain iteation, if there exists one data point such that $y_i * < w^{(t)}, x_i > \leq 0$, we set this flag to 0, otherwise the algorithm will stop. Hence we know the value w cannot successfully separate the dataset yet. Finally since the maximum iterations is set by us and is bigger enough. When we reach the maximum iteration and the flag is still 0, we know that algorithm will not stop and the dataset is non-linear separable.

III: Adaboost



According to the graph above, we can see that the error is decreasing quickily when the number of classifier increased in the range between 1 to 10. And when the number of classifier keeps increasing, the error decreased not too much and kept fluctuating up and down. Based on my experimental observations and elbow method, the number of classifier I choose is eight because eight is the turning point.

IV. ReadMe



```
1.perceptron

1>.To run ERM on linearly-separable-dataset.csv
For example, the file data structure is the following:
HW1 file:
        code file:
                Adaboost.py
                perceptron.py
        data file:
                Breast_cancer_data.csv
                linearly-separable-dataset.csv
If the working directory is HW1 file, run:
python ./code/perceptron.py --dataset ./data/linearly-separable-dataset.csv --mode erm --
max_iterations <Int>(optional,default 1000000) --learning_rate <Float>(optional, default 0.1)
Explain: dataset and mode are required. max_iterations and learning_rate are optional.
2>.To run 10-folds cross validation on linearly-separable-dataset.csv:
python ./code/perceptron.py --dataset ./data/linearly-separable-dataset.csv --mode 10-fold --
max_iterations <Int>(optional,default 1000000) --learning_rate <Float>(optional, default 0.1)
3>.To run ERM on Breast_cancer_data.csv:
python ./code/perceptron.py --dataset ./data/Breast_cancer_data.csv --mode erm --max_iterations
<Int>(optional,default 1000000) --learning_rate <Float>(optional, default 0.1)
4>.To run 10-folds cross validation on Breast_cancer_data.csv:
python ./code/perceptron.py --dataset ./data/Breast_cancer_data.csv --mode 10-fold --
max_iterations <Int>(optional,default 1000000) --learning_rate <Float>(optional, default 0.1)

2. Adaboost
1>.To run ERM on Breast_cancer_data.csv:
python ./code/Adaboost.py --dataset ./data/Breast_cancer_data.csv --mode erm --
number_of_classifier <Int>(optional, default 8)
2>.To run 10-folds cross validation on Breast_cancer_data.csv:
python ./code/Adaboost.py --dataset ./data/Breast_cancer_data.csv --mode 10-fold --
number_of_classifier <Int>(optional, default 8)
```

This is a picture. To get the command, there is another file called ReadMe.txt in the same file.