REPORT

CSE 512: Assignment 1

Chumki Acharya (112683478)

Performance:

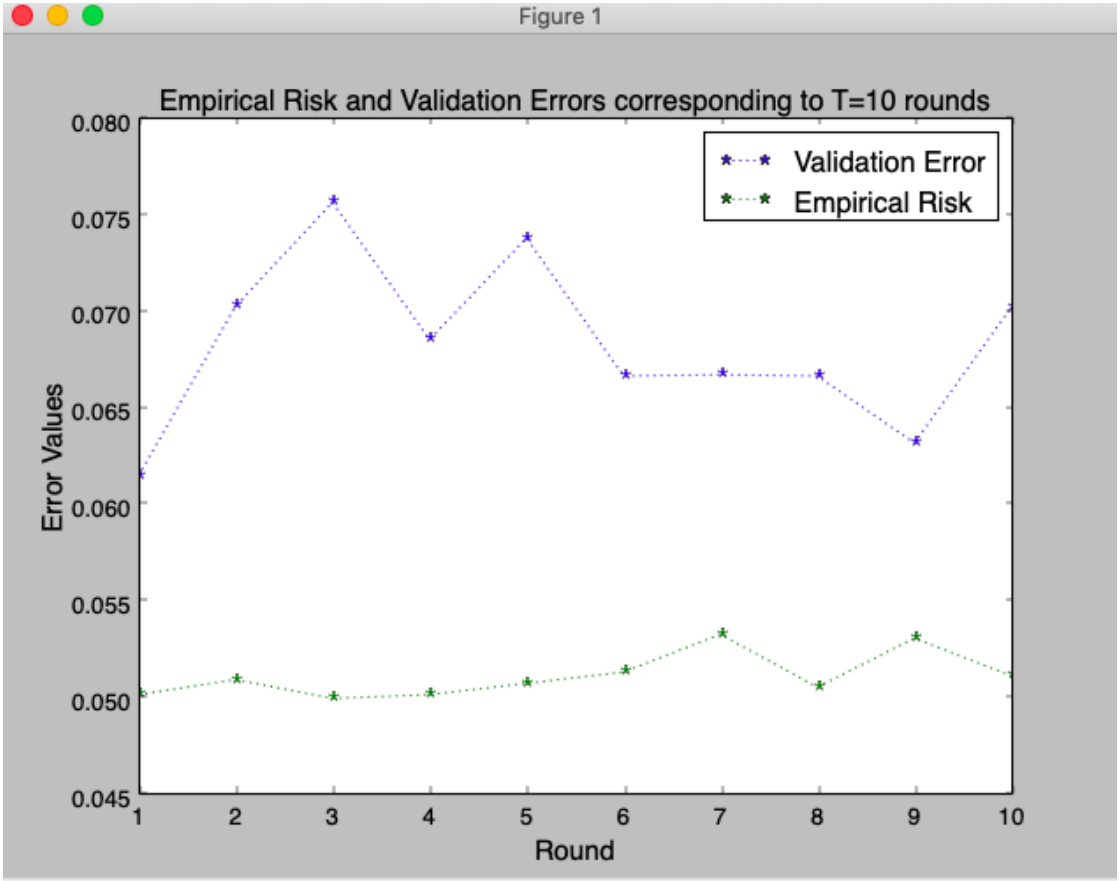
Algorithm	Mode	Dataset	Epochs	Learning Rate	Loss/Error	Accuracy
Perceptron	ERM	Linearly Separable	100	0.1	0.00	1.0
Perceptron	Cross Validation	Linearly Separable	100	0.1	0.056	0.94
Perceptron	ERM	Breast Cancer	100	0.1	0.016	0.84
Perceptron	Cross Validation	Breast Cancer	100	0.1	0.042	0.95
AdaBoost	ERM	Breast Cancer	10	N/A	0.055	0.94
AdaBoost	Cross Validation	Breast Cancer	10	N/A	0.067	0.93

Perceptron:

The perceptron algorithm terminates more faster in case of linearly separable dataset than the Breast cancer dataset. The reason may be that the dataset unlike the previous one is not exactly linearly separable and thus the error of prediction cannot reach zero. The algorithm converges by making the loss value zero i.e. zero error in case of linearly separable dataset. Thus, the heuristic was to stop the algorithm when the error value becomes constant. It has been observed that the algorithm stops when the error becomes constant in case of the Breast Cancer Dataset.

AdaBoost:

The plot of empirical risk and cross validation errors corresponding to T values ranging from 1 to 10 has been shown in the figure below. It has been observed that the cross validation error for all rounds is reasonably higher than the empirical risk values for the same rounds. The reason behind this is that the error on training data would be constant and less than that on the testing data. Here the Empirical Error is the error in the training data while the Validation Error is the error observed on the predictions made on the testing data. Also, the algorithm with different T values reach an optimal accuracy after 10 rounds, so we can conclude 10 rounds of boosting will give us optimal solution.



ReadMe:

1. Perceptron Algorithm

- File Name: perceptron.py
- Datasets: linearly-separable-dataset.csv / Breast_cancer_data.csv
- Modes: erm / crossvalidation
- Run following on command line mentioning the appropriate dataset file and mode:
  python perceptron.py --dataset linearly-separable-dataset.csv --mode erm

2. AdaBoost Algorithm

- File Name: adaBoost.py
- Dataset: Breast_cancer_data.csv
- Modes: erm / crossvalidation
- Run following on command line mentioning the appropriate mode:
  python adaBoost.py --mode erm

References:

- https://jakevdp.github.io/PythonDataScienceHandbook/04.01-simple-line-plots.html
- https://medium.com/towards-artificial-intelligence/matplotlib-complete-beginners-guide-to-lin e-plots-a436e18d69e4
- https://www.youtube.com/watch?v=7VeUPuFGJHk
- https://www.youtube.com/watch?v=LsK-xG1cLYA&t=1027s
- https://medium.com/@thomascountz/19-line-line-by-line-python-perceptron-b6f113b161f3
- https://towardsdatascience.com/perceptron-learning-algorithm-d5db0deab975
- https://www.youtube.com/watch?v=VRcixOuG-TU