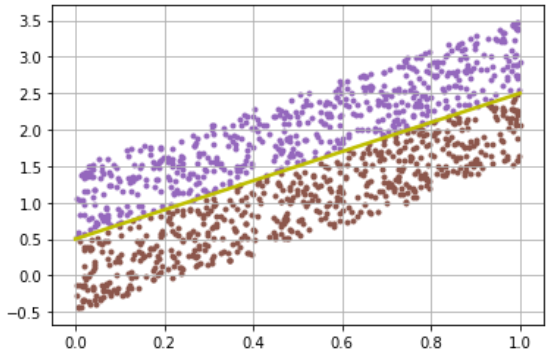
REPORT

I. Running the code: Unzip and extract the homework submission zip file. Open the code directory and in a new command prompt window (Windows)/ new terminal window (Linux/Mac) execute the code by providing dataset and mode as parameters-

1. **python Perceptron.py --dataset linearly-separable-dataset.csv --mode erm**
2. **python Perceptron.py --dataset linearly-separable-dataset.csv --mode nfold**
3. **python Perceptron.py --dataset Breast_cancer_data.csv --mode erm**
4. **python Perceptron.py --dataset Breast_cancer_data.csv --mode nfold**
5. **python Adaboost.py --dataset Breast_cancer_data.csv --mode erm**
6. **python Adaboost.py --dataset Breast_cancer_data.csv --mode nfold**

Linearly Separable Data: Bias = -10.0, Weight vector = [-40.02891027 20.0112814]



II. Performance metrics: Perceptron

Dataset	Mode	Probability of Error
Linearly Separable Dataset	ERM	0.0
Linearly Separable Dataset	10-FOLD	0.0
Breast Cancer Data	ERM	0.24
Breast Cancer Data	10-FOLD	0.17

II. Performance metrics: Adaboost with Decision Stumps

Dataset	Mode	Probability of Error
Breast Cancer Data	ERM	0.076
Breast Cancer Data	10-FOLD	0.083

III. Perceptron Heuristics: For the perceptron algorithm, when 'linearly-separable-dataset' is executed, the learning algorithm is executed with accuracy 100%, in both ERM and 10-fold modes, which is evidence that the data is indeed linearly separable.

While running the same algorithm over the breast_cancer_dataset, which is not linearly separable, the best probability of error in ERM mode was 0.13, when the number of epochs was 1000 or higher. This non-convergence of error to 0.0 is evidence that not all cases are correctly diagnosed.

III. Adaboost Heuristics: For the AdaBoost algorithm, when the number of epochs was too low, the probability of error was higher. On increasing the number of epochs, training error converged to 0, but there were disconnected patches or islands of similarly labelled data. Thus, the model overfit the data which increased the true error.

To determine the optimum number of epochs, I plotted a graph of T vs ERM and found that the ideal number of epochs was 8. I could've set the number to be greater, but it hampered the performance of the model.

Comparing the Perceptron and Adaboost heuristics, it is evident that Adaboost gives lesser error for non-separable datasets like the Breast Cancer Data, and Perceptron is better for linearly separable datasets.