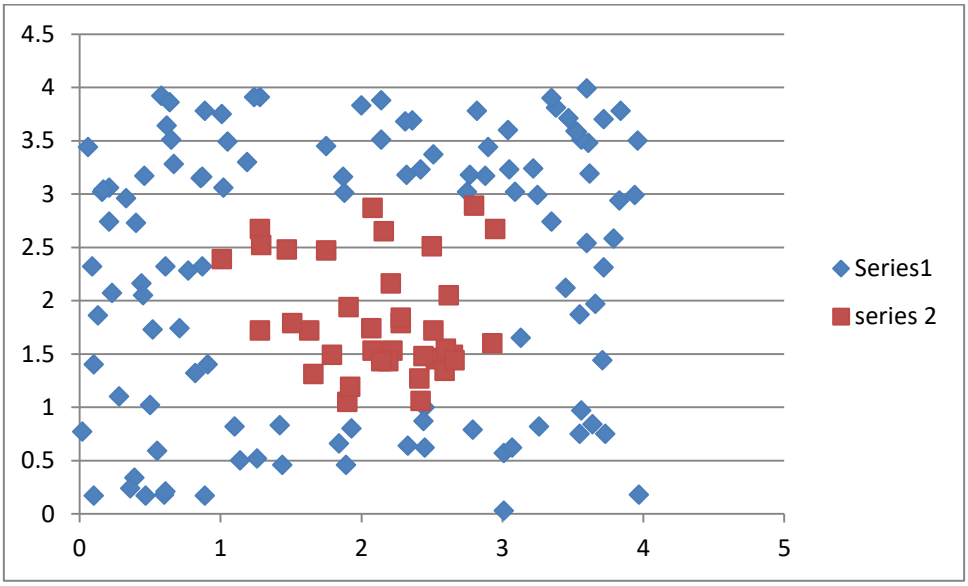
Ariel University

Machine Learning

Homework 3

Take the “rectangle” data set attached. It contains 150 points, separated by a square:



Each pair of points can define a line that passes through the two points. The set of all such lines forms is our hypothesis set, that is set of rules. Implement Adaboost using the above set of rules:

One run of Adaboost is as follows: Split the data randomly into ½ test (T) and ½ train (S). Run Adaboost on S to identify the 8 most important lines h_i and their respective weights α_i. For each k=1,…,8, compute the empirical error of the function H_k on the training set, and the true error of H_k on the test set:

$$H_k(x) = sign(\sum_{i=1}^{k} \alpha_i h_i(x))$$

$$\bar{e}(H_k) = \frac{1}{n}\sum_{x_i \in S}[y_i \neq H_k(x)]$$

$$e(H_k) = \frac{1}{n}\sum_{x_i \in T}[y_i \neq H_k(x)]$$

Execute 100 runs of Adaboost, and report $\bar{e}(H_k)$ and $e(H_k)$ averaged over the 100 runs. Hand in code in python (write at the top which version of python you’re using) and printouts of the value of each H_k for each dataset (total: 16 values). Answer the following:

1. Analyze the behavior of Adaboost on train and test. Do you see any exceptional behavior? Explain.

2. Do you see overfitting? Explain.