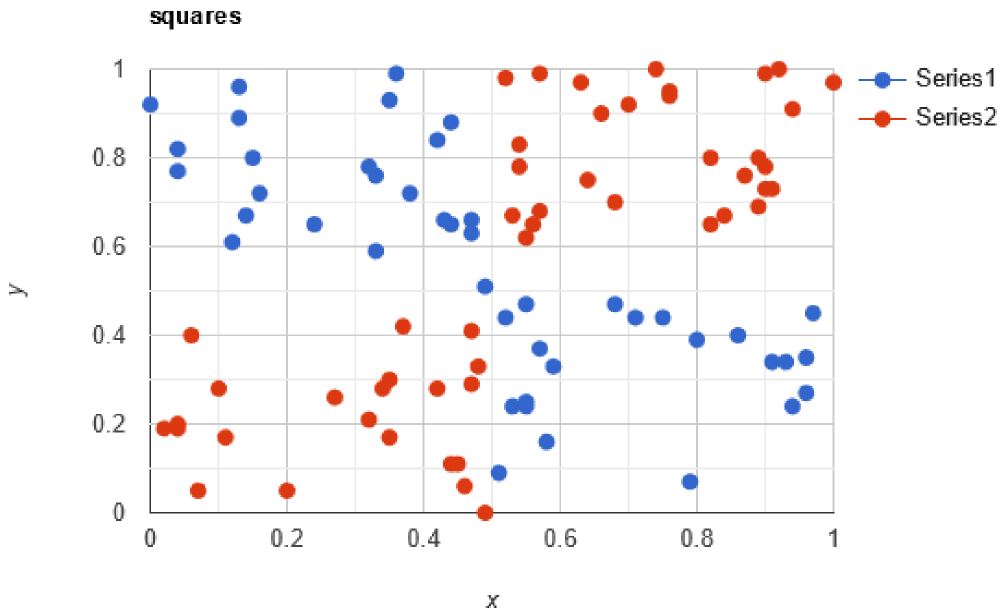
Ariel University

Machine Learning

Homework 3

For the problems below, hand in python code, and also hand in the answers in a separate file.

Problem 1. The “squares” data set contains 100 2-dimensional points, where the last column in the file is the labels:



Each pair of points define a line that passes through them. The set of all such lines is our set of rules. Implement Adaboost using these rules.

One run of Adaboost is as follows: Split the data randomly into ½ test (T) and ½ train (S). Use the points of S (not T) to define the hypothesis set of lines. 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 S, and the true error of H$_k$ on T:

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

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

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

Execute 50 runs of Adaboost, and report $\bar{e}(H_k)$ and $e(H_k)$ for each k, averaged over the 50 runs. Hand in printouts of the values of $\bar{e}(H_k)$ and $e(H_k)$ (total: 16 values). Answer the following:

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

2. Do you see overfitting? Explain.

Problem 2. Now run the above algorithm using circles instead of lines. A circle is defined by two points of S: One point is the center, and the radius is the distance from the center to the second point. In addition, a circle can have two directions: inside is red and outside is blue, or inside is blue and outside is red.

As in Problem 1, hand in printouts of the values of $\bar{e}(H_k)$ and $e(H_k)$ (total: 16 values). Also answer the two questions from Problem 1, but now for circles. And answer the following:

3. How do the results from Problem 1 and Problem 2 differ? Elaborate.