Ariel University

Machine Learning

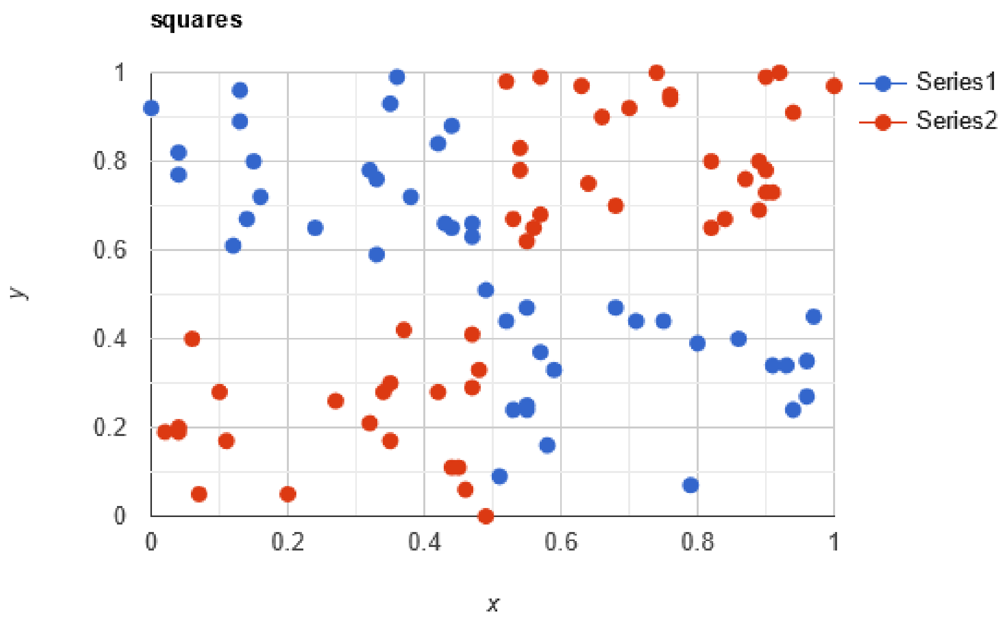
Homework 3

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

Problem 1. Write an implementation of Winnow, and run it on the points in the file "winnow_vectors." Each row contains a 20-dimensional vector, and the final coordinate is the label. You should go through the list of points in order. If the algorithm makes a mistake on a point, update the weights, and go back to the first point on the list. Stop when the algorithm has gone through all points without a mistake.

How many mistakes were made by the algorithm?

Problem 2. 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:

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

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

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

Execute 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.