**Assignment 4 – AdaBoost**

**Question 1.** Use 3 weak learners (decision stumps) to design an AdaBoost classifier trained with the following data (with Play Baseball Today as the target attribute) showing calculations worked out by hand. Provide the final decision function for the classifier.

|  |  |
| --- | --- |
| Number of Previous Days of Rain | Play Baseball Today |
| 0 | Yes |
| 1 | Yes |
| 2 | Yes |
| 3 | No |
| 4 | No |
| 5 | No |
| 6 | Yes |
| 7 | Yes |
| 8 | Yes |
| 9 | No |

**Question 2.** Using the dataset included in the file Tennis.txt, classify the data using the Python implementation of AdaBoost as found in the textbook. The data file is comma delimited. The data includes the results of 114 tennis matches from the Wimbledon 2013. The first column indicates which player won – either player 1 (indicated by 1) or player 2 (indicated by -1). The remaining columns indicate the following statistics measured during the sets. These columns (from 2 to 33) include:

|  |  |
| --- | --- |
| Column | Attribute |
| 2 | Number of sets won in for player 1 |
| 3 | Number of sets won for player 2 |
| 4 | First Serve Percentage for player 1 (Real Number) |
| 5 | First Serve Won by player 1 (Real Number) |
| 6 | Second Serve Percentage for player 1 (Real Number) |
| 7 | Second Serve Won by player 1 (Real Number) |
| 8 | Aces won by player 1 (Numeric-Integer) |
| 9 | Double Faults committed by player 1 (Numeric-Integer) |
| 10 | Winners earned by player 1 (Numeric) |
| 11 | Unforced Errors committed by player 1 (Numeric) |
| 12 | Break Points Created by player 1 (Numeric) |
| 13 | Break Points Won by player 1 (Numeric) |
| 14 | Net Points Attempted by player 1 (Numeric) |
| 15 | Net Points Won by player 1 (Numeric) |
| 16 | Set 1 result for Player 1 (Numeric-Integer) |
| 17 | Set 2 Result for Player 1 (Numeric-Integer) |
| 18 | Set 3 Result for Player 1 (Numeric-Integer) |
| 19 | First Serve Percentage for player 2 (Real Number) |
| 20 | First Serve Won by player 2 (Real Number) |
| 21 | Second Serve Percentage for player 2 (Real Number) |
| 22 | Second Serve Won by player 2 (Real Number) |
| 23 | Aces won by player 2 (Numeric-Integer) |
| 24 | Double Faults committed by player 2 (Numeric-Integer) |
| 25 | Winners earned by player 2 (Numeric) |
| 26 | Unforced Errors committed by player 2 (Numeric) |
| 27 | Break Points Created by player 2 (Numeric) |
| 28 | Break Points Won by player 2 (Numeric) |
| 29 | Net Points Attempted by player 2 (Numeric) |
| 30 | Net Points Won by player 2 (Numeric) |
| 31 | Set 1 result for Player 2 (Numeric-Integer) |
| 32 | Set 2 Result for Player 2 (Numeric-Integer) |
| 33 | Set 3 Result for Player 2 (Numeric-Integer) |

For this question, use the first 100 data points as training and test on the remaining 14 data points. Provide the error rate that you have for the 14 data points (# of errors / # of test points).

Note: There is an error in the Python code from the website: You must change the last line in the function def adaBoostTrainDS(dataArr,classLabels,numIt=40):

from

return weakClassArr,aggClassEst

to

return weakClassArr