CS685: Data Mining

Assignment 2 (100 marks)

Due on: 5th November, 2018, 10:00pm

Aim

The aim of this assignment is figuring out classes from data.

Game

The value y is a sum of three functions f_i of x_i:

$$y = f_1(x_1) + f_2(x_2) + f_3(x_3)$$

The values of x_i are between 0.1 and 1.0 and is rounded to 4 decimal places.

The class c of a tuple $\langle x_1, x_2, x_3 \rangle$ depends on the corresponding y value in the following manner. The total range of y is divided into 10 equal parts, each part being called a *class*. The class labels are 0 to 9. However, there is no ordinal relationship among them. In other words, class 0 does not necessarily represent the lower most range of y with class 1 representing the next range, etc. Thus, although the class labels are consistent, they are shuffled.

Noise

The classes generated for the training data are injected with a little amount of noise as well. Thus, with a probability of 0.05, the true class label of a tuple is randomised to be any of the remaining 9 classes.

Functions

The function f_i is described as $f_i = c_i \times g_i(x_i)$. So,

$$y = c_1 \times g_1(x_1) + c_2 \times g_2(x_2) + c_3 \times g_3(x_3)$$

The function g_i is either logarithmic or a polynomial of x_i. The logarithmic form is $\log_b(x_i + l)$ where l is one of $\{0.1, 0.2, \ldots, 0.9, 1.0\}$ and $b \in \{2, 3, 5\}$. The polynomial form is x_i^{+d} or x_i^{-d} where d is one of $\{+1.1, +1.2, \ldots, +2.9, +3.0\}$.

The coefficients $c_i \neq 0$ are between $+10$ to $+100$ at steps of 0.1. The coefficients can be negative in the same manner, i.e., $\{-100.0, -99.9, \ldots, -10.1, -10.0\}$.

Data

The initial data that is given to you contains 90 points with their associated classes.

Batch and Penalty

If the given amount of data is not enough, you can ask for more batches of data in two forms from the website http://hrishirt.cse.iitk.ac.in/cs685/assign2/. You should login using the username and password supplied to you.

In the first form, you can ask for more tuples of the form $\langle x_1, x_2, x_3 \rangle$. Each time you ask for a batch of tuples, 2.5 marks will be penalised. In each batch, you can ask up to 5 tuples.

In the second form, you can ask for a batch of 5 *random* tuples for a particlar class. Asking for such a batch penalises 2.5 marks.

The first 2 batches that you ask are *free*.

Submission

The file submitted should be an executable/script in a Linux software freely available. It should input a test file containing exactly 100 lines in the following format: x_1, x_2, x_3. It should output 100 class labels, each on a single line in the format c.

You also need to submit the code corresponding to it.

Grading

These will be matched with the *truth file* withheld from you.

The marks obtained is

$$marks = accuracy^2 \times (100 - penalty)$$

So, if you have a penalty of 10 marks and got an accuracy of 80%, your final marks is $0.8 \times 0.8 \times (100 - 10) = 57.60$.