CS 613 - Machine Learning

Avani Jain

Assignment 2 - Classification

PART-1 [Theory Part]

1:

A - Sample Entropy, H(Y) from this training data (using log base 2)

The total number of observations $T = 3+4+4+1+1+3+5 = 21$

The total number of observations in class $Y_+ = 3+4+4+1 = 12$

The total number of observations in class $Y_- = 1+3+5 = 9$

$$H(Y) = -\frac{4}{7} * \log_2 \frac{4}{7} - \frac{3}{7} * \log_2 \frac{3}{7} = 0.9851 \tag{1}$$

$$H(Y) = 0.9851 \tag{2}$$

B - Information Gains for branching on variables x1 and x2

$$H(x_1) = \frac{8}{21}(-\frac{7}{8} * \log_2 \frac{7}{8} - \frac{1}{8} * \log_2 \frac{1}{8}) + \frac{13}{21}(-\frac{5}{13} * \log_2 \frac{5}{13} - \frac{8}{13} * \log_2 \frac{8}{13}) = 0.802 \tag{3}$$

$$H(x_1) = 0.802 \tag{4}$$

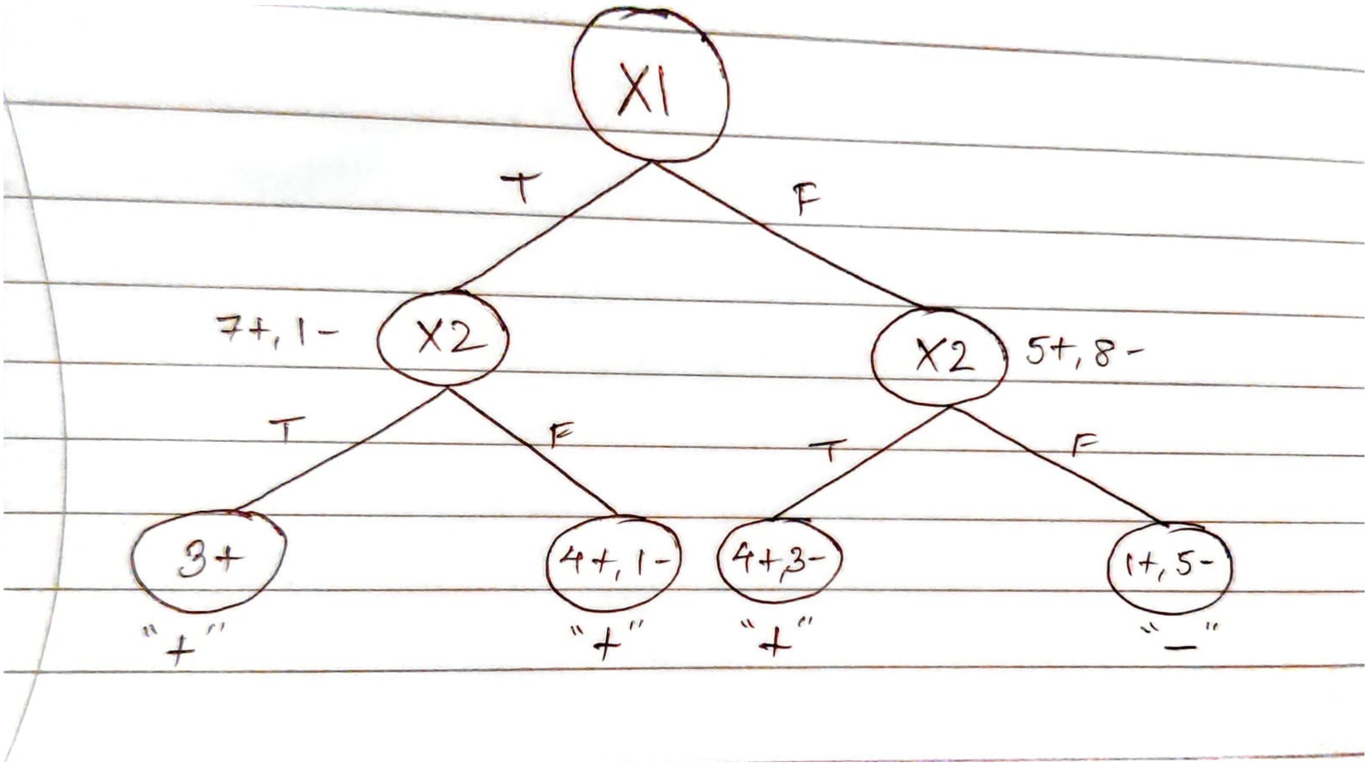
$$InformationGain(x_1) = H(Y) - H(x_1) = 0.1831 \tag{5}$$

$$H(x_2) = \frac{10}{21}(-\frac{7}{10} * \log_2 \frac{7}{10} - \frac{3}{7} * \log_2 \frac{3}{7}) + \frac{11}{21}(-\frac{5}{11} * \log_2 \frac{5}{11} - \frac{6}{11} * \log_2 \frac{6}{11}) = 0.9404 \tag{6}$$

$$H(x_2) == 0.9404 \tag{7}$$

$$InformationGain(x_2) = H(Y) - H(x_2) = 0.0447 \tag{8}$$

C - Decision Tree that would be learned by the ID3 algorithm without pruning from this training data



2:

A - Class Priors:-

$$P(A = Yes) = 0.6, P(A = No) = 0.4 \tag{9}$$

B - Gaussian Parameters (Gaussian Naive Bayes classification)

Let μ_{x1} be mean of "number of chars" feature x_1.

$$\text{We have; } \mu_{x1} = \frac{216+69+302+60+393}{5} = 208$$

Let μ_{x2} be mean of "Average Word Length" feature x_2

$$\text{We have; } \mu_{x2} = \frac{5.68+4.78+2.31+3.16+4.2}{5} = 4.026$$

Let σ_{x1} and σ_{x2} be standard division of x_1 and x_2.

$$\text{We have; } \sigma_{x1} = 145.215 \text{ and } \sigma_{x2} = 1.326$$

We can standardize x_1, x_2 and split into class "yes" and "no" as $x_{1,y}$, $x_{2,y}$ $x_{1,n}$, $x_{2,n}$

We have;

$$x_{1,y} = \begin{bmatrix} 0.05509059 \\ -0.95719904 \\ -1.01917595 \end{bmatrix} \; x_{2,y} = \begin{bmatrix} 1.24771393 \\ 0.56878857 \\ -0.65327706 \end{bmatrix}$$

$$x_{1,n} = \begin{bmatrix} 0.64731446 \\ 1.27396994 \end{bmatrix} \; x_{2,n} = \begin{bmatrix} -1.29448434 \\ 0.1312589 \end{bmatrix}$$

Therefore, we can calculate mean and standard division of each matrix as

$$\mu_{x1,y} = -0.640 \text{ and } \sigma_{x1,y} = 0.603$$
$$\mu_{x2,y} = 0.3877 \text{ and } \sigma_{x2,y} = 0.963$$
$$\mu_{x1,n} = 0.9606 \text{ and } \sigma_{x1,n} = 0.443$$
$$\mu_{x2,n} = -0.5816 \text{ and } \sigma_{x2,n} = 1.008$$

C - Predict Classification

Note:- Test Features Standardized before computing pdf

$$P(A = yes|Chars = 242, A.W.L = 4.56) = P(A = yes).p(Chars = 242|N(\mu_{yes1}, \sigma_{yes1}).p(Chars = 4.56|N(\mu_{yes2}$$

$$P(A = yes|Chars = 242, A.W.L = 4.56) = 0.6 * 0.2312 * 0.4141 = 0.0574 \tag{11}$$

$$P(A = no|Chars = 242, A.W.L = 4.56) = P(A = no).p(Chars = 242|N(\mu_{no1}, \sigma_{no1}).p(Chars = 4.56|N(\mu_{no2}, \sigma_{n}$$

$$P(A = no|Chars = 242, A.W.L = 4.56) = 0.4 * 0.2348 * 0.2457 = 0.0231 \tag{12}$$

$$P(A = yes|Chars = 242, A.W.L = 4.56)_{normalized} = 0.7130 \tag{13}$$

$$P(A = no|Chars = 242, A.W.L = 4.56)_{normalized} = 0.2870 \tag{14}$$

Therefore, Classification = Class - Yes

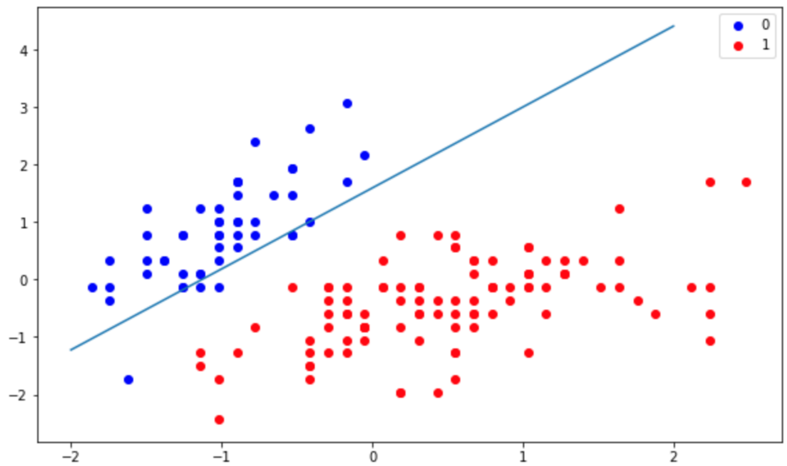
3: A - VALIDATION SET:-

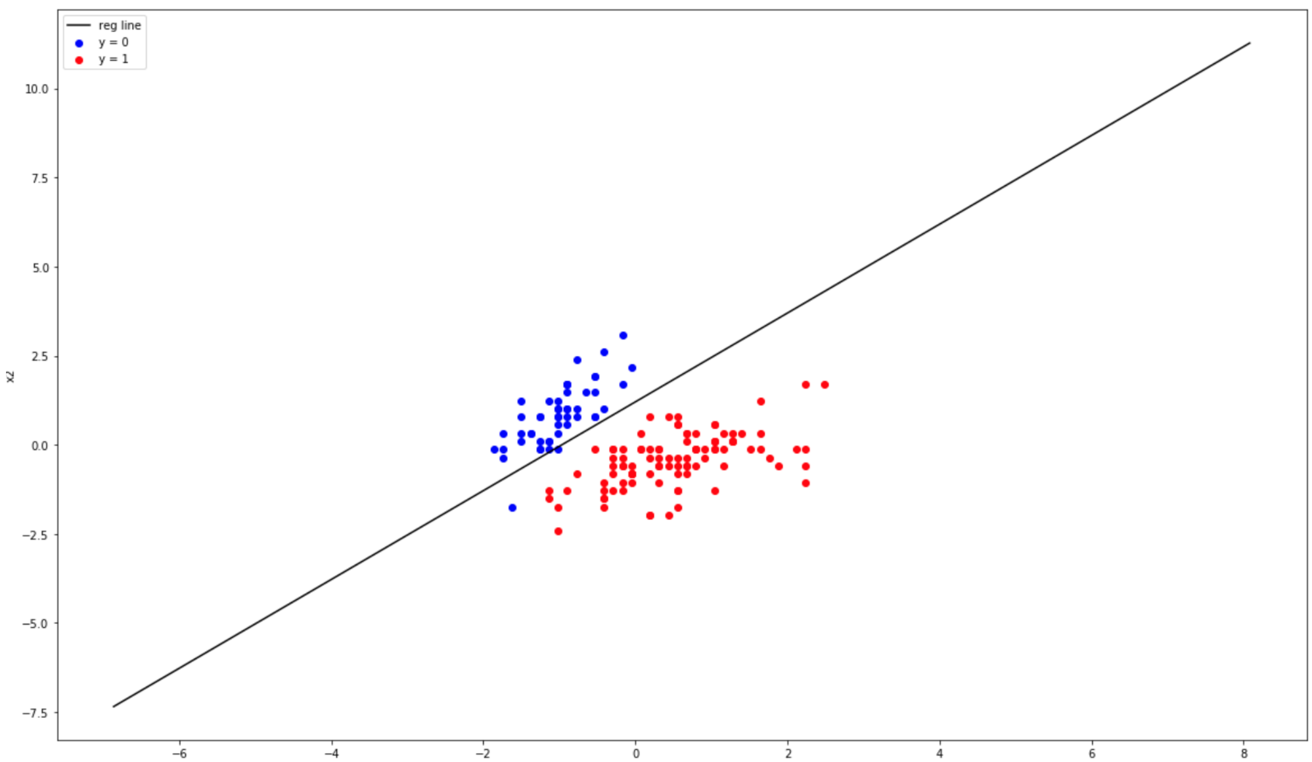
We are using an iterative approach where we are using the training data as the model and test it on the Validation Set. We could set k =1. Then perform KNN on the Validation Set with value k. We expect that the overall error on the Validation Set should decrease with increasing k values i.e. k =1, 2.. etc and then reach a locally optimal point after which the error starts increasing which is a good estimate the user-defined parameter k for the model. This k-value may not be the globally optimum k value depending on the nature of the data.

1 Part 2

1. $\theta = \begin{bmatrix} 0.66666667 & 0.3100605 & -0.24891998 \end{bmatrix}$

2. Plot my method VS sklearn method





2 Part 3

1. Model evaluation statistic

   (a) Precision = 0.63414

(b) Recall = 0.96141
(c) F-measure(F1) = 0.76421
(d) Accuracy = 75.94524

3 Part 4

1. Model evaluation statistic

    (a) Precision = 0.86994
    (b) Recall = 0.87280
    (c) F-measure(F1) = 0.87137
    (d) Accuracy = 84.6805

4 Part 5

1. Model evaluation statistic

    (a) Precision = 0.833009
    (b) Recall = 0.76063
    (c) F-measure(F1) = 0.79518
    (d) Accuracy = 84.23357