CS383 Part#4 theory

Shivansh Suhane
Winter 2020

1. X1, X2 -> Y

   a. The entropy formula is given in homework 1:

   $$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

   H = -(12/21) * log2 (12/21) – 9/21 * log2 (9/21)
   H = 0.5239 + 0.4613
   H = 0.9852

   b. In this question, we first classify the observations per feature based on the True or false. +T is for example, the count of Trues matching a + on Y.

   For feature 1,

+T = 7	4+3
-T = 1	1
+F = 5	4+1
-F = 8	5+3

   For feature 2,

+T = 7	4+3
-T = 3	3
+F = 5	5
-F = 6	5+1

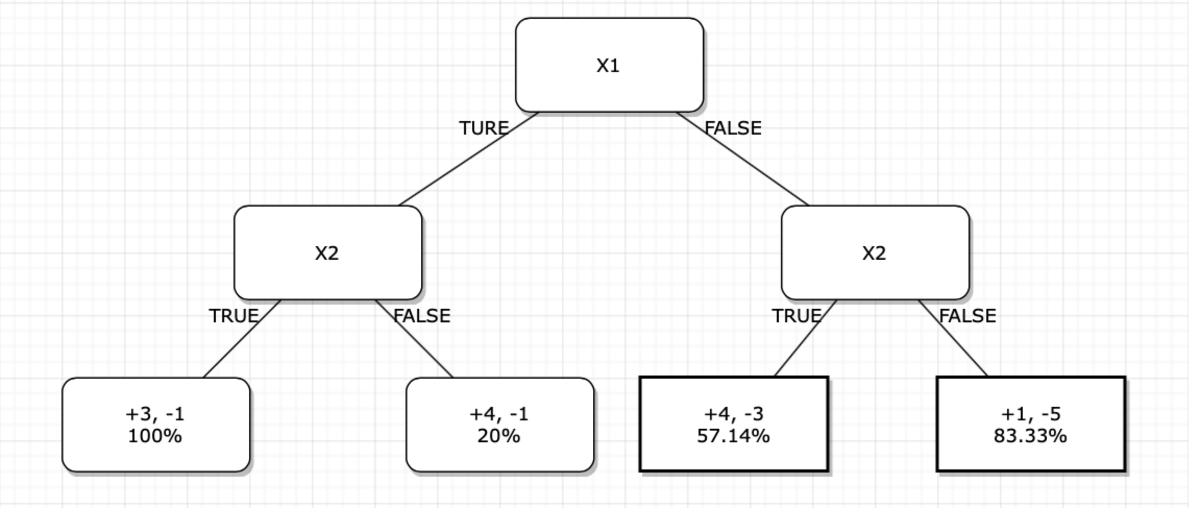
   Now we use the formula,

   $$E(S) = \sum_{i=1}^{c} - p_i \log_2 p_i$$

   R = (8/21 * (-7/8 log2 (7/8) + -1/8(log2(1/8)) + (13/21 * (-5/13 * (-513 log2 (5/13 + -8/13 log2(8/13)))
   <u>= 0.8021</u>

   R = (10/21 * (-7/10 log2(7/10 -3/10 log2(3/10))) + (11/21* (-5/21 * (-5/11 log(5/11) -6/11 log2(6/11)))
   <u>= 0.9403</u>

c. Here’s the decision tree



2. Character count questions:
   a. The priors are:

      P(A) = 3/5

      P(!A) = 2/5

   b. In this question, we try to find the statistics to perform a naïve Bayes classification. We firstly standardize the data for each feature, then find out the parameters for Gaussian for each feature: Mean for num of Chars for YES and NO, Mean of Avg. Word length for YES and NO, Standard Deviation of No. of Chars for YES and NO, and Standard Deviation of Avg. Word Length for YES and NO.

Standardized Data for Number of Chars:

Formulas:

Standardization:

$$z = \frac{x-\mu}{\sigma}$$

with mean:

$$\mu = \frac{1}{N}\sum_{i=1}^{N}(x_i)$$

and standard deviation:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$$

Z(#) = X-mean / Dev

X=[216,
   69,
   302,
   60,
   393]

mean = 208
dev = root(0.25(5.68-4.026)^2 * (4.78-4.026)^2 * ……………… * (4.2-4.026)^2)

Z(char_count) = [0.055,
-0.957,
0.647,
-1.019,
1.273]
This is the standardized feature.

Standardized Data for Avg. word length:
Z(avg_WL) = X-mean / Dev

Dev = root((1/N-1) summation: 1->N(Xi-X)^2)
mean = 4.026

Z(char_ avg) = [1.247,
0.567,
-1.294,
-0.653,
0.131]

This is the standardized feature 2.

Mean for number of Chars:
Mean(A=yes) = 1/3 sum(0.055, -0.957, -1.019) = -0.06404
Mean(A=no) = 1/2 sum(0.6473+1.2739) = 0.9606

Mean for word length:
Mean(A=yes) = 1/3 sum(1.247+ 0.568+-0.653) = 0.3877
Mean(A=no) = 1/2 sum(-1.294+0.131)= -0.5816

Standard Deviation for number of Chars:
Mean(A=yes) = root((1/N-1) summation: 1->N(Xi-X)^2)= 0.6031
Mean(A=no) = root((1/N-1) summation: 1->N(Xi-X)^2)= 0.4431

Standard Deviation for Avg. word length:
Mean(A=yes) = root((1/N-1) summation: 1->N(Xi-X)^2)= 0.9633
Mean(A=no) = root((1/N-1) summation: 1->N(Xi-X)^2)= 1.0081

c. We need to find gaussian with standardized data.

$$P(x_k|y) \propto \frac{1}{\sigma_k\sqrt{2\pi}}e^{-\frac{(x_k-\mu_k)^2}{2\sigma_k^2}}$$

Passed Feature values =[242, 4.56]
We standardize these values to get [0.234,0.403]

Then, we find probability for A=yes and A=no

Y = Yes
Mean = -0.640, dev = 0.603, G = gaussian function F1
Mean = 0.3877, dev = 0.9633, G = gaussian function F2
P(Y=A) = P(A=YES) * G(0.234,-0.640, 0.603) * G(0.403,0.3877,0.9633)
= 0.6*0.2312*0.4142= 0.057441

Y = No
Mean = 0.961, dev = 0.4431, G = gaussian function F1
Mean = -0.582, dev = 1.0081, G = gaussian function F2
P(Y!=A) = P (A=No)G(0.234, 0.9612,0.4431) G(0.403,-0.582, 1.0081)
P = 0.4 * 0.234 * 0.245 = 0.022932

Since the probability of Y = A is greater than Y!=A, the student gets an A.

Coding assignments statistics:

3. Run statistics for Naïve Bayes Classifier:
Run Statistics:
- **a. Accuracy :0.8044328552803129**
- **b. Recall: 0.9633333333333334**
- **c. Precision: 0.6752336448598131**
- d. **f-measure: 0.7939560439560439**

4. Run statistics for Logistic Regression Classifier:
Run Statistics:
- **a. Accuracy :0.923728813559322**
- **b. Recall: 0.8675496688741722**
- **c. Precision: 0.9340463458110517**
- **d. f-measure: 0.8995708154506438**