Avik Bag
Professor Matt Burlick
CS 383 – Assignment 5
2nd March 2017

Theory question

Question 1

Table 1: Raw Data

Y	X_1	X_2	Count
+	T	T	3
+	T	F	4
+	F	T	4
+	F	F	1
-	T	T	0
-	T	F	1
-	F	T	3
-	F	F	5

From the above raw data, we can see that there are a total of 21 data entries, 12 of them being (+) in the Y column, while 9 of them are (-).

Sample entropy on Y, H(Y)

$$H\big(P(v_+), P(v_-)\big) = (-P(v_+)\log_2 P(v_+)) + (-P(v_-)\log_2 P(v_-))$$

$$\Rightarrow H(Y) = \left(-\frac{12}{21}\log_2\frac{12}{21}\right) + \left(-\frac{9}{21}\log_2\frac{9}{21}\right)$$

$$= 0.9852$$

Information Gain on x_1

Table 2: IG on x1

When True	When False
7 Instance of (+)	5 instance of (+)
1 instance of (-)	8 instance of (-)
Total: 8	Total: 13

$$IG(x_1) = H(Y) - \left[\frac{8}{21}H\left(\frac{7}{8},\frac{1}{8}\right) + \frac{13}{21}H\left(\frac{5}{13},\frac{8}{13}\right)\right]$$

$$\Longrightarrow 0.9852 - \left[\frac{8}{21}H\left(\frac{7}{8},\frac{1}{8}\right) + \frac{13}{21}H\left(\frac{5}{13},\frac{8}{13}\right)\right] = \mathbf{0.1832}$$

Information Gain on x_2

Table 3: IG on x2

When True	When False
7 Instance of (+)	5 instance of (+)
3 instance of (-)	6 instance of (-)
Total: 10	Total: 11

$$IG(x_1) = H(Y) - \left[\frac{10}{21}H\left(\frac{7}{10},\frac{3}{10}\right) + \frac{11}{21}H\left(\frac{5}{11},\frac{6}{11}\right)\right]$$
$$\Longrightarrow 0.9852 - \left[\frac{10}{21}H\left(\frac{7}{10},\frac{3}{10}\right) + \frac{11}{21}H\left(\frac{5}{11},\frac{6}{11}\right)\right] = \mathbf{0.0448}$$

Hence, there is a higher information gain on variable x_1, and that will be the dataset that will be used as the decision node.

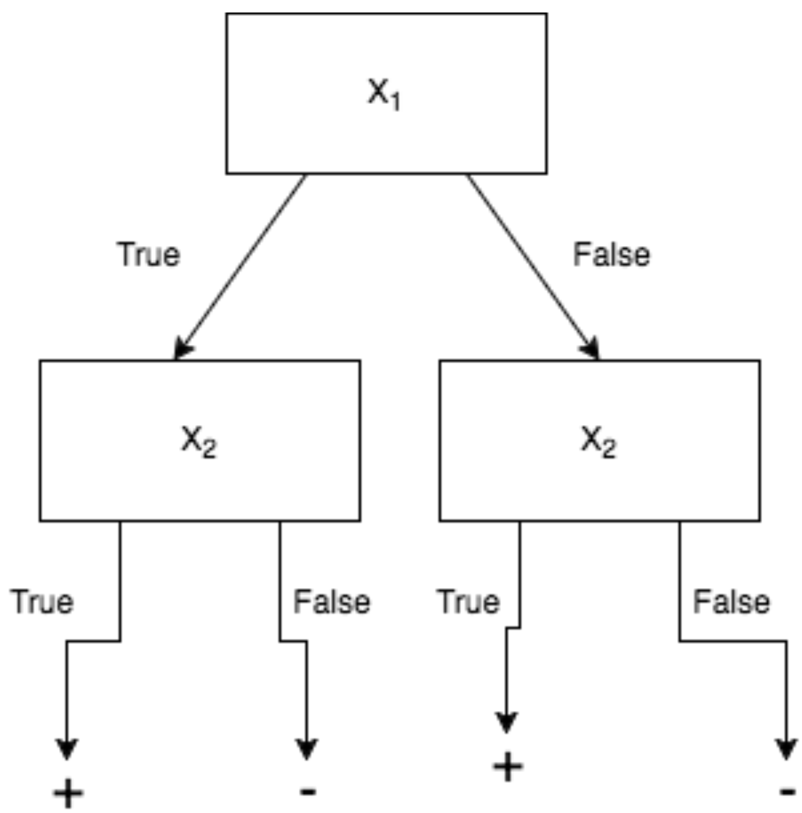

Figure 1: ID3 Decision Tree

Question 2

a) Class priors

$$P(A = Yes) = \frac{3}{5}$$

$$P(A = No) = \frac{2}{5}$$

b) Gaussian parameters.

Table 4: Raw data

	Characters	Word Length	Give an A (1=yes, 0=no)
	216	5.68	1
	69	4.78	1
	302	2.31	0
	60	3.16	1
	393	4.2	0
Mean	208	4.0260	
Std	145.2154	1.3256	

Standardized Data

Table 5: Standardized data

Characters	Word Length	Give an A (1=yes, 0=no)
0.055091	1.2477	1
-0.9572	0.56879	1
0.64731	-1.2945	0
-1.0192	-0.65328	1
1.274	0.13126	0

Split the data into groups (Yes/No)
Data for give a yes

Table 6: Mean and Std for group (yes)

	Characters	Word Length	Give an A (1=yes, 0=no)
	0.055091	1.2477	1
	-0.9572	0.56879	1
	-1.0192	-0.65328	1
Mean	-0.6404	0.2877	
Std	0.6031	0.9633	

Data for give a no

Table 7: Mean and Std for group (no)

	Characters	Word Length	Give an A (1=yes, 0=no)
	0.64731	-1.2945	0
	1.274	0.13126	0
Mean	0.9606	-0.5816	
Std	0.4431	1.0082	

The data highlighted in tables 6 and 7 are what is necessary in the Gaussian Naïve Bayes classification.

c) We then use the data generated in the previous part to predict the classification of the test data.

	Character count(c)	Mean word length(l)
Unstandardized	242	4.56
Standardized	0.2341	0.4028

Using the normal pdf distribution,

Classification	Probability of character count given norm pdf with respective mean and std. of classification	Probability of average word length given norm pdf with respective mean and std. of classification
Yes	0.2312	0.4141
No	0.2348	0.2457

Therefore, to calculate the probabilities of classification,

$$P\left(yes \,\middle|\, c = 242 \,\&\&\, l = 4.56\right) = \frac{3}{5} * 0.2312 * 0.4141 = 0.0574$$

$$P\left(no \,\middle|\, c = 242 \,\&\&\, l = 4.56\right) = \frac{2}{5} * 0.2348 * 0.4141 = 0.0231$$

Hence, since the probability of getting a yes is higher, the essay with the given values will get an A (yes).

Statistics for K Nearest Neighbors

```
Precision: 0.899441340782
Recall    : 0.838541666667
F-Measure: 0.867924528302
Accuracy : 0.904109589041
```