Theory

Question 1

a)

H(Y) = $H(\frac{12}{21}, \frac{9}{21}) = H(\frac{4}{7}, \frac{3}{7}) = -\frac{4}{7}log_2(\frac{4}{7}) - \frac{3}{7}log_2(\frac{3}{7}) = \mathbf{0.9852}$

b)

IG(x1) = $H(\frac{p}{p+n}, \frac{n}{p+n}) - E(x1) = H(\frac{12}{21}, \frac{9}{21}) - E(H(x1))$

= $0.9852 + \frac{8}{21}\left(-\frac{7}{8}log_2\left(\frac{7}{8}\right) - \frac{1}{8}log_2\left(\frac{1}{8}\right)\right) + \frac{13}{21}\left(-\frac{5}{13}log_2\left(\frac{5}{13}\right) - \frac{8}{13}log_2\left(\frac{8}{13}\right)\right)$

= 0.9852 – 0.207 – 0.595

= **0.1832**

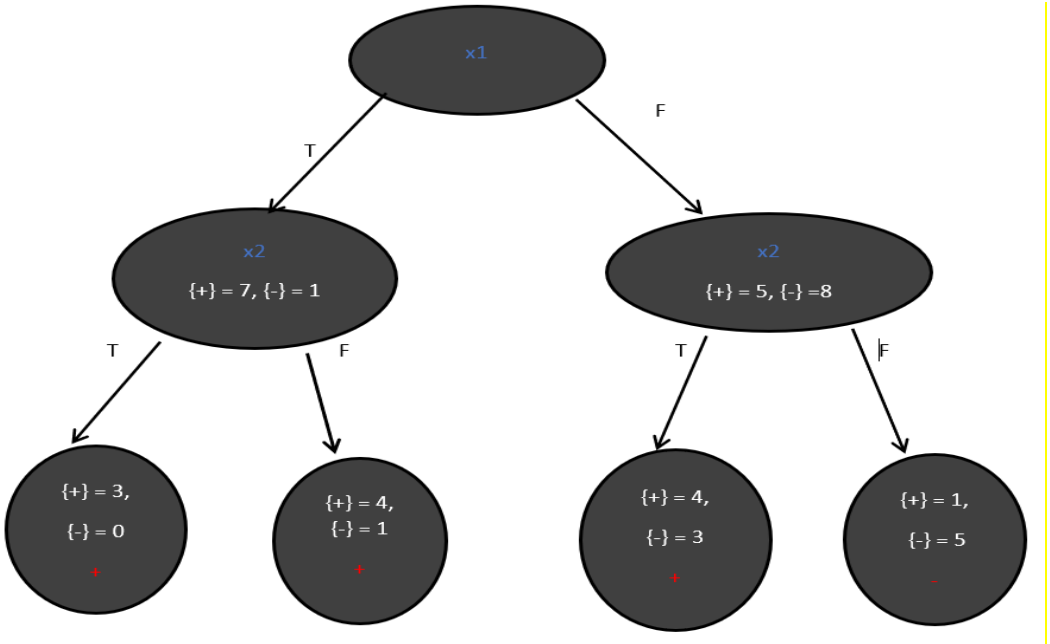
IG(x2) = $H(\frac{p}{p+n}, \frac{n}{p+n}) - E(x2) = H(\frac{12}{21}, \frac{9}{21}) - E(H(x1))$

= $0.9852 + \frac{10}{21}\left(-\frac{7}{10}log_2\left(\frac{7}{10}\right) - \frac{3}{10}log_2\left(\frac{3}{10}\right)\right) + \frac{11}{21}\left(-\frac{5}{11}log_2\left(\frac{5}{11}\right) - \frac{6}{11}log_2\left(\frac{6}{11}\right)\right)$

= 0.9852 – 0.4200 – 0.5207

= **0.0445**

c) The leaf node class choices are in red.



\+ + + -

CS 383 – Machine Learning – Homework 4 – Rakeen Rouf

Question 2

a)

P(A = Yes) = $\frac{3}{5} = 0.6$

P(A = No) = $\frac{2}{5}$ = 0.4

b)

Let's standardize our features.

# of chars = [216, 69, 302, 60, 393]

Mean of # of chars = $\frac{1}{5}\sum_{i=1}^{5} x_i = 208$

Mean of # of chars = $\frac{1}{5}\sum_{i=1}^{5} x_i = 4.026$

Standard deviation of # of chars = $\sqrt{\frac{1}{5-1}\sum_{i=1}^{5}(x_i - mean\_of\_\#\_of\_chars)^2} = 145.22$

Standard deviation of # of chars = $\sqrt{\frac{1}{5-1}\sum_{i=1}^{5}(x_i - mean\_\text{Average\_word\_length})^2} = 1.3256$

Standardized # of chars = # of chars – mean of # of chars / standard deviation of # of chars

Standardized Average word length = Average word length – mean of Average word length / standard deviation of Average word length

Therefore, our standardized input data set is =

0.0551	1.2477	*yes*
−0.9572	0.5688	*yes*
0.6473	−1.2945	*no*
−1.0192	−0.6533	*yes*
1.2740	0.1313	*no*

Next let's find the necessary parameters

Feature 1 # of characters,

$\mu 1yes$ = −0.6404,$\mu 1no$ = 0.9606

$\sigma 1yes$ = 0.6031,$\sigma 1no$ = 0.4431

Feature 2 – Avg word length

$\mu 2yes$ = 0.3877,$\mu 2no$ = −0.5816

$\sigma 2yes$ = 0.9633,$\sigma 2no$ = 1.0082

c)

Standardized value of test pair of features [242 4.56] using mean and standard deviation of input data= [0.2341 0.4028]

$\boldsymbol{P(A = yes|no.of\ char = 242, avg.word\ len = 4.56)}$ = $P(A = yes) * p(0.2341|N(-0.6404, 0.6031) * p(0.4028|N(0.3877, 0.9633)$ = 0.6 ∗ 0.2312 ∗ 0.4141 = **0.0574**

$\boldsymbol{P(A = no|no.of\ char = 242, avg.word\ len = 4.56)}$ = $P(A = no) * p(0.2341|N(0.9606, 0.4431) * p(0.4028|N(-0.5816, 1.0082)$ = 0.4 ∗ 0.2348 ∗ 0.2457 = **0.0231**

Normalized probabilities:

$\boldsymbol{P(A = yes|no.of\ char = 242, avg.word\ len = 4.56)}$ = $\frac{0.0574}{0.0574 + 0.0231}$ = **0.7130**

$\boldsymbol{P(A = no|no.of\ char = 242, avg.word\ len = 4.56)}$ = $\frac{0.0231}{0.0574 + 0.0231}$ = **0.2870**

This essay should get an A

Naïve Bayes Classifier

```
Precision: 66.990291262135 93%
Recall: 95.83333333333334%
F-measure: 78.85714285714288%
Accuracy: 80.69145466405 74%
```

Fig 1: Classification Statistics for Naïve Bayes Classifier

Logistic Regression

```
Precision: 89.75694444444444%
Recall: 89.75694444444444%
F-measure: 89.75694444444444%
Accuracy: 92.30267449445532%
```

Fig 1: Classification Statistics for Logistic Regression