1. Theory Questions

1. Decision Tree Algorithms/ID3

$$X = \begin{bmatrix} T & T & 3 \\ T & F & 4 \\ F & T & 4 \\ F & F & 1 \\ T & T & 0 \\ T & F & 1 \\ F & T & 3 \\ F & F & 5 \end{bmatrix}, Y = \begin{bmatrix} + \\ + \\ + \\ + \\ - \\ - \\ - \\ - \end{bmatrix}$$

$$\mathbb{E}(\mathrm{H(A)}) = \sum_{i=1}^{k} \frac{p_i + n_i}{p+n} * \boldsymbol{H}\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

$$H\big(P(v_1), P(v_2)\big) = \sum_{i}^{n} (-P(v_i) \log_2 P(v_i))$$

a. What is the sample entropy, H(Y)?

Total Observations: 21
Positive Observations: 12
Negative Observation: 9

$$P(y = +) = \frac{12}{21}, P(y = -) = \frac{9}{21}$$

$$H\left(P\left(\frac{12}{21}\right), P\left(\frac{9}{21}\right)\right) = \left(-\frac{12}{21} * \log_2 \frac{12}{21} - \frac{9}{21} * \log_2 \frac{9}{21}\right) = \mathbf{.98522}$$

b. What are the weighted average entropies of the class labels of the subsets created by variables x1 and x2 (5pts)

a. Entropy Based on X1

When x_1 = T

$$\mathrm{X} = \begin{bmatrix} T & T & 3 \\ T & F & 4 \\ T & T & 0 \\ T & F & 1 \end{bmatrix} \quad Y = \begin{bmatrix} + \\ + \\ - \\ - \end{bmatrix}$$

$P_{x1=T}$ = 7, $N_{x1=T}$ = 1, C_i = 8 P (Y = + | x_1 = T) = $\frac{7}{8}$, P (Y = - | x_1 = T) = $\frac{1}{8}$

When x_1 = F

$$\mathrm{X} = \begin{bmatrix} F & T & 4 \\ F & F & 1 \\ F & T & 3 \\ F & F & 5 \end{bmatrix} \quad Y = \begin{bmatrix} + \\ + \\ - \\ - \end{bmatrix}$$

$P_{x1=T} = 5$, $N_{x1=T} = 8$, $C_i = 13$ P (Y = + | x_1 = T) = $\frac{5}{13}$, P (Y = - | x_1 = T) = $\frac{8}{13}$

$$\mathbb{E}\big(\mathrm{H}(x_1)\big) = \left(\frac{7+1}{8+13} * \left(-\frac{7}{8} * \log_2 \frac{7}{8} - \frac{1}{8} * \log_2 \frac{1}{8}\right)\right) + \left(\frac{5+8}{8+13} * \left(-\frac{5}{13} * \log_2 \frac{5}{13} - \frac{8}{13} * \log_2 \frac{8}{13}\right)\right) = \mathbf{.802123}$$

b. Entropy Based on X2

When x_2 = T

$$\mathrm{X} = \begin{bmatrix} T & T & 3 \\ F & T & 4 \\ T & T & 0 \\ F & T & 3 \end{bmatrix} \quad Y = \begin{bmatrix} + \\ + \\ - \\ - \end{bmatrix}$$

$P_{x2=T} = 7$, $N_{x2=T} = 3$, $C_i = 10$ P (Y = + | x_2 = T) = $\frac{7}{10}$, P (Y = - | x_2 = T) = $\frac{3}{10}$

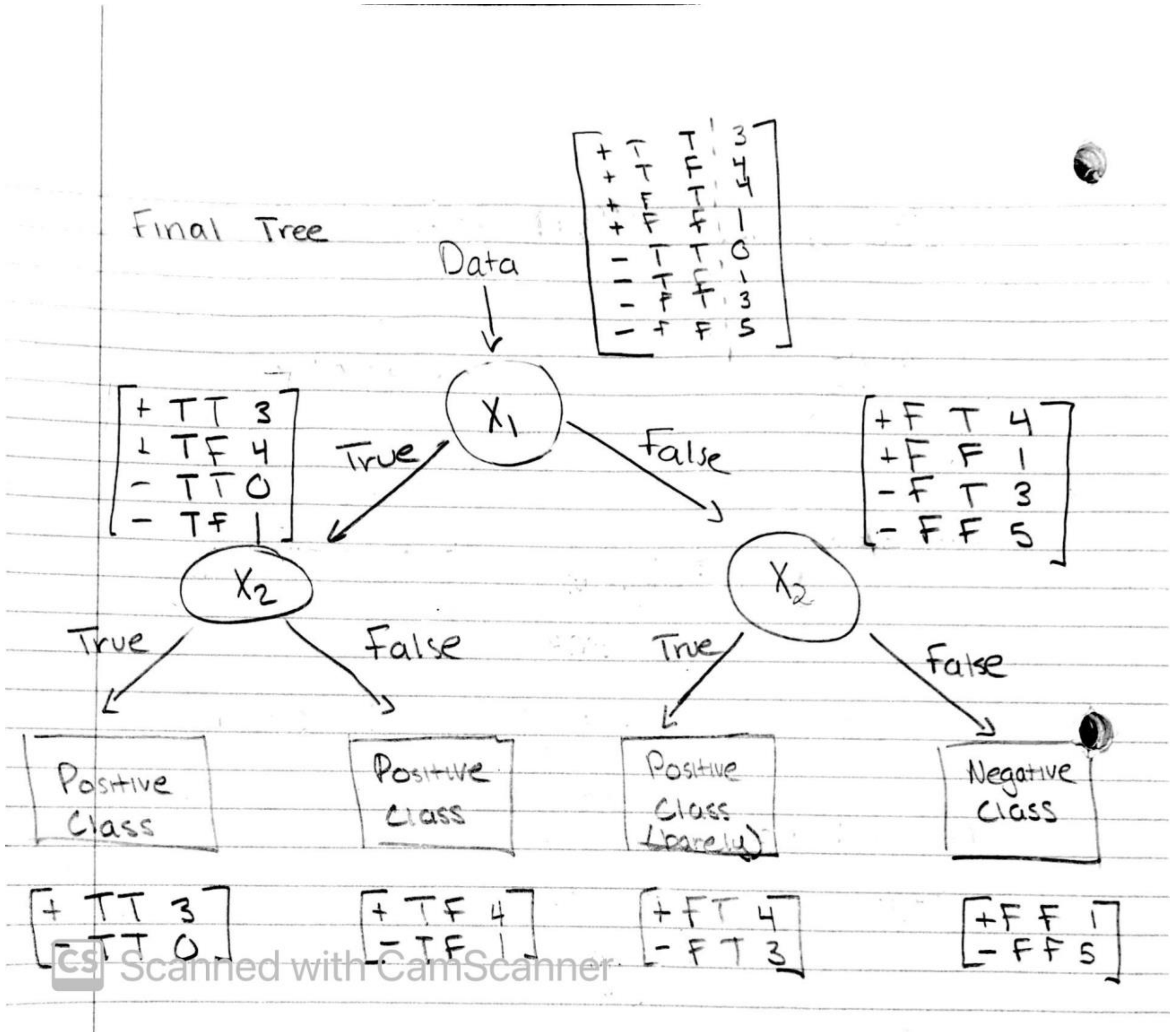
When x_1 = F

$$\mathrm{X} = \begin{bmatrix} T & F & 4 \\ F & F & 1 \\ T & F & 1 \\ F & F & 5 \end{bmatrix} \quad Y = \begin{bmatrix} + \\ + \\ - \\ - \end{bmatrix}$$

$P_{x2=T} = 5$, $N_{x2=T} = 6$, $C_i = 11$ P (Y = + | x_2 = T) = $\frac{5}{11}$, P (Y = - | x_2 = T) = $\frac{6}{11}$

$$\mathbb{E}\big(\mathrm{H}(x_2)\big) = \left(\frac{7+3}{10+11} * \left(-\frac{7}{10} * \log_2 \frac{7}{10} - \frac{3}{10} * \log_2 \frac{3}{10}\right)\right) + \left(\frac{5+6}{10+11} * \left(-\frac{5}{11} * \log_2 \frac{5}{11} - \frac{6}{11} * \log_2 \frac{6}{11}\right)\right) = \mathbf{.940344}$$

c. Since the entropy of splitting up on feature one (.802) is lower than the entropy of splitting on feature two (.940), our first split will be on feature 1.



2. Naïve Bayes Probabilistic Classifier for Essay Grades

$$X = \begin{bmatrix} \#Chars & Average\ Word\ Length \\ 216 & 5.68 \\ 69 & 4.78 \\ 302 & 2.31 \\ 60 & 3.16 \\ 393 & 4.2 \end{bmatrix} \quad Y = \begin{bmatrix} Give\ An\ A \\ Yes \\ Yes \\ No \\ Yes \\ no \end{bmatrix}$$

a. $P(y = Yes) = \frac{3}{5}, P(y = No) = \frac{2}{5}$

b. Standardize

# of Chars:

Mean: 208

Std Dev: 145.215

Average Word Length:

Mean: 4.026

Std Dev: 1.3256

$$stand\_X = \begin{bmatrix} \#Chars & Average\ Word\ Length \\ .05509 & 1.2477 \\ -.957201 & .5688 \\ .64732 & -1.2945 \\ -1.01917 & -.65329 \\ 1.27397 & .13126 \end{bmatrix} \quad Y = \begin{bmatrix} Give\ An\ A \\ Yes \\ Yes \\ No \\ Yes \\ no \end{bmatrix}$$

Give an A = Yes:

$\boldsymbol{\mu_{Yes,x1} = -.640427,\ \sigma_{Yes,x1} = .6031}$

$\boldsymbol{\mu_{Yes,x2} = -.3877,\ \sigma_{Yes,x2} = .9633}$

Give an A = No:

$\boldsymbol{\mu_{No,x1} = .96064,\ \sigma_{No,x1} = .44311}$

$\boldsymbol{\mu_{No,x2} = -.58162,\ \sigma_{No,x2} = 1.005164}$

c. New Data = $[242 \quad 4.56]$
Standardized New Data = $[.2431 \quad .4028]$

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

Prob (Yes) ≈α P (A = Yes) * P ($f_1 = x_1$ | A = Yes) * P ($f_2 = x_2$ | A = Yes)

P (f1 = x1 | A = Yes) = $\frac{1}{.6031*\sqrt{2\pi}} e^{\frac{-(.2311- -.640427)^2}{2*(.6031)^2}}$ = .52827

P (f2 = x2 | A = Yes) = $\frac{1}{.9633*\sqrt{2\pi}} e^{\frac{-(.4028- .3877)^2}{2*(.9633)^2}}$ = .41409

Prob (Yes) ≈α 3/5 * .52827 * .41409 = .13125

Prob (No) ≈α P (A = No) * P ($f_1 = x_1$ | A = No) * P ($f_2 = x_2$ | A = No)

P (f1 = x1 | A = No) = $\frac{1}{.44311*\sqrt{2\pi}} e^{\frac{-(.2311- .96064)^2}{2*(.44311)^2}}$ = .23475

P (f2 = x2 | A = No) = $\frac{1}{1.008164*\sqrt{2\pi}} e^{\frac{-(.4028- -.58162)^2}{2*(1.008164)^2}}$ = .24566

Prob (No) ≈α 2/5 * . 23475* . 24566= .02306

Since the proportional probability of this new observation belonging to the Give An A = Yes Group (P(Yes)) is greater than the proportional probability of this new observation belonging to the Give An A = No Group (P(No)), we should give this new essay an A.

2. Naïve Bayes Classifier

Precision: 0.646
Recall: 0.964
F measure: 0.773
Accuracy: 0.787

3. Logistic Regression Classifier

Precision: 0.921
Recall: 0.868
F measure: 0.894
Accuracy: 0.922