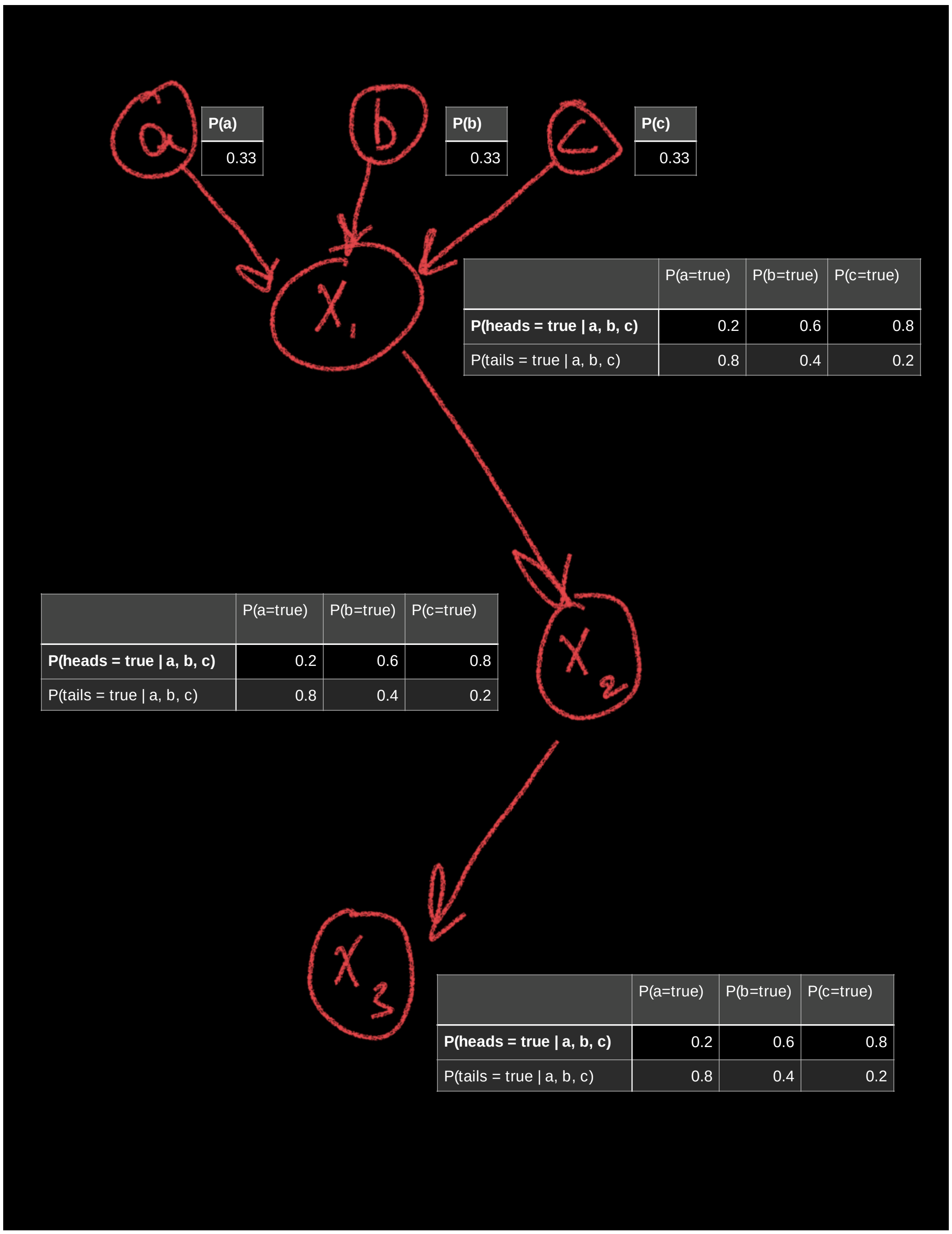
CSCI 4525 Unit 3: Learning Assignment

Patrick Griffin

Q1)

a.



b.

P(a|heads) P(a|heads) P(a|tails) = .2 x .2 x .8 = 0.032

P(b|heads) P(b|heads) P(b|tails) = .6 x .6 x .4 = 0.144

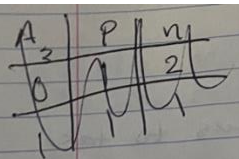
P(c|heads) P(c|heads) P(c|tails) = .8 x .8 x .2 = 0.128

Coin b would most likely have been drawn.

Q2)

Consider the following data set comprised of three binary input attributes (A_1, A_2, and A_3) and one binary output:

Example	A_1	A_2	A_3	Output *y*
x_1	1	0	0	0
x_2	1	0	1	0
x_3	0	1	0	0
x_4	1	1	1	1
x_5	1	1	0	1



y	P	n
	2	3

$$H(y) = B\left(\frac{P}{P+n}\right) = B\left(\frac{2}{2+3}\right) = B\left(\frac{2}{5}\right)$$

$$= B(0.4)$$

rent
ntropy?

$$B(0.4) = -(0.4 \log_2 0.4 + 0.6 \log_2 0.6)$$

$$= -(-0.5288 + (-0.4422))$$

$$= -(-0.971)$$

$$= 0.971 \text{ bits}$$

A_1	P	n
1	2	2
0	0	1

$$\text{Gain} \rightarrow \left(B(0.4) - \sum \frac{P_k + n_k}{p+n} B\left(\frac{P_k}{P_k + n_k}\right)\right)$$

$$- \sum_{k=1} \frac{2+2}{2+3} B\left(\frac{2}{2+2}\right) + \sum_{k=0} \frac{1}{5} B\left(\frac{0}{1}\right)$$

$$= -\left(\frac{4}{5}\left(\frac{1}{2}\log_2 \frac{1}{2} + \frac{1}{2}\log_2 \frac{1}{2}\right) + \frac{1}{5}(0 \log_2 0 + 1 \log_2 1)\right)$$

$$= -\left(4/5\left(-0.5 + -0.5\right) + 1/5\ (0)\right)$$

$$(4/5\ (-1) + 0) = \frac{4}{5}$$

$$0.971 - 0.8 = 0.171 \text{ bits}$$

A_2	P	n
1	2	0
0	0	2

$$\sum_{k=1} \frac{2}{5} B\left(\frac{2}{2}\right) + \sum_{k=0} \frac{3}{5} B\left(\frac{1}{3}\right)$$

$= 0.971 - \left(\frac{2}{5}(1 \log_2 1 + 0) + \frac{3}{5}\left(\frac{1}{3}\log_2\frac{1}{3} + 0.67 \log_2 0.67\right)\right)$

$= 0.971 - (0 + (-0.53 * (-0.39)))$

$= 0.971 - (-0.92) =$ ~~1.891~~

A_2	P	n
1	2	1
0	0	2

look at output y_1 count 1 or 0

$$\sum_{k=1} \frac{P_k + n_k}{P+n} B\left(\frac{P_k}{P_k+n_k}\right)$$

$$\sum_{k=1} \frac{3}{5} B\left(\frac{2}{3}\right) + \sum_{k=0} \frac{2}{5} B(1)$$

$3/5(0.667 \log_2 0.667 + 0.333 \log_2 0.333) + 0$

$3/5(-0.39 + (-0.53))$

$= 0.971 - 0.6(-0.92)$

$= 0.971 - 0.552 = 0.419$ bits ✓

A_3	P	n
1	1	1
0	1	2

$$0.971 - \sum_{k=1} \frac{2}{5} B(0.5) + \sum_{k=0} \frac{3}{5} B\left(\frac{1}{3}\right)$$

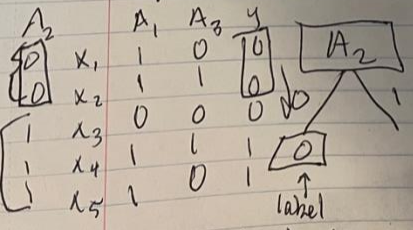
$0.971 - \left(\frac{2}{5}(-(-1)) + \frac{3}{5}(-(0.333 \log_2 0.333 + 0.667 \log_2 0.667))\right)$

$-(-0.53 + -.39)$

$= 0.971 - (0.4 + 0.552)$

$= 0.971 - 0.952 = 0.019$ bits

	A_1	A_3	y
x_1	1	0	0
x_2	1	1	0
x_3	0	0	0
x_4	1	1	1
x_5	1	0	1



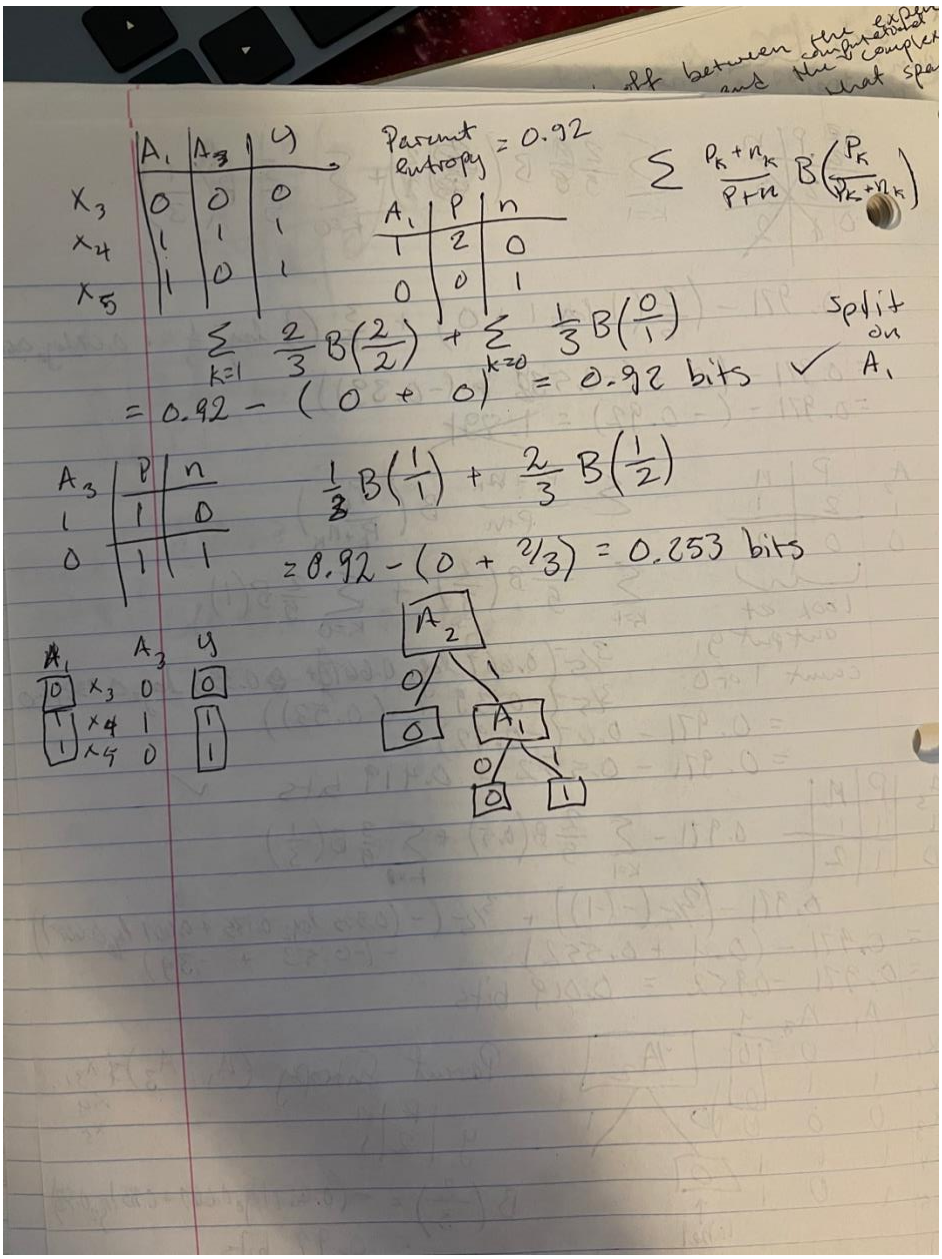
Parent Entropy (A_1, A_3) x_3, x_4, x_5

y	P	n
	2	1

$B\left(\frac{2}{3}\right) = -(0.667 \log_2 0.667 + 0.333 \log_2 0.333)$

$= 0.92$ bits

0's have same classification


A1 A3 y
X3 0 0 0
X4 1 1 1
X5 1 0 1
Parent Entropy = 0.92
A1 P n
1 2 0
0 0 1
Σ (Pk+nk)/(P+n) B(Pk/(Pk+nk))
Σ k=1 2/3 B(2/2) + Σ k=0 1/3 B(0/1)
= 0.92 − (0 + 0) = 0.92 bits
split on A1
A3 P n
1 1 0
0 1 1
1/3 B(1/1) + 2/3 B(1/2)
= 0.92 − (0 + 2/3) = 0.253 bits
A1 A3 y
0 X3 0 0
1 X4 1 1
1 X5 0 1
A2
A1

Q3)

a.) **4 points**: Classify these two unknown flowers using 4 Nearest Neighbor (3NN) and Euclidean distance. For both flowers, you must list their 3 nearest neighbors and the distance between that flower and each neighbor.

Flower X:

Speal Length: 5.4

Speal Width: 3.7cm

Petal Length: 1.5cm

Petal Width: 0.2cm

Flower Y:

Speal Length: 5.9cm

Speal Width: 3.0cm

Petal Length: 5.1cm

Petal Width: 1.8cm

Flower X:

Flower 1 = 0.374

Flower 5 = 0.424

Flower 6 = 0.346

Classified as Iris-setosa.

Flower Y:

Flower 17 = 0.671

Flower 22 = 0.332

Flower 24 = 0.648

Classified as Iris-virginica.

b.) **3 points**: Use k-Means clustering with k= 3 to group all 30 observations into three clusters. Initially, each cluster should contain every third element. So, initially cluster 1 will contain observations 1,4,7,10, etc. Cluster 2 will contain 2,5,8, etc. Repeat the k-means algorithm until it converges. What are the coordinates of the three centroids after convergence? (Note that this is an unsupervised learning problem, so you should ignore the class label when doing this clustering.)

Cluster 1

1	5.1	3.5	1.4	0.2
4	4.6	3.1	1.5	0.2
7	4.6	3.4	1.4	0.3
10	4.9	3.1	1.5	0.1
2	4.9	3	1.4	0.2
5	5	3.6	1.4	0.2
8	5	3.4	1.5	0.2
3	4.7	3.2	1.3	0.2
6	5.4	3.9	1.7	0.4
9	4.4	2.9	1.4	0.2
total	48.6	33.1	14.5	2.2
centroid	4.86	3.31	1.45	0.22

cluster 2

11	7	3.2	4.7	1.4
17	6.3	3.3	4.7	1.6
23	7.1	3	5.9	2.1
26	7.6	3	6.6	2.1
29	6.7	2.5	5.8	1.8
13	6.9	3.1	4.9	1.5
19	6.6	2.9	4.6	1.3
22	5.8	2.7	5.1	1.9
25	6.5	3	5.8	2.2
28	7.3	2.9	6.3	1.8
12	6.4	3.2	4.5	1.5
15	6.5	2.8	4.6	1.5
21	6.3	3.3	6	2.5

24	6.3	2.9	5.6	1.8
30	7.2	3.6	6.1	2.5
total	100.5	45.4	81.2	27.5
centroid	6.7	3.02666667	5.41333333	1.83333333

cluster 3

14	5.5	2.3	4	1.3
20	5.2	2.7	3.9	1.4
27	4.9	2.5	4.5	1.7
18	4.9	2.4	3.3	1
16	5.7	2.8	4.5	1.3
total	26.2	12.7	20.2	6.7
Centroid	5.24	2.54	4.04	1.34