HW 3: Sam Farren

Chapter 2:

2.19

Formulas:

Cosine:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Correlation:

$$\rho_{X,Y} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

$$covariance(x,y) = s_{xy} = \frac{1}{n\text{-}1} \sum_{k=1}^{n} (x_k \text{-} xbar)^2$$

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

Euclidian:

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}.$$

A: x=(1,1,1,1) y=(2,2,2,2)

cosine: (1*2 + 1*2 + 1*2 + 1*2) / (sqrt (1^2 + 1^2 + 1^2 + 1^2) * sqrt(2^2 + 2^2 + 2^2 + 2^2)) = 8 / (sqrt(4)*sqrt(16) = **1**

Correlation: (1/(4-1)) * ((1-1)^2 + (1-1)^2 + (1-1)^2 + (1-1)^2) * ((2-2)^2 + (2-2)^2 + (2-2)^2 + (2-2)^2) = **0**

Euclidian: sqrt((1-2)^2 + (1-2)^2 + (1-2)^2 + (1-2)^2) = sqrt(4) = 2

B: x=(0,1,0,1) y=(1,0,1,0)

cosine: (0*1 + 1*0 + 0*1 + 1*0) / (sqrt (0^2 + 1^2 + 0^2 + 1^2) * sqrt(1^2 + 0^2 + 1^2 + 0^2)) = 0 / (sqrt(2)*sqrt(2) = **0**

Correlation: (1/(4-1)) * ((0-.5)^2 + (1-.5)^2 + (0-.5)^2 + (1-.5)^2) * ((1-.5)^2 + (0-.5)^2 + (1-.5)^2 + (0-.5)^2) = **.333**

Euclidian: sqrt[(0-1)^2 + (1-0)^2 + (0-1)^2 + (1-0)^2] = 2

Jaccard: M11 = 0 -> 0/4 = 0

C: x=(0,-1,0,1) y=(1,0,-1,0)

Cosine: (0*1 + -1*0 + 0*-1 + 1*0) / (sqrt (0^2 + -1^2 + 0^2 + 1^2) * sqrt(1^2 + 0^2 + -1^2 + 0^2)) = 0 / (sqrt(2)*sqrt(2) = **0**

Correlation: (1/(4-1)) * ((0-0)^2 + (-1-0)^2 + (0-0)^2 + (1-0)^2) * ((1-0)^2 + (0-0)^2 + (-1-0)^2 + (0-0)^2) = **1.333**

Euclidian: sqrt[(0-1)^2 + (-1-0)^2 + (0+1)^2 + (1-0)^2] = 2

D: x=(1,1,0,1,0,1) y=(1,1,1,0,0,1)

Cosine: (1*1 + 1*1 + 0*1 + 1*0 + 0*0 + 1*1) / (sqrt (1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2) * sqrt(1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 1^2)) = 3 / (sqrt(4)*sqrt(4) = **.75**

Correlation: (1/(6-1)) * ((1-.66)^2 + (1-.66)^2 … = .25

Jaccard: M11 = 3 M00 = 1 -> M11/(N-M00) = 3/5 = .6

E: x=(2,-1,0,2,0,-3) y=(-1,1,-1,0,0,-1)

Cosine: Cosine: (2*-1 + -1*1 + 0*-1 + 2*0 + 0*0 + -3*-1) / (sqrt (2^2 + -1^2 + 0^2 + 2^2 + 0^2 + -3^2) * sqrt(-1^2 + 1^2 + -1^2 + 0^2 + 0^2 + -1^2)) = 0 / (sqrt(18)*sqrt(4) = **0**

Correlation: Correlation: (1/(6-1)) * ((2-0)^2 + (1-0)^2 … = 0

Chapter 4:

4.2: C0 has 10 C1 has 10

a. GINI = 1 - (10/20)^2 - (10/20)^2 = .5

b. GINI = 1 - Sum(from i=0 to c-1 of) [p(i | t)]^2 = 1 - [(0/1)^2 + (1/1)^2] = 0

c. 10 males GINI (Male with 6 to 4 split) = 1 - [(6/10)^2 + (4/10)^2 = .48
10 Females GINI (Female with 6 to 4 split) = 1 - [(6/10)^2 + (4/10)^2 = .48
Weighted Average: .48

d. Gini(Family Car) = 1 - [(1/4)^2 + (3/4)^2] = .375
Gini(Luxury) = 1 - [(1/8)^2 + (7/8)^2] = .2186
Gini(Sports) = 1 - [(8/8)^2 + (0/8)^2] = 0
Weighted Average: GINI(Car Type) = ((4/20)*.375 + (8/20)*.2186 + 0) = .163

e. Gini(Small) = 1 - [(3/5)^2 + (2/5)^2] = .48
Gini(Medium) = 1 - [(3/7)^2 + (4/7)^2] = .49
Gini(Large) = 1 - [(2/4)^2 + (2/4)^2] = .5
Gini(Extra Large) = 1 - [(2/4)^2 + (2/4)^2] = .5
GINI(Shirt Size) = (5/20)*.48 + (7/20)*.49 + (4/20)*.5 + (4/20)*.5 = .4915

f. Car type would be the best because it's Gini Coefficient is the lower than Gender and Shirt Size.

g. Customer Id has the sole purpose of being unique to each customer and can't help predict any of the classes because of that exact reason. Therefore it shouldn't be used.

4.3

a. There are 4 positive so P(+) = 4/9 and there are 5 negative so P(-)=5/9
Entropy = (-4/9)log(4/9) - (5/9)log(5/9) = .9911

b.

a1:

a1	+	-
T	3	1
F	1	4

Entropy(a1) = (4/9)[-(3/4)log(3/4) - (1/4)log(1/4)] + (5/9)[-(1/5)log(1/5) - (4/5)log(4/5)] = .7616

InformationGain(a1) = .9911-.7616 = .2294

a2	+	-
T	2	3
F	2	2

```
Entropy(a2) = (5/9)[ -(2/5)log(2/5) - (3/5)log(3/5) ] + (4/9)[ -
(2/4)log(2/4) - (2/4)log(2/4) ] = .9839
InformationGain(a2) = .9911-.9839 = .0072

     c.

     Split 1 (.5) :
          > Entropy = -[(4/9)*log(4/9) + (5/9)*log(5/9] = .99107

     Split 2 (2.0):
          <= Entropy = -[(1/1)*log(1/1) + (0)*log(0)] = 0
          > Entropy = -[(3/8)*log(3/8) + (5/8)*log(5/8)] = .95443
          Average: 1/9 * 0 + 8/9 * .95443 = .84839
          Information Gain: .9911 - .8439 = .143

     Split 3 (3.5):
          <= Entropy = -[(1/2)*log(1/2) + (1/2)*log(1/2)] = 1
          > Entropy = -[(3/7)*log(3/7) + (4/7)*log(4/7)] = .98523
          Average: 2/9 * 1 + 7/9 * .98523 = .988512
          Gain: .9911 - .988512 = .00249

     Split 4 (4.5):
          <= Entropy = -[(2/3)*log(2/3) + (1/3)*log(1/3)] = .9183
          > Entropy = -[(2/6)*log(2/6) + (4/6)*log(4/6)] = .9183
          Average: .9183
          Gain: .9911 - .9183 = .0727


     Split 5 (5.5):
          <= Entropy = -[(2/5)*log(2/5) + (3/5)*log(3/5)] = .97095
          > Entropy = -[(2/4)*log(2/4) + (2/4)*log(2/4)] = 1
          Average = 5/9 * .97095 + 4/9 * 1 = .98386
          Gain: .9911 - .98386 = .00724

     Split 6 (6.5):
          <= Entropy = -[(3/6)*log(3/6) + (3/6)*log(3/6)] = 1
          > Entropy = -[(1/3)*log(1/3) + (2/3)*log(2/3)] = .9183
          Average = 6/9 * 1 + 3/9 * .9183 = .9728
          Gain = .9911 - .9728 = .0183

     Split 7 (7.5)
          <= Entropy = -[(4/8)*log(4/8) + (4/8)*log(4/8)] = 1
          > Entropy = -[(0/1)*log(0/1) + (1/1)*log(1/1)] = 0
          Average = 8/9 * 1 + 1/9 * 0 = .8889
          Gain = .9911 - .8889 = .10211
```

Split 8 (8.5):

<= Entropy = -[(4/9)*log(4/9) + (5/9)*log(5/9)] = .99108

> Entropy = -[(0)*log(0) + (0)*log(0)] = 0

Gain = 0

(Wasn't sure whether to include splitting at .5 and 8.5 since it just includes all of the data in one and but I incorporated it anyway.)

d.

a1 produces the best split when comparing information gain.

e.

a1 produces an error rate of 2/9

a2 produces an error rate of 4/9

a1 has a lower error rate so it is better

f. Gini(a1) = (4/9)[1-(3/4)^2 - (1/4)^2] + (5/9)[1 - (1/5)^2 - (4/5)^2] = .3444

Gini(a2) = (5/9)[1-(2/5)^2 - (3/5)^2] + (4/9)[1 - (2/4)^2 - (2/4)^2] = ..4889

Gini(a1) is smaller so a1 produces a better split in the data.

4.8

a. optimistic = 3/10 = .3

Pesimistic = (3 + 4*.5)/10 = .5

Pruning: 4/5 = .8

Chapter 5:

5.5

a. 29 positive and 21 Negative in Data->

R1: 12 postive and 3 negative - >

Expected Frequency of positive: 15*29/50 = 8.7

Expected Frequency of Negative: 15*21/50 = 6.3

Likelihood Ratio: 2 * [12 * log(12/8.7) + 3 * log(3/6.3)] = 4.71

R2: 7 positive and 3 negative

Expected Frequency of positive: 10*29/50 = 5.8

Expected Frequency of Negative: 10*21/50 = 4.2

Likelihood Ratio: 2 * [7 * log(7/5.8) + 3 * log(3/4.2)] = .89

R3: 8 positive and 4 negative
Expected Frequency of positive: 12*29/50 = 6.96
Expected Frequency of Negative: 12*21/50 = 5.04
Likelihood Ratio: 2 * [8 * log(8/6.96 + 4 * log(4/5.04)] = .5472

Since R1 has the highest likelihood value it is the best, and since R3 has the lowest value it is the worst rule

b. Laplace = (f+ + 1)/(n+k)

R1: (12 + 1)/(15 + 2) = 76.47%
R2: (7 + 1)/(10 + 2) = 66.67%
R3: (8 + 1)/(12 + 2) = 64.29%

Since R1 has the highest laplace measure it is the best, and R3 is the worst.

c. M-estimate: (f+ + kp+)/(n+k) with k = 2 and p+ = .58

R1: (12 + 2*.58)/(15 + 2) = 77.41%
R2: (7 + 2*.58)/(10 + 2) = 68%
R3: (8 + 2*.58)/(12 + 2) = 65.43%

Since R1 has the highest m-estimate it is the best, and R3 is the worst.

d. With the R1 examples not being discarded, R2 will be chosen because the accuracy of R2 is higher than R3. (70% > 66.7%)

e. If only the positive examples of R1 are discarded, this will produce new accuracies in R2 and R3 being 70% and 60% respectively. So R2 would be preferred over R3 in this case.

f. With both positive and negative examples being discarded for R1, R2 will get a new accuracy of 70% and R3 will have an accuracy of 75%. Therefore R3 will be preferred.

5.7
a.
P(A=1 | -) = 2/5 = (.4)
P(B=1 | -) = 2/5 = (.4)
P(C=1 | -) = 5/5 = (1.0)
P(A=0 | -) = 3/5 = (.6)
P(B=0 | -) = 3/5 = (.6)
P(C=0 | -)= 0 = (0.0)
P(A=1 | +) = 3/5 = (.6)
P(B=1 | +) = 1/5 = (.2)
P(C=1 | +) = 4/5 = (.8)

P(A=0 | +) = 2/5 = (.4)
P(B=0 | +) = 4/5 = (.8)
P(C=0 | +) = 1/5 = (.2)

b.
(A=0, B=1, C=0)
Positive:
Let K = P(A=0, B=1, C=0)
P(+|A=0, B=1, C=0)
= P(A=0, B=1, C=0|+)*P(+) / P(A=0, B=1, C=0)
= P(A=0|+) P(B=1|+) P(C=0|+)*P(+) / K
= .4*.2*..2*.5/K = .008/K

Negative:
P(-|A=0, B=1, C=0)
= P(A=0, B=1, C=0|-)*P(-) / P(A=0, B=1, C=0)
= P(A=0|-) P(B=1|-) P(C=0|-)*P(-) / K
= 0/K (Since P(C=0 | -)= 0 and this is mult. in numerator)

Based off the above caluclations, this class should be postive.

c. p=1/2 and m=4
P(A=0|+) = (2+2)/(5+4) = 4/9
P(A=0|-) = (3+2)/(5+4) = 5/9
P(B=1|+) = (1+2)/(5+4) = 3/9
P(B=1|-) = (2+2)/(5+4) = 4/9
P(C=0|+) = (1+2)/(5+4) = 3/9
P(C=0|-) = (2)/(5+4) = 2/9

d.
Positive:
Let K = P(A=0, B=1, C=0)
P(+|A=0, B=1, C=0)
= P(A=0, B=1, C=0|+)*P(+) / P(A=0, B=1, C=0)
= P(A=0|+) P(B=1|+) P(C=0|+)*P(+) / K
= (4/9)*(3/9)*(3/9)*.5/K = .0247/K

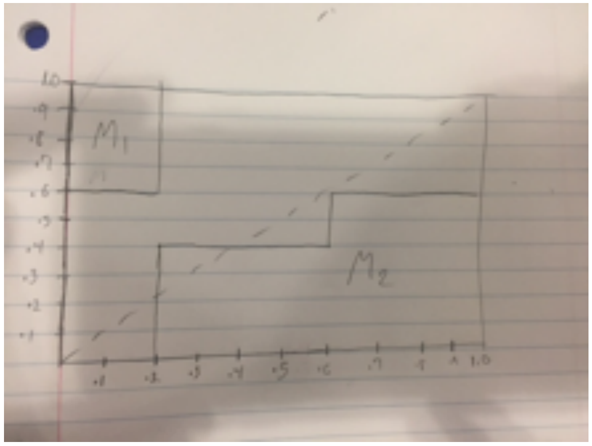
Negative:
P(-|A=0, B=1, C=0)
= P(A=0, B=1, C=0|-)*P(-) / P(A=0, B=1, C=0)
= P(A=0|-) P(B=1|-) P(C=0|-)*P(-) / K
= (5/9)*(4/9)*(2/9)*.5 / K
= .0274/K

The prediction of this class should be Negative.

e. The m-estimate was a better approach, because the conditional probability for P(C=0 | -) was equal to 0, so that completely ruled out the negative class as a possibility. With m-estimate, one conditional didn’t affect the entire calculation so it was better.

5.17

a.



The area under M1 is much greater than the area under M2 therefore M1 is a much better model.

b. Confusion matrix for M1 with cutoff t=.5

		+	-
Actual	+	3	2
	-	1	4

Precision = 3/4 = 75%
Recall = 3/5 = 60%
F-measure = (2*.75*.6)/(.75+.6) = .667

c. Confusion matrix for M2 with cutoff t=.5

		+	-
Actual	+	1	4
	-	1	4

Precision = 1/2 = 50%
Recall = 1/5 = 20%
F-measure = (2*.5*.2)/(.5+.2) = .2857

M1 is better than M2 with regards to f-measure which is consistent with the ROC curve.

d. confusion matrix for M1 with cutoff t=.1

		+	-
Actual	+	5	0
	-	4	1

Precision = 5/9 = 55.55%
Recall = 5/5 = 100%
F-measure = (2*.556*1)/(.556+.1) = .715

Based on f-measure, I prefer a threshold of t=.1 instead of t=.5.

This conclusion is inconsistent with what it should be. With t=.1 fpr=.8 and tpr=1 but with t=.5 fpr=.2 and tpr=.6. With the threshold of .5 this is closer to the ideal model so it is better.

Additional exercise: Given the below confusion matrix for classifier *C*, compute the accuracy rate, error rate, true positive, false positive, precision and F-measure.

Predicted Class				
		+	-	Total
Actual	+	350	122	472
Class	-	344	670	1014
	Total	694	792	1486

Accuracy Rate: (TP + TN)/ (P + N) = (350 + 670)/(472 + 1014) = (1020/1486) = .686

Error Rate: (FN + FP) / (TP + FN + FP + TN) = (122 + 344)/(1486) = .314

True Positive: TP/(TP + FN) = 350/(350 + 122) = .742

False Positive: FP/(FP + TN) = 344/(344 + 670) = .339

Precision: TP/(TP + FP) = 350/(350 + 344) = .504

F-Measure: 2 * ((Precision * Recall)/(Precision + Recall))

Precision = TP/(TP + FP) = .504
Recall = TP/(TP + FN) = 350/(350 + 122) = .742

= 2 * (.504 * .742)/(.504 + .742) = 2 * (.374/1.246) = 2*.3 = .6