Q1 (12 points): Chap 4 Exercise 3 (a) to (f) (2 points each)

Given a table

Table 1 Dataset for Exercise 3

Instance	a_1	a_2	a_3	Target Class
1	T	T	1.0	+
2	T	T	6.0	+
3	T	F	5.0	-
4	F	F	4.0	+
5	F	T	7.0	-
6	F	T	3.0	-
7	F	F	8.0	-
8	T	F	7.0	+
9	F	T	5.0	-

a. What is the entropy of this collection of training examples with respect to the class attribute?
We count the number of (+) and (-) then we have:

C1 (+)	4
C2 (-)	5

Recap :

$$Entropy(t) = -\sum_j p(j \mid t) \log_2 p(j \mid t)$$

P(C1) = 4/9
P(C2) = 5/9

Thus:
Entropy = - (4/9) * $\log_2$ (4/9) - (5/9) * $\log_2$ (5/9) = **0.99107606**

b. What are the information gains of a_1, a_2 relative to these training examples?
Recap:

$$GAIN_{split} = Entropy(p) - \left(\sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \right)$$

First, we calculate entropies for a_1 and a_2.

		Target Class	
		Positive (+)	Negative (-)
a_1	T	3	1
	F	1	4

Entropy (a_1, target class) = P(T) * Entropy (3,1) + P(F) * Entropy (1,4)

$$= \frac{4}{9}\left[-\frac{3}{4}*\log_2(\frac{3}{4}) - \frac{1}{4}*\log_2(\frac{1}{4})\right] + \frac{5}{9}\left[-\frac{1}{5}*\log_2(\frac{1}{5}) - \frac{4}{5}*\log_2(\frac{4}{5})\right] = 0.7616$$

GAIN a_1 = 0.9912 -0.7616 = 0.2296

		Target Class	
		Positive (+)	Negative (-)
a_2	T	2	3
	F	2	2

Entropy (a_2, target class) = P(T) * Entropy (2,3) + P(F) * Entropy (2,2) =

$$\frac{5}{9}\left[-\frac{2}{5}*\log_2(\frac{2}{5}) - \frac{3}{5}*\log_2(\frac{3}{5})\right] + \frac{4}{9}\left[-\frac{2}{4}*\log_2(\frac{2}{4}) - \frac{2}{4}*\log_2(\frac{2}{4})\right] = 0.9839$$

GAIN a_2 = 0.9912 -0.9839 = 0.0072

c. For a_3, which is a continuous attribute, compute the information gain for every possible split.

	1		**3**		**4**		**5**		**6**		**7**		**8**			
Split	**0.5**		**2**		**3.5**		**4.5**		**5.5**		**6.5**		**7.5**		**8.5**	
	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>	≤	>
P(+)	0	4	1	3	1	3	2	2	2	2	3	1	4	0	4	0
P(-)	0	5	0	5	1	4	1	4	3	2	3	2	4	1	5	0

RECAP: 0*log$_2$0 =0

Entropy (1, target class) = P(≤0.5)*Entropy (0,0) + P(>0.5)*Entropy(4,5)

$$0 + \frac{9}{9}\left[-\frac{4}{9}*\log_2(\frac{4}{9}) - \frac{5}{9}*\log_2(\frac{5}{9})\right] = 0.9912$$

GAIN (1) = 0.9912 – 0.9912 = 0

Entropy (3, target class) = P(≤2)*Entropy (1,0) + P(>2)*Entropy(3,5)

$$\frac{1}{9}\left[-\frac{1}{1}*\log_2(\frac{1}{1}) - \frac{0}{1}*\log_2(\frac{0}{1})\right] + \frac{8}{9}\left[-\frac{3}{8}*\log_2(\frac{3}{8}) - \frac{5}{8}*\log_2(\frac{5}{8})\right] = 0.8484$$

GAIN (3) = 0.9912 – 0.8484 = 0.1428

Entropy (4, target class) = P(≤3.5)*Entropy (1,1) + P(>3.5)*Entropy(3,4)

$$\frac{2}{9}\left[-\frac{1}{2}*\log_2(\frac{1}{2})-\frac{1}{2}*\log_2(\frac{1}{2})\right]+\frac{7}{9}\left[-\frac{3}{7}*\log_2(\frac{3}{7})-\frac{4}{7}*\log_2(\frac{4}{7})\right]=0.9885$$

GAIN (4) = 0.9912 – 0.9885 = 0.0027

Entropy (5, target class) = P(≤4.5)*Entropy (2,1) + P(>4.5)*Entropy(2,4)

$$\frac{3}{9}\left[-\frac{2}{3}*\log_2(\frac{2}{3})-\frac{1}{3}*\log_2(\frac{1}{3})\right]+\frac{6}{9}\left[-\frac{2}{6}*\log_2(\frac{2}{6})-\frac{4}{6}*\log_2(\frac{4}{6})\right]=0.9183$$

GAIN (5) = 0.9912 – 0.9183 = 0.0729

Entropy (6, target class) = P(≤5.5)*Entropy (2,3) + P(>5.5)*Entropy(2,2)

$$\frac{5}{9}\left[-\frac{2}{5}*\log_2(\frac{2}{5})-\frac{3}{5}*\log_2(\frac{3}{5})\right]+\frac{4}{9}\left[-\frac{2}{4}*\log_2(\frac{2}{4})-\frac{2}{4}*\log_2(\frac{2}{4})\right]=0.9839$$

GAIN (6) = 0.9912 – 0.9839 = 0.0073

Entropy (7, target class) = P(≤6.5)*Entropy (3,3) + P(>6.5)*Entropy(1,2)

$$\frac{6}{9}\left[-\frac{3}{6}*\log_2(\frac{3}{6})-\frac{3}{6}*\log_2(\frac{3}{6})\right]+\frac{3}{9}\left[-\frac{1}{3}*\log_2(\frac{1}{3})-\frac{2}{3}*\log_2(\frac{2}{3})\right]=0.9728$$

GAIN (7) = 0.9912 – 0.9728 = 0.0184

Entropy (8, target class) = P(≤7.5)*Entropy (4,4) + P(>7.5)*Entropy(0,1)

$$\frac{8}{9}\left[-\frac{4}{8}*\log_2(\frac{4}{8})-\frac{4}{8}*\log_2(\frac{4}{8})\right]+\frac{1}{9}\left[-\frac{0}{1}*\log_2(\frac{0}{1})-\frac{1}{1}*\log_2(\frac{1}{1})\right]=0.8889$$

GAIN (8) = 0.9912 – 0.8889 = 0.1023

d. What is the best split (among a_1, a_2, a_3) according to the information gain?
Best of a_3 at GAIN (3) = 0.1428,
hence

GAIN (a_3) = 0.1428,
GAIN (a_2) = 0.0072,
GAIN (a_1) = 0.2296

We can see that, MAX (GAIN (a_3), GAIN (a_2), GAIN (a_1)) = GAIN (a_1) = 0.2296. Therefore, a_1 provides the best split.

e. What is the best split (between a_1, a_2) according to the classification error rate?

Recap:

Classification Error Rate: = (FP + FN) / (TP + TN + FP + FN)

Classification error rate a_1: (1+1)/ (3+1+1+4)=2/9= 0.2222

Classification error rate a_2: (2+2)/ (2+3+2+2)=4/9= 0.4444

Therefore, **a_1 provides best split** because of lower classification error rate

f. What is the best split (between a_1, a_2) according to Gini Index?

GINI INDEX (a_1) : $\frac{4}{9}\left(1-\left(\frac{3}{4}\right)^2-\left(\frac{1}{4}\right)^2\right)+\frac{5}{9}\left(1-\left(\frac{1}{5}\right)^2-\left(\frac{4}{5}\right)^2\right)=0.3444$

GINI INDEX (a_2) : $\frac{5}{9}\left(1-\left(\frac{2}{5}\right)^2-\left(\frac{3}{5}\right)^2\right)+\frac{4}{9}\left(1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2\right)=0.4889$

Therefore, **a_1 provides best split** because of lower GINI Index

Question 2: Chapter 4, Exercise 5

Consider the following data set for a binary class problem.

A	B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

a. Calculate the information gain when splitting on A and B. Which attribute would the decision tree induction algorithm choose?

Class Label (+)	**4**
Class Label (-)	**6**

P(C1) = 4/10
P(C2) = 6/10
Thus, Entropy = $-4/10*\log_2 4/10 - 6/10\log_2 6/10 = 0.9710$

		Class Label	
		Positive (+)	Negative (-)
A	T	4	3
	F	0	3

Entropy (A, Class Label) = P(T) * Entropy (4,3) + P(F) * Entropy (0,3)
$= \frac{7}{10}\left[-\frac{4}{7}*\log_2(\frac{4}{7}) - \frac{3}{7}*\log_2(\frac{3}{7})\right] + \frac{3}{10}\left[-\frac{0}{3}*\log_2(\frac{0}{3}) - \frac{3}{3}*\log_2(\frac{3}{3})\right] = 0.6896$
GAIN (A) = 0.9710 – 0.6896 = **0.2813**

		Class Label	
		Positive (+)	Negative (-)
B	T	3	1
	F	1	5

Entropy (B, Class Label) = P(T) * Entropy (3,1) + P(F) * Entropy (1,5)
$= \frac{4}{10}\left[-\frac{3}{4}*\log_2(\frac{3}{4}) - \frac{1}{4}*\log_2(\frac{1}{4})\right] + \frac{6}{10}\left[-\frac{1}{6}*\log_2(\frac{1}{6}) - \frac{5}{6}*\log_2(\frac{5}{6})\right] = 0.7145$
GAIN (B) = 0.9710 – 0.7145 = **0.2565**

Because GAIN (A) > GAIN (B), **attribute A** will be chosen to split the node.

b. Calculate the gain in the Gini Index when splitting on A and B. Which attribute would the decision tree induction algorithm choose?
Overall GINI Index: $1 - (4/10)^2 - (6/10)^2 = 0.48$
GINI Index (A) : $\mathbf{\frac{7}{10}\left(1 - \left(\frac{4}{7}\right)^2 - \left(\frac{3}{7}\right)^2\right) + \frac{3}{10}\left(1 - \left(\frac{0}{3}\right)^2 - \left(\frac{3}{3}\right)^2\right) = 0.3428}$
GINI Index (B) : $\mathbf{\frac{4}{10}\left(1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2\right) + \frac{6}{10}\left(1 - \left(\frac{1}{6}\right)^2 - \left(\frac{5}{6}\right)^2\right) = 0.3167}$

Because GINI (B) < GINI (A), **attribute B** will be chosen to split the node

c. Figure 4.13 shows that entropy and the Gini Index are both monotonously increasing on the range [0, 0.5] and they are both monotonously decreasing on the range [0.5, 1]. Is it possible that information gain and the gain in the Gini Index favor different attribute? Explain
Yes, as indicated in previous question a) and b) due to the use of different functions.

Question 3. Chapter 4 Exercise 7

The following table summarizes a data set with three attributes A, B, C and two class labels +, -. Build a two-level decision tree

A	B	C	Number of instances	
			+	-
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

a. According to the classification error rate, which attribute would be chosen as the first splitting attribute? For each attribute, show the contingency table and the gains in classification error rate.

	Number of instances
+	50
-	50

Original classification error rate: 1 – MAX (50/100,50/100) = 0.5
Entropy = -50/100*$\log_2$50/100 -50/100*$\log_2$50/100 =1

Entropy (A, Class Label) = P(T) * Entropy (25,0) + P(F) * Entropy (25,50)
$= \frac{25}{100}\left[-\frac{25}{25}*\log_2(\frac{25}{25})-\frac{0}{25}*\log_2(\frac{0}{25})\right]+\frac{75}{100}\left[-\frac{25}{75}*\log_2(\frac{25}{75})-\frac{50}{75}*\log_2(\frac{50}{75})\right]$
= 0.6887
GAIN (A) = 1 – 0.6887 = ***0.3113***

Entropy (B, Class Label) = P(T) * Entropy (30,20) + P(F) * Entropy (20,30)
$= \frac{50}{100}\left[-\frac{30}{50}*\log_2(\frac{30}{50})-\frac{20}{50}*\log_2(\frac{20}{50})\right]+\frac{50}{100}\left[-\frac{20}{50}*\log_2(\frac{20}{50})-\frac{30}{50}*\log_2(\frac{30}{50})\right]$

= 0.9710
GAIN (B) = 1 – 0.9710 = ***0.029***

Entropy (C, Class Label) = P(T) * Entropy (25,25) + P(F) * Entropy (25,25)
$= \frac{50}{100}\left[-\frac{25}{50} * \log_2(\frac{25}{50}) - \frac{25}{50} * \log_2(\frac{25}{50})\right] + \frac{50}{100}\left[-\frac{25}{50} * \log_2(\frac{25}{50}) - \frac{25}{50} * \log_2(\frac{25}{50})\right]$
= 1
GAIN (C) = 1 – 1 = ***0***

Classification Error Rate: = (FP + FN) / (TP + TN + FP + FN)

		+	-
A	**T**	25	0
	F	25	50

		+	-
B	**T**	30	20
	F	20	30

		+	-
C	**T**	25	25
	F	25	25

Classification Error Rate A: (25+0)/(25+0+25+50) = 25/100
Classification Error Rate B: (20+20)/(30+20+20+30) = 40/100
Classification Error Rate C: (25+25)/(25+25+25+25) = 50/100

Attribute A provides the lowest classification error rate (or highest information gain); therefore, it will be chosen to split the node.

b. Repeat for the two children of the root node
Because A_T = 25 (+) and 0 (-). It is pure for T's values; no further action is required!
For $A_{T=F}$ we construct new table as follow by removing the shaded rows

A	B	C	Number of instances	
			+	-
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

We have:

A	B	C	Number of instances	
			+	-
F	T	T	0	20
F	F	T	0	5
F	T	F	25	0
F	F	F	0	25

	Number of instances
+	25
-	50

		+	-
B	**T**	25	20
	F	0	30

		+	-
C	**T**	0	25
	F	25	25

Classification Error Rate B: (20+0)/(25+20+0+30) = 20/75
Classification Error Rate C: (25+25)/(0+25+25+25) = 50/75

Entropy = $-25/75*\log_2 25/75 \ -50/75*\log_2 50/75 = 0.9183$

Entropy (B, Class Label) = P(T) * Entropy (25, 20) + P(F) * Entropy (0,30)
$= \frac{45}{75}\left[-\frac{25}{45}*\log_2(\frac{25}{45}) - \frac{20}{45}*\log_2(\frac{20}{45})\right] + \frac{30}{75}\left[-\frac{0}{30}*\log_2(\frac{0}{30}) - \frac{30}{30}*\log_2(\frac{30}{30})\right]$
= 0.5946
GAIN (B) = 0.9183 – 0.5946 = 0.3237

Entropy (C, Class Label) = P(T) * Entropy (0, 25) + P(F) * Entropy (25,25)
$= \frac{25}{75}\left[-\frac{0}{25} * \log_2(\frac{0}{25}) - \frac{25}{25} * \log_2(\frac{25}{45})\right] + \frac{50}{75}\left[-\frac{25}{50} * \log_2(\frac{25}{50}) - \frac{25}{50} * \log_2(\frac{25}{50})\right]$
= 0.6667
GAIN (B) = 0.9183 – 0.6667 = 0.2156

From both classification error rate and information gain, we can see that, **attribute B** will be chosen to split node.

c. How many instances are misclassified by the resulting decision tree?

		+	-
B	**T**	25	20
	F	0	30

Because B_F = 0 (+) and 30 (-). It is pure for F's values; Thus, the resulting decision tree has 20 misclassified instances.

d. Repeat parts a), b), c) using C as the first root/splitting attribute.
Recap:

		+	-
C	**T**	25	25
	F	25	25

Entropy (C, Class Label) = P(T) * Entropy (25,25) + P(F) * Entropy (25,25)
$= \frac{50}{100}\left[-\frac{25}{50} * \log_2(\frac{25}{50}) - \frac{25}{50} * \log_2(\frac{25}{50})\right] + \frac{50}{100}\left[-\frac{25}{50} * \log_2(\frac{25}{50}) - \frac{25}{50} * \log_2(\frac{25}{50})\right]$
= 1
GAIN (C) = 1 – 1 = ***0***
Classification Error Rate C: (25+25)/(25+25+25+25) = 50/100

After splitting node C, we have two sets (distinguished by shaded region)

A	**B**	**C**	**Number of instances**	
			+	-
T	T	T	5	0
F	T	T	0	20
T	F	T	20	0
F	F	T	0	5
T	T	F	0	0
F	T	F	25	0
T	F	F	0	0
F	F	F	0	25

For the 1st set

		+	-
A	**T**	25	0
	F	0	25

		+	-
B	**T**	5	20
	F	20	5

Classification error rate A: 0/50 = 0
Classification error rate B: 40/50

Entropy (A, Class Label) = P(T) * Entropy (25, 0) + P(F) * Entropy (0,25)
$= \frac{25}{50}\left[-\frac{25}{25} * \log_2(\frac{25}{25}) - \frac{0}{25} * \log_2(\frac{0}{25})\right] + \frac{25}{50}\left[-\frac{0}{25} * \log_2(\frac{0}{25}) - \frac{25}{25} * \log_2(\frac{25}{25})\right]$
= 0
GAIN (B) = 1 – 0 = 1.

Entropy (B, Class Label) = P(T) * Entropy (5, 20) + P(F) * Entropy (20,5)
$= \frac{25}{50}\left[-\frac{5}{25} * \log_2(\frac{5}{25}) - \frac{20}{25} * \log_2(\frac{20}{25})\right] + \frac{25}{50}\left[-\frac{20}{25} * \log_2(\frac{20}{25}) - \frac{5}{25} * \log_2(\frac{5}{25})\right]$
= 0.7220
GAIN (B) = 1 – 0.7220 = 0.278

Attribute A will be chosen to split the node with 0 misclassified instances

For the 2nd set

		+	-
A	**T**	0	0
	F	25	25

		+	-
B	**T**	25	0
	F	0	25

Classification error rate A: 25/50 = 0.5
Classification error rate B: 0/50 = 0

Entropy (A, Class Label) = P(T) * Entropy (0, 0) + P(F) * Entropy (25,25)
$= \left[-\frac{25}{50} * \log_2(\frac{25}{50}) - \frac{25}{50} * \log_2(\frac{25}{50})\right]$
= 1

GAIN (A) = 1 – 1 = 0.

Entropy (B, Class Label) = P(T) * Entropy (25, 0) + P(F) * Entropy (0,25)
$= \frac{25}{50}\left[-\frac{25}{25} * \log_2(\frac{25}{25}) - \frac{0}{25} * \log_2(\frac{0}{25})\right] + \frac{25}{50}\left[-\frac{0}{25} * \log_2(\frac{0}{25}) - \frac{25}{25} * \log_2(\frac{25}{25})\right]$
$= 0$
GAIN (B) = 1 – 0 = 1

Attribute B will be chosen to split the node with 0 misclassified instance

e. Use the results in parts (c) and (d) to conclude about the greedy nature of decision tree induction algorithm.
It can be seen that, starting from attribute A can lead to 20 misclassified instances but starting from attribute C, overall classification error rate is 0. This concludes that greedy heuristic does not always lead to the best decision tree.

Question 4: Chapter 5 Exercise 7

Consider the data set shown in Table 5.10

Record	A	B	C	Class
1	0	0	0	+
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-
5	0	0	1	+
6	1	0	1	+
7	1	0	1	-
8	1	0	1	-
9	1	1	1	+
10	1	0	1	+

a. Estimate the conditional probabilities for P (A| +), P(B|+), P(C|+), P (A| -), P(B|-), P(C|-)?
P (A_0| +) = 2/5 = 0.4 ; P (A_0| -) = 3/5 = 0.6
P (A_1| +) = 3/5 = 0.6 ; P (A_1| -) = 2/5 = 0.4

P (B_0| +) = 4/5 ; P (B_0| -) = 3/5
P (B_1| +) = 1/5 ; P (B_1| -) = 2/5

P (C_0| +) = 1/5 ; P (C_0| -) = 0/5
P (C_1| +) = 4/5 ; P (C_1| -) = 5/5

b. Use the estimate of conditional probabilities given in the previous question to predict the class label for a test sample (A=0, B=1, C=0) using the naïve Bayes approach.

To predict the class label of a test record X = (A=0, B=1, C=0) we need to compute the posterior probabilities P(+|X) and P(-|X)
P(+) = P (-) = 0.5
P (X|+)= P(A=0|+) * P(B=1|+)*P(C=0|+) = 2/5 * 1/5 * 1/5 = 2/75
P (X|-) = P(A=0|-) * P(B=1|-)*P(C=0|-) = 3/5 * 2/5 * 0/5 = 0

P(+|X) = P(X|+) * P(+) / P(X) = *a**0.5*2/75 where *a* =1/P(X) is a constant term
P(-|X) = P(X|-) * P(-) / P(X) = 0
Since P(+|X) > P(-|X), the record is classified as **'+'**

c. Estimate the conditional probabilities using the m-estimate approach, with p=1/2, and m=4
Recap:
$P(x_i|y_i) = (n_c + mp)/(n+m)$

$P(A_0\| +) = (2 + 4*1/2)/5+4 = 4/9$	$P(A_0\| -) = (3 + 4*1/2)/5+4 = 5/9$
$P(A_1\| +) = (3 + 4*1/2)/5+4 = 5/9$	$P(A_1\| -) = (2 + 4*1/2)/5+4 = 4/9$
$P(B_0\| +) = (4 + 4*1/2)/5+4 = 6/9$	$P(B_0\| -) = (3 + 4*1/2)/5+4 = 5/9$
$P(B_1\| +) = (1 + 4*1/2)/5+4 = 3/9$	$P(B_1\| -) = (2 + 4*1/2)/5+4 = 4/9$
$P(C_0\| +) = (1 + 4*1/2)/5+4 = 3/9$	$P(C_0\| -) = (0 + 4*1/2)/5+4 = 2/9$
$P(C_1\| +) = (4 + 4*1/2)/5+4 = 6/9$	$P(C_1\| -) = (5 + 4*1/2)/5+4 = 7/9$

d. Repeat part (b) using the conditional probabilities given in part (c)

P (X|+)= P(A=0|+) * P(B=1|+)*P(C=0|+) = 4/9 * 3/9 * 3/9 = 4/81 = 0.04938

P (X|-) = P(A=0|-) * P(B=1|-)*P(C=0|-) = 5/9 * 4/9 * 2/9 = 40/729 = 0.05487

P(+|X) = P(X|+) * P(+) / P(X) = *a**0.5*0.04938

where *a* =1/P(X) is a constant term

P(-|X) = P(X|-) * P(-) / P(X) = *a**0.5*0.05487

Since P(+|X) < P(-|X), the record is classified as **'-'**

e. Compare the two methods for estimating probabilities. Which method is better and why?

The reason to have m-estimate of conditional probability is to overcome the vanish of the overall posterior probability if the class-conditional probability for one of the attribute is zero. This problem is shown in part (b). So **it is obvious that the m-estimate approach is better.**

Question 5: Chapter 5 Exercise 12

Given the Bayesian network shown in Figure 5.48, compute the following probabilities.

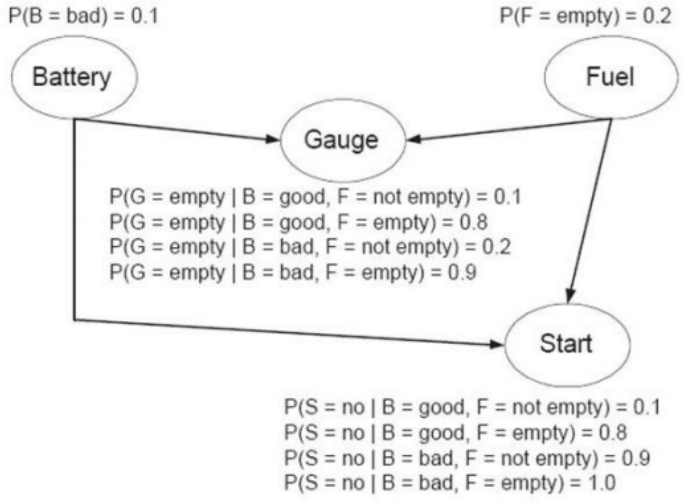


Figure 5.48. Bayesian belief network for Exercise 12.

a. P(B=good, F=empty, G=empty, S=yes)
   = P(G,S|B,F)*P(B,F) = P(G|BF)*P(S|BF)*P(B)*P(F)
   = 0.8 * (1-0.8)* (1-0.1)*0.2 = 0.8*0.2*0.9*0.2 = **0.0288**
b. P (B=bad, F=empty, G=not empty, S = no)
   = P(G,S|B,F)*P(B,F) = P(G|BF)*P(S|BF)*P(B)*P(F)
   = (1-0.9) * 1.0 *0.1 *0.2 =**0.002**
c. Given that the battery is bad, compute the probability that the cart will start.
   Let $\alpha \in \{\text{empty, not empty}\}$. Given the conditions above we need to calculate P(S=start|B=bad, F= α) .
   First we have
   P(S=start|B=bad, F=empty) = 1- P(S=no|B=bad, F=empty) =1-1=0

   Then
   P(S=start|B=bad, F= α)
   $= \sum_{\alpha} P(S = \text{start}|B = \text{bad}, F = \alpha)P(B = \text{bad})P(F = \alpha)$
   $= P(S = \text{start}|B = \text{bad}, F = \text{empty})P(B = \text{bad})P(F = \text{empty}) +$
   $P(S = \text{start}|B = \text{bad}, F = \text{not empty})P(B = \text{bad})P(F = \text{not empty})$

$= 0 + (1- P(S = no|B = bad, F = not\ empty))*P(B=bad)*P(F=not\ empty)$

$= (1-0.9)*0.1*(1-0.2) = 0.1*0.1*0.8 =$ **0.008**

Question 6: Chapter 5 Exercise 17

You are asked to evaluate the performance of two classification models, M1 and M2. The test set you have chosen contains 26 binary attributes, labeled as A through Z.

Table 5.14 shows the posterior probabilities obtained by applying the models to the test set. (Only the posterior probabilities for the positive class are shown). As this is a two-class problem, P(-) = 1- P(+) and P(-|A,…,Z) = 1-P(+|A,…,Z). Assume that we are mostly interested in detecting instances from the positive class.

a. Plotting the ROC curve for both M1 and M2 (You should plot them on the same graph). Which model do you think is better? Explain your reasons?

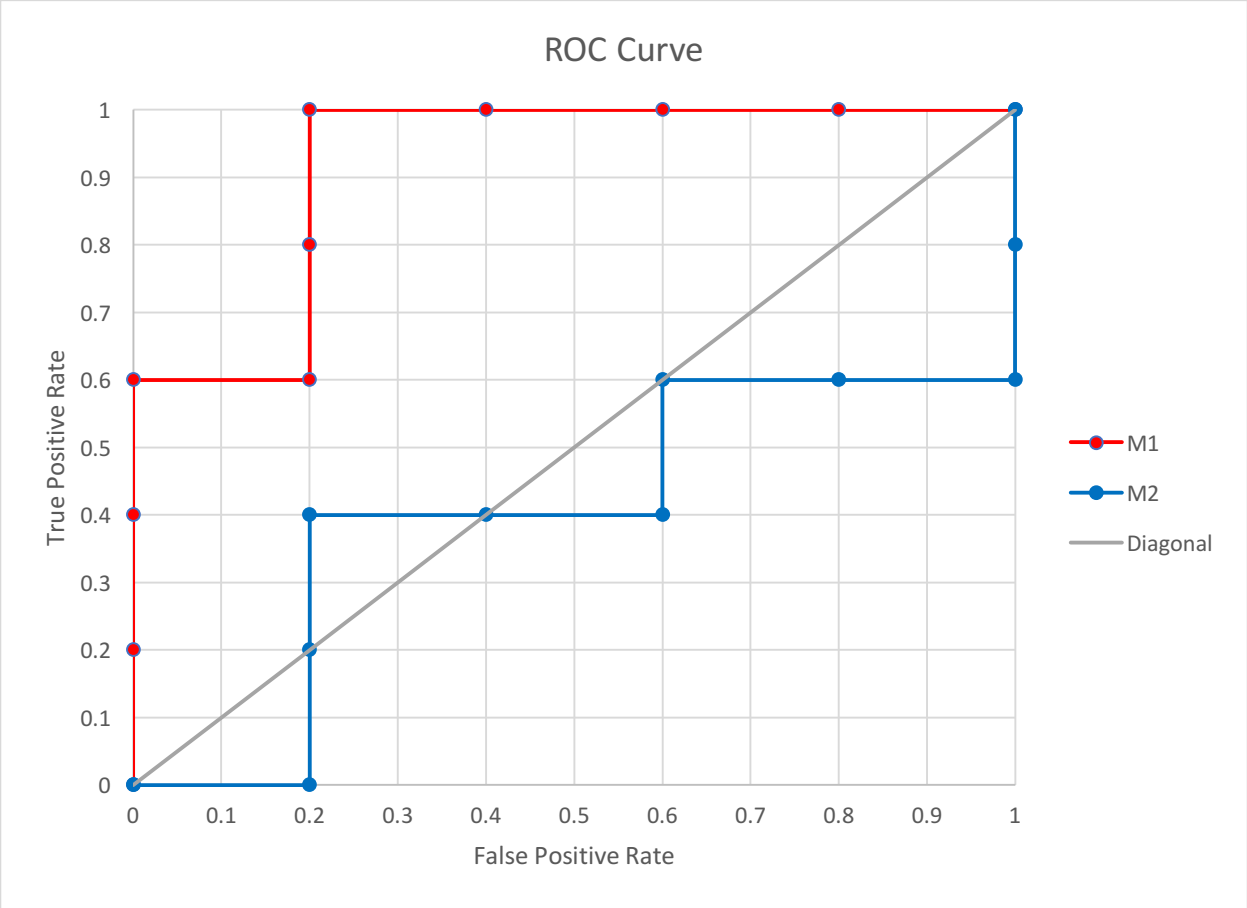
Table 5.14. Posterior probabilities for Exercise 17.

Instance	True Class	$P(+\|A,\ldots,Z,M_1)$	$P(+\|A,\ldots,Z,M_2)$
1	+	0.73	0.61
2	+	0.69	0.03
3	−	0.44	0.68
4	−	0.55	0.31
5	+	0.67	0.45
6	+	0.47	0.09
7	−	0.08	0.38
8	−	0.15	0.05
9	+	0.45	0.01
10	−	0.35	0.04

We construct two new tables as follow

Class	-	-	-	-	+	+	-	+	+	+	
M1	0.08	0.15	0.35	0.44	0.45	0.47	0.55	0.6	0.69	0.73	1
TP	5	5	5	5	5	4	3	3	2	1	0
FP	5	4	3	2	1	1	1	0	0	0	0
TN	0	1	2	3	4	4	4	5	5	5	5
FN	0	0	0	0	0	1	2	2	3	4	5
TPR	1	1	1	1	1	0.8	0.6	0.6	0.4	0.2	0
FPR	1	0.8	0.6	0.4	0.2	0.2	0.2	0	0	0	0

Class	+	+	-	-	+	-	-	+	+	-	
M2	0.01	0.03	0.04	0.05	0.09	0.31	0.38	0.45	0.61	0.68	1
TP	5	4	3	3	3	2	2	2	1	0	0
FP	5	5	5	4	3	3	2	1	1	1	0
TN	0	0	0	1	2	2	3	4	4	4	5
FN	0	1	2	2	2	3	3	3	4	5	5
TPR	1	0.8	0.6	0.6	0.6	0.4	0.4	0.4	0.2	0	0
FPR	1	1	1	0.8	0.6	0.6	0.4	0.2	0.2	0.2	0



From the graph above, it is clearly **that M1 performs better than M2** because its area under ROC curve is larger than that of M2's area.

b. For model M1, suppose you choose the cutoff threshold to be t=0.5. In other words, any test instances whose posterior probability is greater than t will be classified as a positive example. Compute the precision, recall, and F-measure for the model at this threshold value

At threshold t = 0.5 we have the following table.

M1	Actual Class	Predicted Class
0.08	-	-
0.15	-	-
0.35	-	-
0.44	-	-
0.45	+	-
0.47	+	-
0.55	-	+
0.67	+	+
0.69	+	+
0.73	+	+

We have confusion table as follow

	Predicted Class		
Actual Class (M1)		Class = +	Class = -
	Class =+	a = 3	b = 2
	Class =-	c = 1	d = 4

Precision (p) = a/(a+c) = ¾ =0.75
Recall (r) = a/(a+b) = 3/(3+2) = 3/5 = 0.6
F-measure (F) = 2rp/(r+p) = 2a/(2a+b+c) = 2*3/(2*3+ 2+1) = 6/9 = 0.667

c. Repeat the analysis for part (c) using the same cutoff threshold on model M2. Compare the F-measure results for both models. Which model is better? Are the results consistent with what you expect from the ROC curve?

M2	Actual Class	Predicted Class
0.01	+	-
0.03	+	-
0.04	-	-
0.05	-	-
0.09	+	-
0.31	-	-
0.38	-	-
0.45	+	-
0.61	+	+
0.68	-	+

	Predicted Class		
Actual Class (M2		Class = +	Class = -
	Class =+	a = 1	b = 4
	Class =-	c = 1	d = 4

Precision (p) = a/(a+c) = 1/2 =0.5
Recall (r) = a/(a+b) = 1/(1+4) =1/5 = 0.2
F-measure (F) = 2rp/(r+p) = 2a/(2a+b+c) = 2*1/(2*1+ 4+1) = 2/7 = 0.2857

Because F-measure (M1=0.667) > F-measure (M2=0.2857) so M1 is better than M2. Without looking at the ROC curve, we expect that M1 is closer to (0,1) than M2 at t=0.5.
We have M1 = (0.2, 0.6), M2 = (0.2, 0.2).
From the ROC curve, this result is consistent. Indeed
$d(M1) = \text{sqrt}((1-0.6)^2 + 0.2^2) = \text{sqrt}(0.2) = 0.447$
$d(M2) = \text{sqrt}((1-0.2)^2 + 0.2^2) = \text{sqrt}(0.68) = 0.4624$
Therefore d(M1) < d(M2) meaning that M1 is closer

d. Repeat part (c) for model M1 using the threshold t=0.1. Which threshold do you prefer, t=0.5 or t=0.1? Are the results consistent with what you expect from the ROC curve?

M1	**Actual Class**	**Predicted Class**
0.08	-	-
0.15	-	+
0.35	-	+
0.44	-	+
0.45	+	+
0.47	+	+
0.55	-	+
0.67	+	+
0.69	+	+
0.73	+	+

	Predicted Class		
Actual Class (M1)		Class = +	Class = -
	Class =+	a = 5	b = 0
	Class =-	c = 4	d = 1

Precision (p) = a/(a+c) = 5/9 =0.556
Recall (r) = a/(a+b) = 5/(5+0) =5/5 = 1
F-measure (F) = 2rp/(r+p) = 2a/(2a+b+c) = 2*5/(2*5+ 0+4) = 10/14 = 0.7143

In terms of F measure, t=0.1 gives the highest F value so it is preferable. We would expect that M1 at t=0.1 is closer to the point (0,1) than M1 at t=0.5 or d(M1=0.1) < d(M1=0.5)

We have
M1(0.1) = (0.8,1), M1(0.5) = (0.2,0.6)

d(M1=0.1) = 0.8 and d(M1=0.5) = 0.447.

Here d(M1) > d(M2) that is contradict to what we expect. Therefore, this result is NOT consistent with the ROC curve.

Since F-measure and ROC curve are NOT consistent. Area under the curve is a good indicator to select threshold t

Area M1 (t=0.1) = 1*(1-0.8) = 0.2

Area M1 (t=0.5) = 0.6*(1-0.2) = 0.48

Since Area M1 (t=0.5) > M1 (t=0.1) **so t=0.5 is preferable.**

Question 7. Given a Twitter dataset, how do you builder a classifier which can identify Harvey rescue tweets automatically (you can define your target if your data does not contain Harvey rescue data)

To classify Harvey rescue tweets automatically, we use "bag-of-words" approach. That is, we define some words that have similar meaning to rescue such as "save, recover, release, relief" and compute the distance between each document with "bag-of-words" vector. If the distance is non-zero value, we classify this document as "Rescue", otherwise "None-Rescue".

a. What is your distance function? (3 points)
   Since we are working with document, each document will be represented as a vector where each attribute denotes the frequency of a term that occurs in the document. The distance function we are going to use is **"Euclidean distance"**
b. Can you run KNN on your dataset? How much training set do you have? (7 points)
   Yes, since our documents are labelled with "Rescue" and "None-Rescue" we can run KNN algorithm.
   Our original data contains approximately 5,300 records, after preprocessing step (remove duplicate tweet or retweets), our final data contains 1700 records. As a rule of thumb, 70% of data will be used as training data 1190 documents.

c. Can you evaluate your KNN classifier using precision, recall and F-measure? (4 points).
   Yes, we can evaluate KNN classifier based on the prediction output vs testing data

d. How to determine the best K? Using different K, compare their recall, precision and F-measure? (4 points)
   As a rule of thumb, the best K is the **square root** of the total records in the dataset. Note that we will select K as an **odd number**.
e. Visualize classification results? (2 points)