DECISION TREE

Exercises

1. Consider the training examples shown in **Table 3.5** for a binary classification problem.

Table 3.5

Customer ID	Gender	Car Type	Shirt Size	Class
1	M	Family	Small	C0
2	M	Sports	Medium	C0
3	M	Sports	Medium	C0
4	M	Sports	Large	C0
5	M	Sports	Extra Large	C0
6	M	Sports	Extra Large	C0
7	F	Sports	Small	C0
8	F	Sports	Small	C0
9	F	Sports	Medium	C0
10	F	Luxury	Large	C0
11	M	Family	Large	C1
12	M	Family	Extra Large	C1
13	M	Family	Medium	C1
14	M	Luxury	Extra Large	C1
15	F	Luxury	Small	C1
16	F	Luxury	Small	C1
17	F	Luxury	Medium	C1
18	F	Luxury	Medium	C1
19	F	Luxury	Medium	C1
20	F	Luxury	Large	C1

a. Compute the Gini index for the overall collection of training examples.

Answer: In this case, C0 and C1 have the same relative frequencies (p = 1 − p =0.5)

Gini = $1 - p^2 - (1-p)^2$= 2*p*(1 − p)

Gini = 2*0.5*0.5= **0.5**

b. Compute the Gini index for the `Customer ID` attribute.

	1	2	……	19	20
C0	1	1		0	0
C1	0	0		1	1

The Gini index for the Customer ID attribute is **0**.

c. Compute the Gini index for the `Gender` attribute.

	Male	Female
C0	6	4
C1	4	6
	10	10

Gini(M)= 1 − 0.6^2 -0.4^2 = **0.48**
Gini(F)= 1 - 0.4^2- 0.6^2 = **0.48**
Gini of Gender attribute= 0.5 × 0.48 + 0.5 × 0.48 = **0.48**

d. Compute the Gini index for the `Car Type` attribute using multiway split.

	Family	Sports	Luxury
C0	1	8	1
C1	3	0	7
	4	8	8

Gini(Family)= 1 − $1/4^2$ -$3/4^2$ = **0.375**
Gini(Sports)= 1- 1^2 - 0^2= **0**
Gini(Luxury)= 1- $1/8^2$ - $7/8^2$=**0.2188**
Gini index of Car Type attribute= 4/20*0.375+8/20*0.2188= **0.1625**

e. Compute the Gini index for the `Shirt Size` attribute using multiway split.

	Small	Medium	Large	Extra Large
C0	3	3	2	2
C1	2	4	2	2
	5	7	4	4

Gini(Small)= 2*3/5*2/5=**0.48**

Gini(Medium)= 2* 3/7*4/7= **0.4898**

Gini(Large)= 2* 0.5*0.5= **0.5**

Gini(Extra Large)= 2* 0.5*0.5= **0.5**

Gini index of Shirt Size attribute= ¼*0.48+(7/20)*0.4898+2*((1/5)*0.5)= **0.4919**

f. Which attribute is better, `Gender`, `Car Type` or `Shirt Size` ?

Gini (Car Type) = 0.1625 **<** Gini (Gender)=0.48 < Gini (Shirt Size) = 0.49

⇨ **Car Type because it has the lowest Gini index.**

g. Explain why `Customer ID` should not be used as the attribute test condition even though it has the lowest Gini.

.

The gini index of Customer Id is 0 and if you add more IDs to the table will only increase the number of partitions, resulting in no further purity gain.

2 . Consider the training examples shown in **Table 3.6** for a binary classification problem.

Table 3.6. Data set

Instance	a1	a2	a3	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?

Entropy (Class) = p*$\log_2$ (p) − (1 − p) *$\log_2$ (1 − p)

$$= -\frac{4}{9} * \log_2(\frac{4}{9}) - \frac{5}{9} *\log_2(\frac{5}{9}) = 0.99$$

b. What are the information gains of a1 and a2 relative to these training examples?

a1	T	F
+	3	1
-	1	4

Entropy(a1) = $\frac{4}{9}$[- $\frac{3}{4}$ * $\log_2(\frac{3}{4})$ - $\frac{1}{4}$ *$\log_2(\frac{1}{4})$]+ $\frac{5}{9}$[- $\frac{1}{5}$ * $\log_2(\frac{1}{5})$ - $\frac{4}{5}$ *$\log_2(\frac{4}{5})$]

= 0.7616

Gain(a1) = 0.99-0.76= 0.2294

a2	T	F
+	2	2
-	3	2

Entropy(a2) = $\frac{5}{9}$[- $\frac{2}{5}$ * $\log_2(\frac{2}{5})$ - $\frac{3}{5}$ *$\log_2(\frac{3}{5})$]+ $\frac{4}{9}$[- $\frac{1}{2}$ * $\log_2(\frac{1}{2})$ - $\frac{1}{2}$ *$\log_2(\frac{1}{2})$]

= 0.9838

Gain(a2)= 0.99-0.9838 = 0.0072

c. For a3, which is a continuous attribute, compute the information gain for every possible split.

a_3	Class label	Split point	Entropy	Info Gain
1.0	+	2.0	0.8484	0.1427
3.0	-	3.5	0.9885	0.0026
4.0	+	4.5	0.9183	0.0728
5.0 5.0	- -	5.5	0.9839	0.0072
6.0	+	6.5	0.9728	0.0183
7.0 7.0	+ -	7.5	0.8889	0.1022

d. What is the best split (among a1, a2, and a3) according to the information gain?

According to information gain, a1 produces the best split.

e. What is the best split (between a1 and a2) according to the misclassification error rate?

For attribute a1: error rate = 2/9. For attribute a2: error rate = 4/9.
Therefore, according to error rate, a1 produces the best split

f. What is the best split (between a1 and a2) according to the Gini index?

For attribute a_1, the gini index is

$$\frac{4}{9}\left[1-(3/4)^2-(1/4)^2\right]+\frac{5}{9}\left[1-(1/5)^2-(4/5)^2\right]=0.3444.$$

For attribute a_2, the gini index is

$$\frac{5}{9}\left[1-(2/5)^2-(3/5)^2\right]+\frac{4}{9}\left[1-(2/4)^2-(2/4)^2\right]=0.4889.$$

Since the gini index for a_1 is smaller, it produces the better split.