DECISION TREE

Compte Rendu

Réalisé par :

Intissar SIDAOUI

3 DNI / G2

Exercice °1

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

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

⇨ J’ai résumée tout le travail dans ce tableau

⇨ Les démonstrations sont au-dessous.

	Gini index	Gini Totale
Overall collection	**0.5**	----------------------
Customer ID	**0**	**0**
Gender	**Gini(M)=0.48** **Gini(F)=0.48**	**0.48**
Car Type	**Gini(Family)= 0.375** **Gini(Sport) = 0** **Gini(Luxury) = 0.218**	**0.1622**
Shirt Size	**Gini(Small) = 0.48** **Gini(Medium) = 0.489** **Gini(Large) = =0.5** **Gini(Extra Large) =0.5**	**0.4911**

	Gender	**Car Type**	**Shirt Size**
Gain	**0.02**	**0.3378**	**0.0089**

On a :

Les Class totale = 20

10 class ''C0'' C0=$\frac{10}{20} = \frac{1}{2}$

Et ➔

10 class ''C1'' C1=$\frac{10}{20} = \frac{1}{2}$

Alors,

On a :

Gini= 1 – P(c0) ^2 – (Pc1) ^2

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

Gini= 1 – P(c0) ^2 – (Pc1) ^2

$= 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2$

$= 1 - \frac{1}{2}$

$=\frac{1}{2}$

b. *Compute the Gini index for the 'Customer ID' attribute.*

Dans la class C0

Les Customer ID {1,2,3,4,5,6,7,8,9,10}, chacun entre eux prend comme probabilité 1 en C0 et 0 en C1

Donc,

Gini= 1 – P(c0) ^2 – (Pc1) ^2

=1 – 1 – 0

=0

Dans la class C1

Les Customer ID {11,12,13,14,15,16,17,18,19,20}, chacun entre eux prend comme probabilité 1 en C1 et 0 en C0

Donc,

Gini= 1 – P(c0) ^2 – (Pc1) ^2

=1 – 0 – 1

=0

- ⇨ La somme des probabilités est égale à 0

Gini=0

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

Pour les 10 class de C0, on a :

6 Masculins

4 féminins

D'où

M = $\frac{6}{10}$

F = $\frac{4}{10}$

Donc,

Gini (M)= 1 – $\left(\frac{4}{10}\right)^2 - \left(\frac{6}{10}\right)^2$

$=1 - \frac{16}{100} - \frac{36}{100}$

$=1 - \frac{52}{100}$

$=\frac{48}{100} = 0.48$

Gini (F) = $1-\left(\frac{6}{10}\right)^2-\left(\frac{4}{10}\right)^2$

$=1-\frac{52}{100}$

$=\frac{48}{100}$ = 0.48

⇨ Gini total = $\frac{10}{20}*Gini(M)+\frac{10}{20}*Gini(F)=\frac{10}{20}*0.48+\frac{10}{20}*0.48$ = 0.48

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

Pour tous les 20 classes, on a :

4 Family

8 Sport

8 Luxury

Pour le class C0, on a :

1 family ➔ P(Family) = $\frac{1}{4}$

8 Sport ➔ P(Sport) = 1

1 Luxury ➔ P(Luxury) = $\frac{1}{8}$

Pour le class C1, on a :

3 Family ➔ P(Family) = $\frac{3}{4}$

0 Sport ➔ P(Sport) = 0

7 Luxury ➔ P(Luxury) = $\frac{7}{8}$

Gini(Family)= 1- $\left(\frac{1}{4}\right)^2 - \left(\frac{3}{4}\right)^2 = 0.375$

Gini(Sport) = 1 – $1 - 0$=0

Gini(Luxury) = 1 – $\left(\frac{1}{8}\right)^2 - \left(\frac{7}{8}\right)^2 = 0.218$

⇨ Gini Total = $\frac{4}{20} * Gini(Family) + \frac{8}{20} * Gini\ (Sport) + \frac{8}{20} * Gini\ (Luxury) = \frac{4}{20} * 0.375 + \frac{8}{20} * 0 + \frac{7}{8} * 0.218 = 0.1622$

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

5 Small

7 Medium

4 Large

4 Extra Large

Pour C0 :

P(Small) = $\frac{3}{5}$

P(Medium) = $\frac{3}{7}$

P(Large) = $\frac{2}{4}$

P(Extra Large) = $\frac{2}{4}$

Pour C1 :

P(Small) = $\frac{2}{5}$

P(Medium) = $\frac{4}{7}$

P(Large) = $\frac{2}{4}$

P(Extra Large) = $\frac{2}{4}$

⇨

Gini(Small) = $1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2 = 0.48$

Gini(Medium) = $1-\left(\frac{3}{7}\right)^2-\left(\frac{4}{7}\right)^2 = 0.489$

Gini(Large) = $1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2 = 0.5$

Gini(Extra Large) = $1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2 = 0.5$

⇨ Gini Totale = $(\left(\frac{5}{20}*0.48\right)+\left(\frac{7}{20}*0.489\right)+\left(\frac{4}{20}*0.5\right)+(\frac{4}{20}*0.5)) = 0.4911$

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

The Gini index for the overall collection = 0.5

The Gini index for the Gender = 0.48

The Gini index for the Car Type =0.1622

The Gini index for the Shirt Size =0.491

➔

Gain(Gender) = 0.5 − 0.48 = 0.02

Gain (Car Type) = 0.5 − 0.1622 = 0.3378

Gain (Shirt Size) = 0.5 − 0.4911 = 0.0089

D’après les résultats on constate que, le gain de Car Type est le plus élevé

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

Car chaque Customer Id est se répète une seule fois et dans une seule classe.

Alors chacun, prend comme probabilité l'une = 1 et l'autre = 0

Exercice ° 2

---

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

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	−

Dans cette Data-set, on a :

9 target class

4 Positives

5 Négatives

Avec :

(Pc+) = $\frac{4}{9}$

(Pc-) = $\frac{5}{9}$

On a :

Entropy = $-\frac{p}{p+n}\ log2\ \left(\frac{p}{p+n}\right) - \frac{n}{p+n}\ log2(\frac{n}{p+n})$

a. What is the entropy of this collection of training examples with respect to the class attribute ?

Entropy (Class) = (Pc+) log2 (Pc+) − (Pc-) log2 (Pc-)

$$=-\frac{4}{9} log2\ \left(\frac{4}{9}\right) - \frac{5}{9} log2\ \left(\frac{5}{9}\right) = 0.9911$$

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

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

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

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

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