



DECISION TREE

Compte Rendu

Réalisé par :

Intissar SIDAOU

3 DNI / G2

Exercise °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 :

$$\text{Gini} = 1 - P(c0)^2 - (Pc1)^2$$

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

$$\begin{aligned}\text{Gini} &= 1 - P(c0)^2 - (Pc1)^2 \\ &= 1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2 \\ &= 1 - \frac{1}{2} \\ &= \frac{1}{2}\end{aligned}$$

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,

$$\text{Gini} = 1 - P(c0)^2 - (Pc1)^2$$

$$= 1 - 1 - 0$$

$$= 0$$

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,

$$\text{Gini} = 1 - P(c_0)^2 - (P(c_1))^2$$

$$= 1 - 0 - 1$$

$$= 0$$

⇒ La somme des probabilités est égale à 1

$$\text{Gini} = 0$$

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,

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

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

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

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

$$\begin{aligned}
 \text{Gini (F)} &= 1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2 \\
 &= 1 - \frac{52}{100} \\
 &= \frac{48}{100} = 0.48
 \end{aligned}$$

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

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 $\rightarrow P(\text{Family}) = \frac{1}{4}$

8 Sport $\rightarrow P(\text{Sport}) = 1$

1 Luxury $\rightarrow P(\text{Luxury}) = \frac{1}{8}$

Pour le class C1, on a :

3 Family $\rightarrow P(\text{Family}) = \frac{3}{4}$

0 Sport $\rightarrow P(\text{Sport}) = 0$

7 Luxury $\rightarrow P(\text{Luxury}) = \frac{7}{8}$

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

$$\text{Gini(Sport)} = 1 - 1 - 0 = 0$$

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

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

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

5 Small

7 Medium

4 Large

4 Extra Large

Pour C0 :

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

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

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

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

Pour C1 :

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

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

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

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

⇒

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

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

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

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

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

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

➔

$$\text{Gain}(\text{Gender}) = 0.5 - 0.48 = 0.02$$

$$\text{Gain (Car Type)} = 0.5 - 0.1622 = 0.3378$$

$$\text{Gain (Shirt Size)} = 0.5 - 0.4911 = 0.0089$$

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 :

$$(P_{c+}) = \frac{4}{9}$$

$$(P_{c-}) = \frac{5}{9}$$

On a :

$$\text{Entropy} = -\frac{p}{p+n} \log_2 \left(\frac{p}{p+n} \right) - \frac{n}{p+n} \log_2 \left(\frac{n}{p+n} \right)$$

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

$$\text{Entropy (Class)} = (P_{c+}) \log_2 (P_{c+}) - (P_{c-}) \log_2 (P_{c-})$$

$$= -\frac{4}{9} \log_2 \left(\frac{4}{9} \right) - \frac{5}{9} \log_2 \left(\frac{5}{9} \right) = 0.9911$$

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

Pour a1 :

(4T) avec

$$(3T+) \quad P(T+) = \frac{3}{4}$$

Et \rightarrow

$$(1T-) \quad P(T-) = \frac{1}{4}$$

(5F) avec

$$(1F+) \quad P(F+) = \frac{1}{5}$$

Et \rightarrow

$$(4F-) \quad P(F-) = \frac{4}{5}$$

$$\text{Entropy (a1)} = \frac{4}{9} \left(-\frac{3}{4} \log_2 \left(\frac{3}{4} \right) - \frac{1}{4} \log_2 \left(\frac{1}{4} \right) \right) + \frac{5}{9} \left(-\frac{1}{5} \log_2 \left(\frac{1}{5} \right) - \frac{4}{5} \log_2 \left(\frac{4}{5} \right) \right) = 0.7616$$

Pour a2 :

(5T) avec

$$(2T+) \quad P(T+) = \frac{2}{5}$$

Et \rightarrow

$$(3T-) \quad P(T-) = \frac{3}{5}$$

(4F) avec

$$(2F+) \quad P(F+) = \frac{2}{4}$$

Et \rightarrow

(2F-)

$$P(F-) = \frac{2}{4}$$

$$\text{Entropy}(a_2) = \frac{5}{9} \left(-\frac{2}{5} \log_2 \left(\frac{2}{5} \right) - \frac{3}{5} \log_2 \left(\frac{3}{5} \right) \right) + \frac{4}{9} \left(-\frac{2}{4} \log_2 \left(\frac{2}{4} \right) - \frac{2}{4} \log_2 \left(\frac{2}{4} \right) \right) = 0.9838$$

$$\text{Gain}(a_1) = 0.9911 - 0.7616 = \underline{\underline{0.2294}}$$

$$\text{Gain}(a_2) = 0.9911 - 0.9838 = \underline{\underline{0.0072}}$$

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

On organise les valeurs de a_3 , de plus grand à plus petit,

| Instance | A3 | Target class |
|----------|-----|--------------|
| 1 | 1.0 | + |
| 6 | 3.0 | - |
| 4 | 4.0 | + |
| 3 | 5.0 | - |
| 9 | 5.0 | - |
| 2 | 6.0 | + |
| 5 | 7.0 | - |
| 8 | 7.0 | + |
| 7 | 8.0 | - |

On constate que la valeur 2.0 n'est pas trouvée dans le tableau.

On calcule le gain de a_3 , en supposant la valeur 2.0, d'où

D'après les valeurs de a_3 , on a 3 valeurs qui sont positives et supérieures à 2.0 et une seule valeur qui est positive et inférieure à 2.0

Aussi on a 5 valeurs qui sont négatives et supérieurs à 2.0 et 0 valeurs qui est négative et inférieur à 2.0.

Donc,

$$\text{Gain (a3)} = 0.1427$$

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

$$\text{Gain (a1)} = \underline{0.2294}$$

$$\text{Gain (a2)} = \underline{0.0072}$$

$$\text{Gain (a3)} = \underline{0.1427}$$

⇒ **A1 est la meilleure.**

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

On rappelle que :

Pour a1 :

(4T) avec

$$(3T+) \quad P(T+) = \frac{3}{4}$$

Et →

$$(1T-) \quad P(T-) = \frac{1}{4}$$

(5F) avec

(1F+)

$$P(F+) = \frac{1}{5}$$

Et →

(4F-)

$$P(F-) = \frac{4}{5}$$

Pour a2 :

(5T) avec

(2T+)

$$P(T+) = \frac{2}{5}$$

Et →

(3T-)

$$P(T-) = \frac{3}{5}$$

(4F) avec

(2F+)

$$P(F+) = \frac{2}{4}$$

Et →

(2F-)

$$P(F-) = \frac{2}{4}$$

On va calculer les erreurs de chacun d'eux :

$$erreur(a1) = \frac{4}{9} \left(1 - \frac{3}{4}\right) + \frac{5}{9} \left(1 - \frac{4}{5}\right) = \frac{1}{9} + \frac{1}{9} = \frac{2}{9}$$

Et

$$erreur(a2) = \frac{5}{9} \left(1 - \frac{3}{5}\right) + \frac{4}{9} \left(1 - \frac{2}{4}\right) = \frac{2}{9} + \frac{2}{9} = \frac{4}{9}$$

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

On va calculer les Gini index :

$$Gini(a1) = \frac{4}{9} \left(1 - \frac{3^2}{4^2} - \frac{1^2}{4^2} \right) + \frac{5}{9} \left(1 - \frac{1^2}{5^2} - \frac{4^2}{5^2} \right) = 0.3444$$

$$Gini(a2) = \frac{5}{9} \left(1 - \frac{2^2}{5^2} - \frac{3^2}{5^2} \right) + \frac{4}{9} \left(1 - \frac{2^2}{4^2} - \frac{2^2}{4^2} \right) = 0.4889$$

Merci pour Votre attention