INSTITUT SUPERIEUR D'INFORMATIQUE
ET DES TECHNIQUES DE COMMUNICATION – HAMMAM SOUSSE

Data Mining
TD 3: Decision Tree

Réalisé par : Ghada GHANNEY
Enseignant : Mr. khmaies Abdallah
Classe: 3DNI « 2 »

Année Universitaire 2021 – 2022

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

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

Gini Index for a given node t :

$$Gini\ Index = 1 - \sum_{i=0}^{c-1} p_i(t)^2$$

P(C0) = 10/20 = 0.5
P(C1) = 10/20 = 0.5

Gini Index = $1 - ((0.5)^2 + (0.5)^2) = 0.5$

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

Dans C0: {{Gini(1) = Gini(2) = Gini(3)= Gini(4)= Gini(4)= Gini(5)= Gini(6)= Gini(7)= Gini(8)= Gini(9)= Gini(10) = 0}}

Dans C1: {{Gini(11) = Gini(12) = Gini(13)= Gini(14)=Gini(15)= Gini(16)= Gini(17)= Gini(18)= Gini(19)= Gini(20) = 0}}

Gini Index de ID attribute =0

c. Compute the Gini index for the Gender attribute.

	M	F
C0	6	4
C1	4	6

P(M) = (6/10) + (4/10)
P(F) = (4/10) + (6/10)

Gini Index de M = $1 - ((0.6)^2 + (0.4)^2) = 0.48$
Gini Index de F = $1 - ((0.4)^2 + (0.6)^2) = 0.48$

Gini Index Globale = ½ * Gini M + ½ * Gini F = 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

Gini Index (Family Car)= $1 - ((1/4)^2 + (3/4)^2) = 0.375$
Gini Index (Sports Car)= 0
Gini Index (Luxury Car)= $1 - ((1/8)^2 + (7/8)^2) = 0.21875$

Gini Index Globale = 1/3 * Gini F +1/3 * Gini S +1/3 * Gini L = 0.1979

e. Compute the Gini index for Shirt size the attribute using multiway

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

Gini Index (Small)= 1 – ((3/5)² + (2/5)²) = 0.48
Gini Index (Medium)= 1 – ((3/7)² + (4/7)²) =0.4898
Gini Index (Large)= 1 – ((2/4)² + (2/4)²) =0.5
Gini Index (Extra Small)= 1 – ((2/4)² + (2/4)²) =0.5

Gini Index Globale = 1/4 * Gini S +1/4 * Gini M +1/4 * Gini L + ¼ * = 0.49245

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

Type is better 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.

Customer ID cannot be used for prediction since Gini Index = 0 andustomers are assigned to Customer ID

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

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

Entropy

$$Entropy = -\sum_{i=0}^{c-1} p_i(t) log_2 p_i(t)$$

Entropy= −4/9 log2(4/9) − 5/9 log2(5/9) = 0.9911.

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

Gain = Entropy(S) – I(Attribute)

Entropy de a1 = 4/9 * [-(3/4)log2(3/4) – (1/4)log2(1/4)] + 5/9 * [- (1/5)log2(1/5) – (4/5)log2(4/5)] = 0.7616

Gain de a1 = 0.9911-0.7616 = 0.2294

Entropy de a2 = 5/9 * [-(2/5)log2(2/5) – (3/5)log2(3/5)] + 4/9 * [- (2/4)log2(2/4) – (2/4)log2(2/4)] = 0.9839

Gain de a2 = 0.9911-0.9839= 0.0072

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

a3	Class label	Entropy	Info Gain
1.0	+	0.8484	0.1427
3.0	-	0.9885	0.0026
4.0	+	0.9183	0.0728
5.0 5.0	- -	0.9839	0.0072
6.0	+	0.9728	0.0183
7.0 7.0	+ -	0.8889	0.1022

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

The best split is a1

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

According to the misclassification error rate, a1 is the best

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

Pour l'attribut a1, the gini Index est :
4/9 *[1-(3/4) ^{2}-(1/4) ^{2}] + 5/9 * [1- (1/5) ^{2} –(4/5) ^{2}] = 0.3444

Pour l'attribut a2, the gini Index est :
5/9 *[1-(2/5) ^{2}-(3/5) ^{2}] + 4/9 * [1- (2/4) ^{2} –(2/4) ^{2}] = 0.4889

a1 is the best split according to the Gini Index