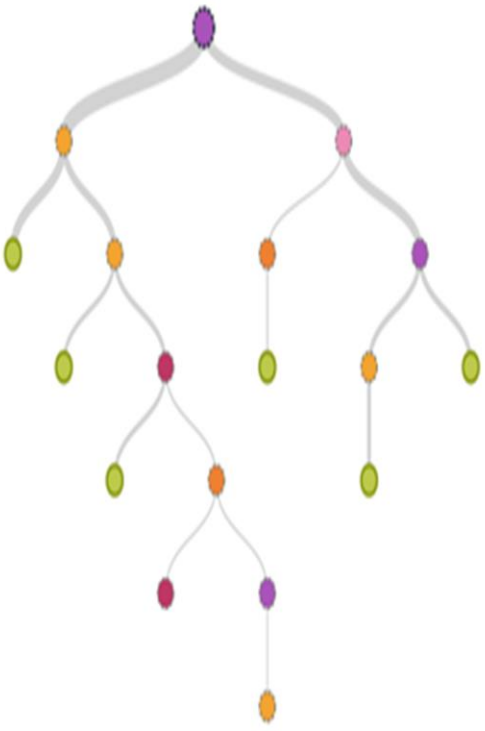
UNIVERSITE DE SOUSSE
Institut Supérieur d'Informatique
Et des Techniques de Communication
Hammam Sousse

جـامعة سوسة
المعهد العالي للإعلامية وتقنيات الاتصال
بحمام سوسة

# Travail TD3

# DECISION TREE



Réalisé par:
Ibtihel Dardouri
3 DNI 2

Année_Universitaire

2021_2022

## Exercises

1. **Consider the training examples shown in Table 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 |

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

⇨ The Gini index for the overall examples is $1 - (5/10)^2 - (5/10)^2 = 0.5$.

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

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

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

⇨ The gini for Male (of Female) is $1 - 0.4^2 - 0.6^2 = 0.48$.
⇨ The Gini index for the Gender attribute is $0.5 \times 0.48 + 0.5 \times 0.48 = 0.48$.

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

⇨ The gini for Family car is $1-(1/4)^2 - (3/4)^2 = 0.375$, Sports car is 0, and Luxury car is 0.2188.

⇨ Gini index is 0.1625.

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

⇨ The gini for Small shirt size is 0.48, Medium shirt size is 0.4898, Large shirt size is 0.5, and Extra Large shirt size is 0.5.
⇨ Gini index for Shirt Size attribute is 0.4914.

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

⇨ 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 attribute cannot be used for prediction (it has no predictive power) since new customers are assigned to new Customer IDs.

**2. Consider the training examples shown in Table 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 | − |

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

⇨ The entropy of the training examples is $-4/9 \log2(4/9) - 5/9 \log2(5/9) = 0.9911$.

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

⇨

The entropy for $a_1$ is

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

Therefore, the information gain for $a_1$ is $0.9911 - 0.7616 = 0.2294$.

The entropy for $a_2$ is

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

Therefore, the information gain for $a_2$ is $0.9911 - 0.9839 = 0.0072$.

**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 |

The best split for $a_3$ occurs at split point equals to 2.

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

⇨ a1

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

⇨ The error rate for a1 is 2/9 and that for a2 is 4/9 so that a1 is the best split attribute.

**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.

| 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 | − |