Préparé par :
Mohammed Yacine BOUHEDADJA

A rendre le mercredi 21/03/2018

Table des matières

Chapitre 1

Indice de GINI

1.1 Definition

Le coefficient de Gini est une mesure statistique, mis au point afin de calculer le degré d'inégalité au sein d'une société (Salaire, niveau de vie, ...). Sa valeur varie entre 0 et 1, il est égal à 0 dans une situation d'égalité parfaite entre les individu de l'ensemble, et est égal à 1 dans dans la situation la plus inégalitaire possible.
Ce critère est utilisé dans la classification binaire, par l'algorithme *CART* (CLASSIFICATION AND REGRESSION DECISION TREE), de façon a calculer l'indice de Gini de tous les testes, et d'en choisir celui ayant la valeur la plus basse.

1.2 Formule

$$\boxed{Gini(S) = 1 - \sum_{j=0}^{n} 1 - P_j^2}$$

Ou P_j représente la fréquence relative à la classe j dans l'ensemble D

1.3 Exemple

Soit l'ensemble de donné suivant :

— Recherche du teste de division

1. Temps
   $Gini(Temps = ensoleillé) = 1 - ((\frac{2}{5})^2 + (\frac{3}{5})^2) = 0.48$
   $Gini(Temps = pluvieux) = 1 - ((\frac{3}{5})^2 + (\frac{2}{5})^2) = 0.48$
   $Gini(Temps = couvert) = 1 - ((\frac{5}{5})^2 = 0$
2. Température
   $Gini(Temperature = haute) = 1 - ((\frac{2}{4})^2 + (\frac{2}{4})^2) = 0.5$
   $Gini(Temperature = basse) = 1 - ((\frac{4}{5})^2 + (\frac{1}{5})^2) = 0.32$
   $Gini(Temperature = moyenne) = 1 - ((\frac{3}{5})^2 + (\frac{2}{5})^2) = 0.39$
3. Humidité
   $Gini(Humidité = haute) = 1 - ((\frac{2}{7})^2 + (\frac{5}{7})^2) = 0.408$
   $Gini(Humidité = normal) = 1 - ((\frac{3}{7})^2 + (\frac{4}{7})^2) = 0.489$
4. Vent
   $Gini(Vent = vrai) = 1 - ((\frac{3}{6})^2 + (\frac{3}{6})^2) = 0.5$
   $Gini(Vent = faux) = 1 - ((\frac{6}{8})^2 + (\frac{2}{8})^2) = 0.378$

FIGURE 1.1 – Ensemble de données de départ

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	Haute	Haute	Faux	Non
Ensoleillé	Haute	Haute	Vrai	Non
Couvert	Haute	Haute	Faux	Oui
Pluvieux	Basse	Haute	Faux	Oui
Pluvieux	Moyenne	Normal	Faux	Oui
Pluvieux	Moyenne	Normal	Vrai	Non
Couvert	Moyenne	Normal	Vrai	Oui
Ensoleillé	Basse	Haute	Faux	Non
Ensoleillé	Moyenne	Normal	Faux	Oui
Pluvieux	Basse	Normal	Faux	Oui
Ensoleillé	Basse	Normal	Vrai	Oui
Couvert	Basse	Haute	Vrai	Oui
Couvert	Haute	Normal	Faux	Oui
Pluvieux	Moyenne	Haute	Vrai	Non

FIGURE 1.2 – Temps = couvert

Temps	Température	Humidité	Vent	Jouer
Couvert	Haute	Haute	Faux	Oui
Couvert	Moyenne	Normal	Vrai	Oui
Couvert	Basse	Haute	Vrai	Oui
Couvert	Haute	Normal	Faux	Oui

Sous ensemble homogène par rapport à la variable de classe, arrêt.

FIGURE 1.3 – Temps != couvert

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	Haute	Haute	Faux	Non
Ensoleillé	Haute	Haute	Vrai	Non
Pluvieux	Basse	Haute	Faux	Oui
Pluvieux	Moyenne	Normal	Faux	Oui
Pluvieux	Moyenne	Normal	Vrai	Non
Ensoleillé	Basse	Haute	Faux	Non
Ensoleillé	Moyenne	Normal	Faux	Oui
Pluvieux	Basse	Normal	Faux	Oui
Ensoleillé	Basse	Normal	Vrai	Oui
Pluvieux	Moyenne	Haute	Vrai	Non

La division se fera par rapport au teste Temps = couvert

Sous ensemble hétérogène par rapport a la variable de classe, continuer son développement.
Recherche de teste de division

1. Temps
   $Gini(Temps = ensoleillé) = 1 - ((\frac{2}{5})^2 + (\frac{3}{5})^2) = 0.48$
   $Gini(Temps = pluvieux) = 1 - ((\frac{3}{5})^2 + (\frac{2}{5})^2) = 0.48$

2. Température
   $Gini(Temperature = haute) = 1 - ((\frac{2}{2})^2 + (0) = 0$
   $Gini(Temperature = basse) = 1 - ((\frac{3}{4})^2 + (\frac{1}{4})^2) = 0.375$
   $Gini(Temperature = moyenne) = 1 - ((\frac{2}{4})^2 + (\frac{2}{4})^2) = 0.5$

3. Humidité
   $Gini(Humidité = haute) = 1 - ((\frac{1}{5})^2 + (\frac{4}{5})^2) = 0.32$
   $Gini(Humidité = normal) = 1 - ((\frac{4}{5})^2 + (\frac{1}{5})^2) = 0.32$

4. Vent
   $Gini(Vent = faux) = 1 - ((\frac{4}{6})^2 + (\frac{2}{6})^2) = 0.45$
   $Gini(Vent = vrai) = 1 - ((\frac{1}{4})^2 + (\frac{3}{4})^2) = 0.375$

La division se fera par rapport au teste Température=haute

FIGURE 1.4 – Température = haute

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	Haute	Haute	Faux	Non
Ensoleillé	Haute	Haute	Vrai	Non

Sous ensemble homogène par rapport à la variable de classe, arrêt.

FIGURE 1.5 – Température != haute

Temps	Température	Humidité	Vent	Jouer
Pluvieux	Basse	Haute	Faux	Oui
Pluvieux	Moyenne	Normal	Faux	Oui
Pluvieux	Moyenne	Normal	Vrai	Non
Ensoleillé	Basse	Haute	Faux	Non
Ensoleillé	Moyenne	Normal	Faux	Oui
Pluvieux	Basse	Normal	Faux	Oui
Ensoleillé	Basse	Normal	Vrai	Oui
Pluvieux	Moyenne	Haute	Vrai	Non

Sous ensemble hétérogène par rapport a la variable de classe, continuer son développement.

Recherche de teste de division

1. Temps
   $Gini(Temps = ensoleillé) = 1 - ((\frac{1}{3})^2 + (\frac{2}{3})^2) = 0.45$
   $Gini(Temps = pluvieux) = 1 - ((\frac{3}{5})^2 + (\frac{2}{5})^2) = 0.48$

2. Température
   $Gini(Temperature = basse) = 1 - ((\frac{3}{4})^2 + (\frac{1}{4})^2) = 0.375$
   $Gini(Temperature = moyenne) = 1 - ((\frac{2}{4})^2 + (\frac{2}{4})^2) = 0.5$

3. Humidité
   $Gini(Humidité = haute) = 1 - ((\frac{1}{3})^2 + (\frac{2}{3})^2) = 0.45$

$Gini(Humidité = normal) = 1 - ((\frac{4}{5})^2 + (\frac{1}{5})^2) = 0.32$

4. Vent
   $Gini(Vent = faux) = 1 - ((\frac{4}{5})^2 + (\frac{1}{5})^2) = 0.32$
   $Gini(Vent = vrai) = 1 - ((\frac{1}{3})^2 + (\frac{2}{3})^2) = 0.45$

La division se fera par rapport à un des testes suivant : Température=haute, vent=faux.
Nous allons le faire par rapport au premier (Humidité=normal).

FIGURE 1.6 – Humidité = normal

Temps	Température	Humidité	Vent	Jouer
Pluvieux	Moyenne	Normal	Faux	Oui
Pluvieux	Moyenne	Normal	Vrai	Non
Ensoleillé	Moyenne	Normal	Faux	Oui
Pluvieux	Basse	Normal	Faux	Oui
Ensoleillé	Basse	Normal	Vrai	Oui

Sous ensemble (A) hétérogène par rapport a la variable de classe, continuer son développement.

FIGURE 1.7 – Humidité != normal

Temps	Température	Humidité	Vent	Jouer
Pluvieux	Basse	Haute	Faux	Oui
Ensoleillé	Basse	Haute	Faux	Non
Pluvieux	Moyenne	Haute	Vrai	Non

Sous ensemble (B) hétérogène par rapport a la variable de classe, continuer son développement.

1. Sous ensemble A
   Recherche de teste de division

   (a) Temps
   $Gini(Temps = ensoleillé) = 1 - ((\frac{1}{3})^2 + (\frac{2}{3})^2) = 0.45$
   $Gini(Temps = pluvieux) = 1 - ((\frac{2}{2})^2 + (0) = 0$

   (b) Température
   $Gini(Temperature = basse) = 1 - ((\frac{2}{2})^2 + (0) = 0$
   $Gini(Temperature = moyenne) = 1 - ((\frac{2}{3})^2 + (\frac{1}{3})^2) = 0.45$

   (c) Vent
   $Gini(Vent = faux) = 1 - ((\frac{3}{3})^2 + (0) = 0$
   $Gini(Vent = vrai) = 1 - ((\frac{1}{2})^2 + (\frac{1}{2})^2) = 0.5$

   La division se ferra par rapport à un des testes suivants : Temps=pluvieux ou Vent=faux, nous allons la faire par rapport au premier teste (Temps=pluvieux).
   Recherche de teste de division

   (a) Température
   $Gini(Temperature = basse) = 1 - ((\frac{1}{1})^2 + (0) = 0$
   $Gini(Temperature = moyenne) = 1 - ((\frac{1}{2})^2 + (\frac{1}{2})^2) = 0.5$

FIGURE 1.8 – Temps != pluvieux

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	Moyenne	Normal	Faux	Oui
Ensoleillé	Basse	Normal	Vrai	Oui

Sous ensemble homogène par rapport a la variable de classe, arrêt.

FIGURE 1.9 – Temps = pluvieux

Temps	Température	Humidité	Vent	Jouer
Pluvieux	Moyenne	Normal	Faux	Oui
Pluvieux	Moyenne	Normal	Vrai	Non
Pluvieux	Basse	Normal	Faux	Oui

Sous ensemble hétérogène par rapport a la variable de classe, continuer son développement.

(b) Vent
$Gini(Vent = faux) = 1 - ((\frac{2}{2})^2 + (0) = 0$
$Gini(Vent = vrai) = 1 - ((\frac{1}{1})^2 + (0) = 0$

Division par rapport au teste Vent = vrai

FIGURE 1.10 – Vent = vrai

Temps	Température	Humidité	Vent	Jouer
Pluvieux	Moyenne	Normal	Vrai	Non

Sous ensemble homogène par rapport à la variable de classe, arrêt. Sous ensemble homogène

FIGURE 1.11 – Vent != Vrai

Temps	Température	Humidité	Vent	Jouer
Pluvieux	Moyenne	Normal	Faux	Oui
Pluvieux	Basse	Normal	Faux	Oui

par rapport à la variable de classe, arrêt.

2. Sous ensemble B
Recherche de teste de division

(a) Temps
$Gini(Temps = ensoleillé) = 1 - ((0) + (\frac{1}{1})^2 = 0$
$Gini(Temps = pluvieux) = 1 - ((\frac{1}{2})^2 + (\frac{1}{2})^2) = 0.5$

(b) Température
$Gini(Temperature = basse) = 1 - ((\frac{1}{2})^2 + (\frac{1}{2})^2) = 0.5$
$Gini(Temperature = moyenne) = 1 - (0) + (\frac{1}{1})^2) = 0$

(c) Vent
$Gini(Vent = faux) = 1 - ((\frac{1}{2})^2 + (\frac{1}{2})^2) = 0.5$
$Gini(Vent = vrai) = 1 - (0)^2 + (\frac{1}{1})^2) = 0$

FIGURE 1.12 – Temps = ensoleillé

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	Basse	Haute	Faux	Non

Division par rapport au teste Temps=ensoleillé
Sous ensemble homogène par rapport à la variable de classe, arrêt.

FIGURE 1.13 – Temps != ensoleillé

Temps	Température	Humidité	Vent	Jouer
Pluvieux	Basse	Haute	Faux	Oui
Pluvieux	Moyenne	Haute	Vrai	Non

Sous ensemble hétérogène par rapport à la variable de classe, continuer son développement.

Recherche de teste de division

(a) Température
$Gini(Temperature = basse) = 1 - ((\frac{1}{1})^2 + (0)) = 0$
$Gini(Temperature = moyenne) = 1 - (0) + (\frac{1}{1})^2) = 0$

(b) Vent
$Gini(Vent = faux) = 1 - ((\frac{1}{2})^2 + (0) = 0$
$Gini(Vent = vrai) = 1 - (0)^2 + (\frac{1}{1})^2) = 0$

Division par rapport au teste Température=basse

FIGURE 1.14 – Température = basse

Temps	Température	Humidité	Vent	Jouer
Pluvieux	Basse	Haute	Faux	Oui

Sous ensemble homogène par rapport à l'attribut de classe, arrêt

FIGURE 1.15 – Température != basse

Temps	Température	Humidité	Vent	Jouer
Pluvieux	Moyenne	Haute	Vrai	Non

Sous ensemble homogène par rapport à l'attribut de classe, arrêt

TOUS LES SOUS ENSEMBLE SONS HOMOGÈNES : CRITÈRE D'ARRÊT ATTEINT.

1.4 L'arbre de décision

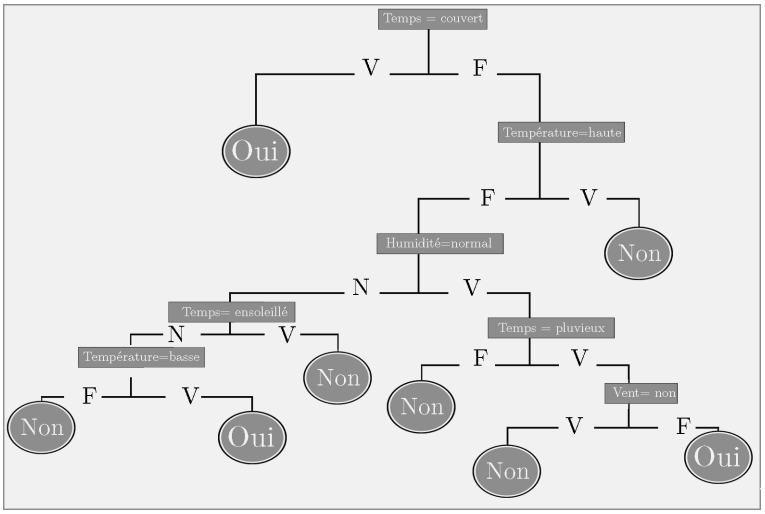

FIGURE 1.16 – Arbre de décision binaire crée avec CART en utilisant l'indice de Gini

Chapitre 2

Gain d'information

2.1 Definition

Le gain d'information est un critère basé sur l'impureté, il utilise la mesure d'entropie comme mesure d'impureté (mesure mise au point par Quinlan en 1987). Ce critère de division est utilisé par l'algorithme ID3.

2.2 Formule

Le Gain d'information d'un attribut T est égal a la différence entre l'entropie pré-division (nœud père p), et l'entropie post-division (par rapport à l'attribut T).

$$\boxed{Gain(p,T) = Entropie(P) - \sum_{j=1}^{n}(P_j * Log_2(P_j))}$$

2.3 Exemple

Soit l'ensemble de donné suivant :

FIGURE 2.1 – Ensemble de données de départ

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	Haute	Haute	Faux	Non
Ensoleillé	Haute	Haute	Vrai	Non
Couvert	Haute	Haute	Faux	Oui
Pluvieux	Basse	Haute	Faux	Oui
Pluvieux	Moyenne	Normal	Faux	Oui
Pluvieux	Moyenne	Normal	Vrai	Non
Couvert	Moyenne	Normal	Vrai	Oui
Ensoleillé	Basse	Haute	Faux	Non
Ensoleillé	Moyenne	Normal	Faux	Oui
Pluvieux	Basse	Normal	Faux	Oui
Ensoleillé	Basse	Normal	Vrai	Oui
Couvert	Basse	Haute	Vrai	Oui
Couvert	Haute	Normal	Faux	Oui
Pluvieux	Moyenne	Haute	Vrai	Non

— Calcule de l'entropie de l'ensemble
Nous avons :
Nombre d'instance : 14

$$Nombre\ de\ classes : 2 \begin{cases} Oui : & 9 \\ Non : & 5 \end{cases}$$

$Entropie(P) = -((\frac{9}{14})Log_2(\frac{9}{14}) + (\frac{5}{14})Log_2(\frac{5}{14})) = 0.94$

— Calcule de l'entropie de chacun des attributs

1. Temps

	Oui	Non	Nb d'instance
Ensoleillé	2	3	5
Pluvieux	3	2	5
Couvert	4	0	4

TABLE 2.1 – Table de comptage de l'attribut Temps

$Entropie(Temps) = \frac{5}{14}(-(\frac{2}{5}Log_2(\frac{2}{5}) + \frac{3}{5}Log_2(\frac{3}{5}))) +$
$\frac{5}{14}(-(\frac{3}{5}Log_2(\frac{3}{5}) + \frac{2}{5}Log_2(\frac{2}{5}))) +$
$\frac{4}{14}(-\frac{4}{4}Log_2(\frac{4}{4}))$
$Entropie(Temps) = 0.69$

2. Température

	Oui	Non	Nb d'instance
Basse	4	1	5
Moyenne	3	2	5
Haute	2	2	4

TABLE 2.2 – Table de comptage de l'attribut Température

$Entropie(Température) = \frac{5}{14}(-(\frac{4}{5}Log_2(\frac{4}{5}) + \frac{1}{5}Log_2(\frac{1}{5}))) +$
$\frac{5}{14}(-(\frac{3}{5}Log_2(\frac{3}{5}) + \frac{2}{5}Log_2(\frac{2}{5}))) +$
$\frac{4}{14}(-(\frac{2}{4}Log_2(\frac{2}{4}) + \frac{2}{4}Log_2(\frac{2}{4})))$
$Entropie(Temps) = 0.89$

3. Humidité

	Oui	Non	Nb d'instance
Haute	4	3	7
Normal	6	1	7

TABLE 2.3 – Table de comptage de l'attribut Humidité

$Entropie(Humidité) = \frac{7}{14}(-(\frac{4}{7}Log_2(\frac{4}{7}) + \frac{3}{7}Log_2(\frac{3}{7}))) +$
$\frac{7}{14}(-(\frac{6}{7}Log_2(\frac{6}{7}) + \frac{1}{7}Log_2(\frac{1}{7})))Entropie(Temps) = 0.79$

4. Vent

$Entropie(Humidité) = \frac{6}{14}(-(\frac{3}{6}Log_2(\frac{3}{6}) + \frac{3}{6}Log_2(\frac{3}{6}))) +$
$\frac{8}{14}(-(\frac{6}{8}Log_2(\frac{6}{8}) + \frac{2}{8}Log_2(\frac{2}{8})))Entropie(Temps) = 0.89$

	Oui	Non	Nb d'instance
Vrai	3	3	6
Faux	6	2	8

TABLE 2.4 – Table de comptage de l'attribut Vent

— Calcul du gain d'information de chaque attribut

$$\left.\begin{array}{l} Gain(Temps) = 0.94 - 0.69 = 0.25 \\ Gain(Température) = 0.94 - 0.89 = 0.05 \\ Gain(Humidité) = 0.94 - 0.69 = 0.15 \\ Gain(Vent) = 0.94 - 0.69 = 0.05 \end{array}\right\} L'attributchoisipourladivisionest : Temps$$

Étant donné que cet attribut se compose de trois valeurs, la division va générer trois sous ensemble, qui sont les suivant :

Temps	Température	Humidité	Vent	Jouer
Couvert	Haute	Haute	Faux	Oui
Couvert	Moyenne	Normal	Vrai	Oui
Couvert	Basse	Haute	Vrai	Oui
Couvert	Haute	Normal	Faux	Oui

FIGURE 2.2 – Sous ensemble A

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	Haute	Haute	Faux	Non
Ensoleillé	Haute	Haute	Vrai	Non
Ensoleillé	Basse	Haute	Faux	Non
Ensoleillé	Moyenne	Normal	Faux	Oui
Ensoleillé	Basse	Normal	Vrai	Oui

TABLE 2.5 – Sous ensemble C

Temps	Température	Humidité	Vent	Jouer
Pluvieux	Basse	Haute	Faux	Oui
Pluvieux	Moyenne	Normal	Faux	Oui
Pluvieux	Moyenne	Normal	Vrai	Non
Pluvieux	Basse	Normal	Faux	Oui
Pluvieux	Moyenne	Haute	Vrai	Non

TABLE 2.6 – Sous ensemble B

Le sous ensemble A, correspondant a la valeur "Temps = couvert" est homogène, son développement s'arrête, contrairement au sous ensemble B et C.

— Développement du sous ensemble B

— Calcul de l'entropie de l'ensemble $Entropie(P) = -((\frac{3}{5})Log_2(\frac{3}{5}) + (\frac{2}{5})Log_2(\frac{2}{5})) = 0.97$

— Calcul de l'entropie de chaque attribut

1. Température $Entropie(Temperature) = \frac{2}{5}(-(\frac{1}{2}Log_2(\frac{1}{2}) + \frac{1}{2}Log_2(\frac{1}{2}))) +$
   $\frac{1}{5}(-(\frac{1}{1}Log_2(\frac{1}{1}))) +$
   $\frac{2}{4}(-(\frac{2}{2}Log_2(\frac{2}{2})))$
   $Entropie(Temperature) = 0.4$
2. Humidité $Entropie(Humidité) = \frac{3}{5}(-(\frac{3}{3}Log_2(\frac{3}{3})) +$
   $\frac{2}{5}(-(\frac{2}{2}Log_2(\frac{2}{2}))$
   $Entropie(Humidité) = 0$
3. Vent $Entropie(Vent) = \frac{3}{5}(-(\frac{1}{3}Log_2(\frac{1}{3}) + \frac{2}{3}Log_2(\frac{2}{3}))) +$
   $\frac{2}{5}(-(\frac{1}{2}Log_2(\frac{1}{2}) + \frac{1}{2}Log_2(\frac{1}{2})))$
   $Entropie(Temps) = 0.95$

— Calcul du gain d'information de chaque attribut

$$\begin{array}{l} Gain(Température) = 0.97 - 0.04 = 0.93 \\ Gain(Humidité) = 0.97 - 0 = 0.97 \\ Gain(Vent) = 0.97 - 0.95 = 0.02 \end{array} \left\{ L'attribut\ choisi\ pour\ la\ division\ est\ :\ Humidité \right.$$

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	Haute	Haute	Faux	Non
Ensoleillé	Haute	Haute	Vrai	Non
Ensoleillé	Basse	Haute	Faux	Non

TABLE 2.7 – Sous ensemble D

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	Moyenne	Normal	Faux	Oui
Ensoleillé	Basse	Normal	Vrai	Oui

TABLE 2.8 – Sous ensemble E

Les sous ensemble résultant de cette division (D, E) sont homogènes, arrêt

— Développement du sous ensemble C

— Calcul de l'entropie de l'ensemble $Entropie(P) = -((\frac{2}{5})Log_2(\frac{2}{5}) + (\frac{3}{5})Log_2(\frac{3}{5})) = 0.97$

— Calcul de l'entropie de chaque attribut

1. Température
   $Entropie(Température) = \frac{2}{5}(-(\frac{2}{2}Log_2(\frac{2}{2}))) +$
   $\frac{3}{5}(-(\frac{2}{3}Log_2(\frac{2}{3}) + \frac{1}{3}Log_2(\frac{1}{3})))$
   $Entropie(Temps) = 0.55$
2. Humidité $Entropie(Humidité) = \frac{3}{5}(-(\frac{2}{3}Log_2(\frac{2}{3}) + \frac{1}{3}Log_2(\frac{1}{3}))) +$
   $\frac{2}{5}(-(\frac{1}{2}Log_2(\frac{1}{2}) + \frac{1}{2}Log_2(\frac{1}{2})))$
   $Entropie(Temps) = 0.95$
3. Vent $Entropie(Vent) = \frac{3}{5}(-(\frac{3}{3}Log_2(\frac{3}{3}))) +$
   $\frac{2}{5}(-(\frac{2}{2}Log_2(\frac{2}{2})))$
   $Entropie(Temps) = 0$

— Calcul du gain d'information de chaque attribut

$$\begin{array}{l} Gain(Température) = 0.97 - 0.55 = 0.42 \\ Gain(Humidité) = 0.97 - 0.95 = 0.92 \\ Gain(Vent) = 0.97 - 0 = 0.97 \end{array} \left\{ L'attributchoisipourladivisionest : Vent \right.$$

Temps	Température	Humidité	Vent	Jouer
Pluvieux	Basse	Haute	Faux	Oui
Pluvieux	Moyenne	Normal	Faux	Oui
Pluvieux	Basse	Normal	Faux	Oui

TABLE 2.9 – Sous ensemble F

Temps	Température	Humidité	Vent	Jouer
Pluvieux	Moyenne	Normal	Vrai	Non
Pluvieux	Moyenne	Haute	Vrai	Non

TABLE 2.10 – Sous ensemble G

TOUS LES SOUS ENSEMBLE SONS HOMOGÈNES : CRITÈRE ARRÊTER ATTEINT.

2.4 L'arbre de décision

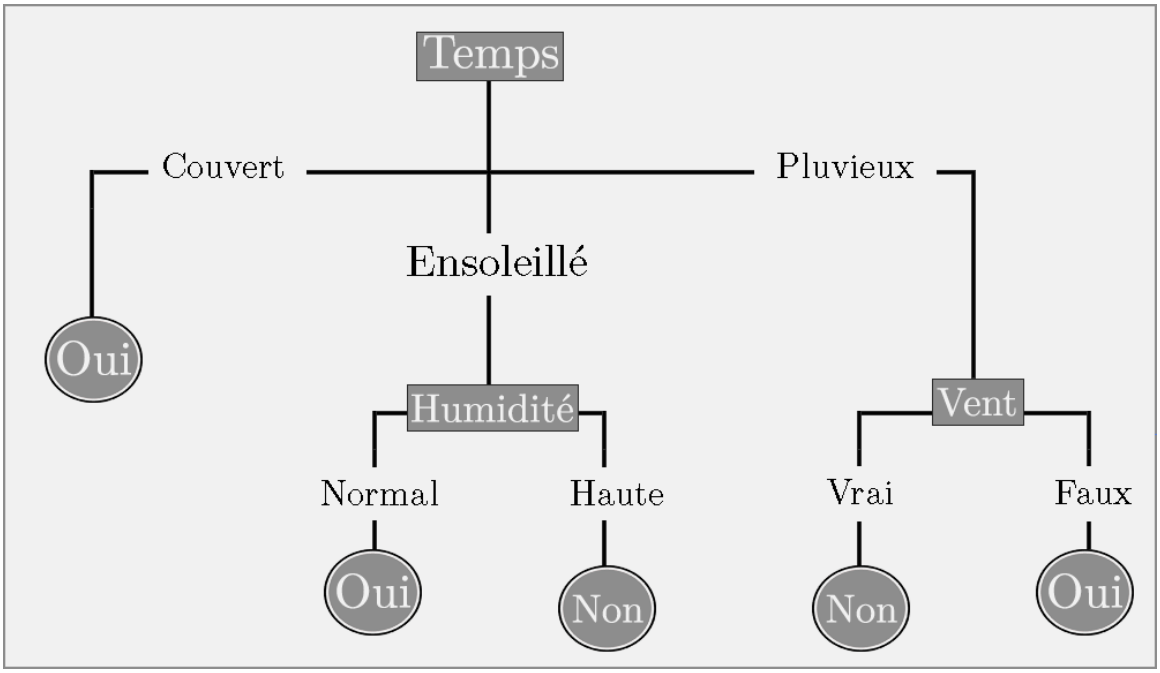

FIGURE 2.3 – Arbre de décision construit avec ID3 en utilisant le Gain d'information

Chapitre 3

RAPPORT DE GAIN

3.1 Definition

Mis au point toujours par QUINLAN, le but de ce critère est de normaliser le rapport de gain, car ce dernier a tendance a favoriser les attributs ayant beaucoup de modalités. Prenons l'exemple d'un attribut qui a autant de modalité qu'il y a de valeur (disons N), avec la Gain d'information, cette attribut sera choisi comme attribut de division, donnant ainsi N sous arbre. Or, le but de la classification est de créer des classes d'individus, et non pas créer une classe pour chaque individu. Avec un calcule normalisé, cet attribut ne sera pas privilégié. Le rapport de gain est utilisé dans l'algorithme C4.5, qui l'amélioration de l'algorithme ID3.

3.2 Formule

Une mesure appelé SPLITINFO est utilisée afin de normaliser le Gain d'information, la formule est la suivante :

$$\boxed{\begin{array}{l} GainRatio(p,T) = \frac{Gain(p,T)}{InfoSplit(p,T)} \\ SplitInfo(p,T) = -\sum_{j=1}^{n} P_j * log_2(P_j) \end{array}}$$

3.3 Exemple

Reprenons le même ensemble de données :
Remarque :
Rappelons que le gain d'information à deja été calculé, nous pouvons donc utiliser directement les résultats suivants :

1. Gain(Temps) = 0.25
2. Gain(Température) = 0.05
3. Gain(Humidité) = 0.15
4. Gain(Vent) = 0.05

FIGURE 3.1 – Ensemble de données de départ

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	Haute	Haute	Faux	Non
Ensoleillé	Haute	Haute	Vrai	Non
Couvert	Haute	Haute	Faux	Oui
Pluvieux	Basse	Haute	Faux	Oui
Pluvieux	Moyenne	Normal	Faux	Oui
Pluvieux	Moyenne	Normal	Vrai	Non
Couvert	Moyenne	Normal	Vrai	Oui
Ensoleillé	Basse	Haute	Faux	Non
Ensoleillé	Moyenne	Normal	Faux	Oui
Pluvieux	Basse	Normal	Faux	Oui
Ensoleillé	Basse	Normal	Vrai	Oui
Couvert	Basse	Haute	Vrai	Oui
Couvert	Haute	Normal	Faux	Oui
Pluvieux	Moyenne	Haute	Vrai	Non

— Calcul d'InfoSplit de chaque attribut

— Temps
Nous avons :

$$Nombre\ de\ modalité : 3 \begin{cases} Ensoleillé : & 5 \\ Pluvieux : & 5 \\ Couvert : & 4 \end{cases}$$

$InfoSplit(Temps) = -(\frac{5}{14}Log_2(\frac{5}{14}) + \frac{5}{14}Log_2(\frac{5}{14}) + \frac{4}{14}Log_2(\frac{4}{14}))$
$InfoSplit(Temps) = 1.577$
— Température $InfoSplit(Température) = -(\frac{5}{14}Log_2(\frac{5}{14}) + \frac{5}{14}Log_2(\frac{5}{14}) + \frac{4}{14}Log_2(\frac{4}{14}))$
$InfoSplit(Température) = 1.577$
— Humidité $InfoSplit(Température) = -(\frac{7}{14}Log_2(\frac{7}{14}) + \frac{7}{14}Log_2(\frac{7}{14})$
$InfoSplit(Température) = 1$
— Vent $InfoSplit(Température) = -(\frac{8}{14}Log_2(\frac{8}{14}) + \frac{6}{14}Log_2(\frac{6}{14})$
$InfoSplit(Température) = 0.98$
— Calcul du rapport de gain de chaque attribut

1. $GainRatio(Temps) = \frac{0.25}{1.577} = 0.15$
2. $GainRatio(Température) = \frac{0.05}{1.577} = 0.03$
3. $GainRatio(Humidité) = \frac{0.15}{1} = 0.15$
4. $GainRatio(Vent) = \frac{0.05}{0.98} = 0.151$

Nous remarquons que le résultat de la recherche de l'attribut de division a changé, nous pouvons donc continuer le calcul avec l'attribut Humidité, qui, avec ses deux valeurs, été désavantagé comparé à l'attribut Température (3 valeurs) lors de l'utilisation du Gain d'information.

Chapitre 4

Chi-Square

4.1 Définition

Le Chi-square est un test statistique, il permet, à partir d'un tableau de contingence, de comparer l'indépendance des variabes, afin de savoir si elles sont liées, autrement dis, il permet de vérifier si les distribution des variables différent les unes des autres. Ce test est utilisé comme critère d'arrêt dans l'algorithme d'arbre de décision CHAID (CHI-SQUARED AUTOMATIC INTERACTION DETECTOR)

4.2 Formule

Le calcule du Chi-square (Xr repose sur deux tableau, ça savoir :
— Tableau d'observation
— Tableau d'estimation

4.3 Exemple

Soit l'ensemble de données suivants :

FIGURE 4.1 – Ensemble de données de départ

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	Haute	Haute	Faux	Non
Ensoleillé	Haute	Haute	Vrai	Non
Couvert	Haute	Haute	Faux	Oui
Pluvieux	Basse	Haute	Faux	Oui
Pluvieux	Moyenne	Normal	Faux	Oui
Pluvieux	Moyenne	Normal	Vrai	Non
Couvert	Moyenne	Normal	Vrai	Oui
Ensoleillé	Basse	Haute	Faux	Non
Ensoleillé	Moyenne	Normal	Faux	Oui
Pluvieux	Basse	Normal	Faux	Oui
Ensoleillé	Basse	Normal	Vrai	Oui
Couvert	Basse	Haute	Vrai	Oui
Couvert	Haute	Normal	Faux	Oui
Pluvieux	Moyenne	Haute	Vrai	Non

— Calcul du Ch-square (X^2) de chaque attribut

1. Temps

	Oui	Non	Total
Ensoleillé	2	3	5
Pluvieux	3	2	5
Couvert	4	0	4
Total	9	5	14

TABLE 4.1 – Tableau d'observation

	Oui	Non
Ensoleillé	$\frac{5*9}{14} = 3.21$	$\frac{5*5}{14} = 1.78$
Pluvieux	$\frac{5*9}{14} = 3.21$	$\frac{5*5}{14} = 1.78$
Couvert	$\frac{4*9}{14} = 2.57$	$\frac{4*5}{14} = 1.43$

TABLE 4.2 – Tableau d'estimation

$$X^2(Temps) = \frac{(2-3.21)^2}{3.21} + \frac{(3-1.78)^2}{1.78} + \frac{(3-3.21)^2}{3.21} + \frac{(2-1.78)^2}{1.78} + \frac{(4-2.57)^2}{2.57} + \frac{(0-1.43)^2}{1.43}$$
$$X^2(Temps) = 3.55$$

2. Température

	Oui	Non	Total
Haute	2	2	4
Basse	4	1	5
Moyenne	3	2	5
Total	9	5	14

TABLE 4.3 – Tableau d'observation

	Oui	Non
Haute	$\frac{4*9}{14} = 2.57$	$\frac{4*5}{14} = 1.43$
Basse	$\frac{5*9}{14} = 3.21$	$\frac{5*5}{14} = 1.78$
Moyenne	$\frac{5*9}{14} = 3.21$	$\frac{5*5}{14} = 1.78$

TABLE 4.4 – Tableau d'estimation

$$X^2(Température) = \frac{(2-2.57)^2}{2.57} + \frac{(2-1.43)^2}{1.43} + \frac{(4-3.21)^2}{3.21} + \frac{(2-1.78)^2}{1.78} + \frac{(3-3.21)^2}{3.21} + \frac{(2-1.78)^2}{1.78}$$
$$X^2(Température) = 0.62$$

3. Humidité

	Oui	Non	Total
Haute	3	4	7
Normal	6	1	7
Total	9	5	14

TABLE 4.5 – Tableau d'observation

	Oui	Non
Haute	$\frac{7*9}{14} = 4.5$	$\frac{7*5}{14} = 2.5$
Basse	$\frac{7*9}{14} = 4.5$	$\frac{7*5}{14} = 2.5$

TABLE 4.6 – Tableau d'estimation

$$X^2(Humidité) = \frac{(3-4.5)^2}{4.5} + \frac{(4-2.5)^2}{2.5} + \frac{(6-4.5)^2}{4.5} + \frac{(1-2.5)^2}{2.5}$$
$$X^2(Humidité) = 2.8$$

4. Vent

	Oui	Non	Total
Vrai	3	3	6
Faux	6	2	8
Total	9	5	14

TABLE 4.7 – Tableau d'observation

	Oui	Non
Haute	$\frac{6*9}{14} = 3.86$	$\frac{6*5}{14} = 2.14$
Basse	$\frac{8*9}{14} = 5.14$	$\frac{8*5}{14} = 2.86$

TABLE 4.8 – Tableau d'estimation

$$X^2(Humidité) = \frac{(3-3.86)^2}{3.86} + \frac{(3-2.14)^2}{2.14} + \frac{(6-5.14)^2}{5.14} + \frac{(2-2.86)^2}{2.86}$$
$$X^2(Temps) = 0.94$$

La division se fera par rapport à l'attribut Temps (celui ayant la plus grande valeur X^2

— Division

Temps	Température	Humidité	Vent	Jouer
Couvert	Haute	Haute	Faux	Oui
Couvert	Moyenne	Normal	Vrai	Oui
Couvert	Basse	Haute	Vrai	Oui
Couvert	Haute	Normal	Faux	Oui

FIGURE 4.2 – Sous ensemble A

Sous ensemble homogène par rapport à la variable de classes, arrêt.

Temps	Température	Humidité	Vent	Jouer
Pluvieux	Basse	Haute	Faux	Oui
Pluvieux	Moyenne	Normal	Faux	Oui
Pluvieux	Moyenne	Normal	Vrai	Non
Pluvieux	Basse	Normal	Faux	Oui
Pluvieux	Moyenne	Haute	Vrai	Non

TABLE 4.9 – Sous ensemble B

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	Haute	Haute	Faux	Non
Ensoleillé	Haute	Haute	Vrai	Non
Ensoleillé	Basse	Haute	Faux	Non
Ensoleillé	Moyenne	Normal	Faux	Oui
Ensoleillé	Basse	Normal	Vrai	Oui

TABLE 4.10 – Sous ensemble C

— Développement du sous ensemble B

1. Température

	Oui	Non	Total
Moyenne	1	2	3
Basse	2	0	2
Total	3	2	5

TABLE 4.11 – Tableau d'observation

	Oui	Non
Haute	$\frac{3*3}{5} = 1.8$	$\frac{3*2}{5} = 1.2$
Basse	$\frac{2*3}{5} = 1.2$	$\frac{2*2}{5} = 0.8$

TABLE 4.12 – Tableau d'estimation

$$X^2(Température) = \frac{(1-1.8)^2}{1.8} + \frac{(2-1.2)^2}{1.2} + \frac{(2-1.2)^2}{1.2} + \frac{(0-0.8)^2}{0.8}$$
$$X^2(Température) = 2.22$$

2. Humidité

	Oui	Non	Total
Haute	1	1	2
Normal	2	1	3
Total	3	2	5

TABLE 4.13 – Tableau d'observation

	Oui	Non
Haute	$\frac{2*3}{5} = 1.2$	$\frac{2*2}{5} = 0.8$
Basse	$\frac{3*3}{5} = 1.8$	$\frac{3*2}{5} = 1.2$

TABLE 4.14 – Tableau d'estimation

$$X^2(Humidité) = \frac{(1-1.2)^2}{1.2} + \frac{(1-0.8)^2}{0.8} + \frac{(2-1.8)^2}{1.8} + \frac{(1-1.2)^2}{1.2}$$
$$X^2(Humidité) = 0.13$$

3. Vent

	Oui	Non	Total
Vrai	0	2	2
Faux	3	0	3
Total	3	2	5

TABLE 4.15 – Tableau d'observation

	Oui	Non
Haute	$\frac{2*3}{5} = 1.2$	$\frac{2*2}{5} = 0.8$
Basse	$\frac{3*3}{5} = 1.8$	$\frac{3*2}{5} = 1.2$

TABLE 4.16 – Tableau d'estimation

$$X^2(Vent) = \frac{(0-1.2)^2}{1.2} + \frac{(2-0.8)^2}{0.8} + \frac{(3-1.8)^2}{1.8} + \frac{(0-1.2)^2}{1.2}$$
$$X^2(Vent) = 5$$

Le sous ensemble B sera donc divisé par rapport a l'attribut Vent.

Temps	Température	Humidité	Vent	Jouer
Pluvieux	Moyenne	Normal	Vrai	Non
Pluvieux	Moyenne	Haute	Vrai	Non

TABLE 4.17 – Sous ensemble D

Temps	Température	Humidité	Vent	Jouer
Pluvieux	Basse	Haute	Faux	Oui
Pluvieux	Moyenne	Normal	Faux	Oui
Pluvieux	Basse	Normal	Faux	Oui

TABLE 4.18 – Sous ensemble E

Les deux sous ensemble C et D sont homogène par rapport a la variable de classe, arrêt.

Développement du sous ensemble C. Faire exactement le même calcul, et l'attribut ayant la plus grande valeur de X^2 sera l'attribut Humidité. Ce qui donnera les deux sous ensemble suivant :

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	Moyenne	Normal	Faux	Oui
Ensoleillé	Basse	Normal	Vrai	Oui

TABLE 4.19 – Sous ensemble F

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	Haute	Haute	Faux	Non
Ensoleillé	Haute	Haute	Vrai	Non
Ensoleillé	Basse	Haute	Faux	Non

TABLE 4.20 – Sous ensemble G

Les deux sous ensemble D et E sont homogène par rapport a la variable de classe, arrêt.

TOUS LES SOUS ENSEMBLE SONS HOMOGÈNES : CRITÈRE D'ARRÊT ATTEINT.

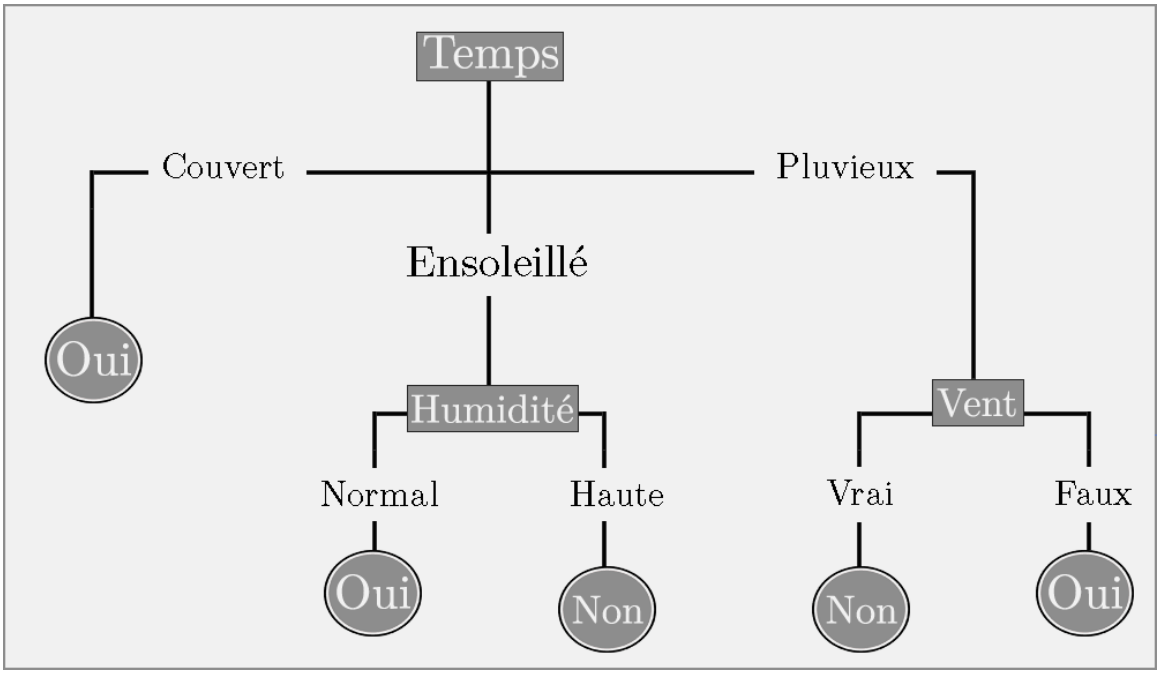

FIGURE 4.3 – Arbre de décision construit avec ID3 en utilisant le Gain d'information