Université Constantine 2
جامعة قسنطينة 2

Apprentissage machine 1

Chapitre 3 : Classification supervisée
Les arbres de décision

Ouadfel Salima
Faculté NTIC/IFA
salima.ouadfel@univ-constantine2.dz

Université Constantine 2
جامعة قسنطينة 2

Apprentissage machine 1

Chapitre 3 : Classification supervisée
Les arbres de décision

Faculté NTIC/IFA
Salima.ouadfel@univ-constantine2.dz

Etudiants concernés

Faculté/Institut	Département	Niveau	Spécialité
Nouvelles technologies	IFA	Master1	STIC

Sélection du meilleur l'attribut avec l'indice de Gini

Soit S l'ensemble de données, l'indice de Gini mesure le degré d'impureté de S et il est exprimé par:

$$\mathbf{Gini(S)} = \sum_{i=1}^{C} p_i(1 - p_i) = 1 - \sum_{i=1}^{C} p_i^2$$

avec $p_i = \frac{|C_i|}{|S|}$ **et** C = nombre de classes

G(S) =0 veut dire que S est pure et G(S) =1 veut dire que S est impure

Sélection du meilleur l'attribut avec l'indice de Gini

La sélection d'un attribut a_j divise l'ensemble S en deux sous-ensembles S_1 et S_2 dont l'indice de Gini défini par $\boldsymbol{Gini}_{a_j}(S_1, S_2)$ est exprimé comme suit:

$$\boldsymbol{Gini}_{a_j}(S_1, S_2) = \frac{|S_1|}{|S|} Gini(S_1) + \frac{|S_2|}{|S|} Gini(S_2)$$

On sélectionne l'attribut a_j qui diminue le plus la valeur de l'impureté.

On cherche a_j tel que

$\boldsymbol{Gini}_{a_j}(S_1, S_2)$ **est minimum**

et $\Delta\boldsymbol{Gini} = \mathbf{Gini}(\mathbf{S}) - \boldsymbol{Gini}_{a_j}(S_1, S_2)$ est **maximum**

Les arbres de décision

Exemple de construction de l'arbre de décision avec l'indice de Gini

Soit S l'ensemble de données du jeu de tennis,

Jour	Temps	Température	Humidité	Vent	Jouer
1	Ensoleillé	chaude	Élevée	faible	non
2	Ensoleillé	chaude	Élevée	fort	non
3	Couvert	chaude	Élevée	faible	oui
4	pluie	douce	Élevée	faible	oui
5	pluie	fraiche	normale	faible	oui
6	pluie	fraiche	normale	fort	non
7	Couvert	fraiche	normale	fort	oui
8	Ensoleillé	douce	Élevée	faible	non
9	Ensoleillé	fraiche	normale	faible	oui
10	Pluie	douce	normale	faible	Oui
11	Ensoleillé	douce	normale	Fort	Oui
12	Couvert	douce	Élevée	Fort	Oui
13	Couvert	chaude	normale	Faible	Oui
14	pluie	douce	Élevée	fort	Non

1- calcul de L'indice de Gini de l'ensemble S

On a 14 données , 9 oui et 5 non

$$\text{Gini}(S) = 1 - \sum_{i=1}^{2} p_i^2 = 1 - \left(\frac{9}{14}\right)^2 - \left(\frac{5}{14}\right)^2$$

$$\boxed{\text{Gini}(S) = 0.459}$$

Exemple de construction de l'arbre de décision avec l'indice de Gini

2- Sélection du meilleur attribut de S

$$Gini_{a_j}(S_1, S_2) = \frac{|S_1|}{|S|} Gini(S_1) + \frac{|S_2|}{|S|} Gini(S_2)$$

a_j=temps

	S_1= {Ensoleillé, couvert} S_2= {pluie}	S_1= {Ensoleillé, pluie} S_2= {couvert}	S_1= {Ensoleillé} S_2= {couvert,pluie}
Gini(S_1)	$1 - \left(\frac{6}{9}\right)^2 - \left(\frac{3}{9}\right)^2 = 0.444$	$1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2 = 0.5$	$1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$
Gini(S_2)	$1 - \left(\frac{3}{5}\right)^2 - \left(\frac{2}{5}\right)^2 = 0.48$	$1 - \left(\frac{4}{4}\right)^2 - \left(\frac{0}{4}\right)^2 =$ **0 (pure)**	$1 - \left(\frac{6}{9}\right)^2 - \left(\frac{3}{9}\right)^2 = 0.444$
$Gini_{a_j}(S_1, S_2)$	$\frac{9}{14} * 0.444 + \frac{5}{14} * 0{,}48 = 0.457$	$\frac{10}{14} * 0.5 + \frac{4}{14} * 0 = 0{,}359$	$\frac{5}{14} * 0.48 + \frac{9}{14} * 0{,}444 = 0.457$

a_j=température

	S_1= {Chaude, fraiche} S_2= {douce}	S_1= {Chaude, douce} S_2= {fraiche}	S_1= {Chaude} S_2= {fraiche, douce}
Gini(S_1)	$1 - \left(\frac{5}{8}\right)^2 - \left(\frac{3}{8}\right)^2 = 0.468$	$1 - \left(\frac{6}{10}\right)^2 - \left(\frac{4}{10}\right)^2 = 0.48$	$1 - \left(\frac{2}{4}\right)^2 - \left(\frac{2}{4}\right)^2 = 0.5$
Gini(S_2)	$1 - \left(\frac{4}{6}\right)^2 - \left(\frac{2}{6}\right)^2 = 0.444$	$1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2 = 0.375$	$1 - \left(\frac{7}{10}\right)^2 - \left(\frac{3}{10}\right)^2 = 0.42$
$Gini_{a_j}(S_1, S_2)$	$\frac{8}{14} * 0.468 + \frac{6}{14} * 0{,}444 = 0.458$	$\frac{10}{14} * 0.48 + \frac{4}{14} * 0 = 0.4498$	$\frac{4}{14} * 0.5 + \frac{10}{14} * 0{,}42 = 0.443$

Exemple de construction de l'arbre de décision avec l'indice de Gini

2- Sélection du meilleur attribut de S

$$\boldsymbol{Gini}_{a_j}(S_1, S_2) = \frac{|S_1|}{|S|} Gini(S_1) + \frac{|S_2|}{|S|} Gini(S_2)$$

a_j=humidité

	S_1= {élevée} S_2= {normale}
Gini(S_1)	$1 - \left(\frac{3}{7}\right)^2 - \left(\frac{4}{7}\right)^2 = 0.489$
Gini(S_2)	$1 - \left(\frac{6}{7}\right)^2 - \left(\frac{1}{7}\right)^2 = 0.244$
$\boldsymbol{Gini}_{a_j}$ (S_1, S_2)	$\frac{7}{14} * 0.489 + \frac{7}{14} * 0{,}244 = 0.368$

a_j=vent

	S_1= {fort} S_2= {faible}
Gini(S_1)	$1 - \left(\frac{3}{6}\right)^2 - \left(\frac{3}{6}\right)^2 = 0.5$
Gini(S_2)	$1 - \left(\frac{6}{8}\right)^2 - \left(\frac{2}{6}\right)^2 = 0.375$
$\boldsymbol{Gini}_{a_j}$ (S_1, S_2)	$\frac{6}{14} * 0.5 + \frac{7}{84} * 0{,}375 = 0.429$

Exemple de construction de l'arbre de décision avec l'indice de Gini

2- Sélection du meilleur attribut de S

On cherche a_j tel que $\Delta \boldsymbol{Gini} = \mathbf{Gini}(\mathbf{S}) - \boldsymbol{Gini}_{a_j}(S_1, S_2)$ est maximum

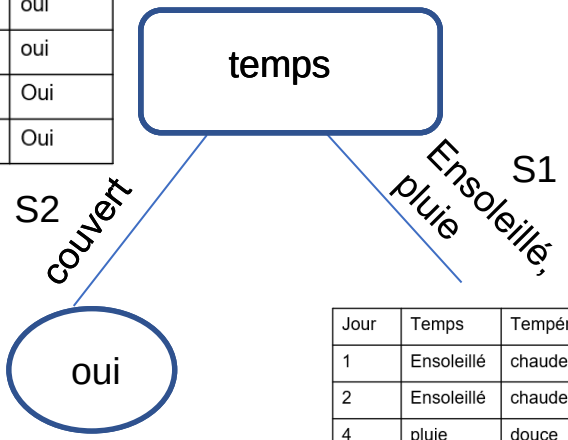
Attribut	(S_1, S_2)	$\boldsymbol{Gini}_{a_j}(S_1, S_2)$	$\Delta \boldsymbol{Gini}$
temps	**S_1= {Ensoleillé, pluie}** **S_2= {couvert}**	$\mathbf{0,359}$	**0.459-0.359 = 0,10**
temperature	S_1= {Chaude} S_2= {fraiche, douce}	0.443	0.459-0.443 = 0.016
humidité	S_1= {élevée} S_2= {normale}	0.368	0.459-0.368 = 0.091
vent	S_1= {fort} S_2= {faible}	0.429	0.459-0.429 = 0.03

L'attribut temps est sélectionné comme nœud racine de l'arbre de décision

Exemple de construction de l'arbre de décision avec l'indice de Gini

3- division de l'ensemble S selon l'attribut temps

Jour	Temps	Température	Humidité	Vent	Jouer
3	Couvert	chaude	Élevée	faible	oui
7	Couvert	fraiche	normale	fort	oui
12	Couvert	fraiche	Élevée	Fort	Oui
13	Couvert	chaude	normale	Faible	Oui



Nœud pure
donc
c'est une feuille

Jour	Temps	Température	Humidité	Vent	Jouer
1	Ensoleillé	chaude	Élevée	faible	non
2	Ensoleillé	chaude	Élevée	fort	non
4	pluie	douce	Élevée	faible	oui
5	pluie	fraiche	normale	faible	oui
6	pluie	fraiche	normale	fort	non
8	Ensoleillé	douce	Élevée	faible	non
9	Ensoleillé	fraiche	normale	faible	oui
10	Pluie	douce	normale	faible	oui
11	Ensoleillé	douce	normale	fort	oui
14	pluie	douce	Élevée	fort	non

Nœud impure
donc à partager

Exemple de construction de l'arbre de décision avec l'indice de Gini

1- calcul de L'indice de Gini de l'ensemble S_1

On a 10 données , 5 oui et 5 non

$$Gini(S_1) = 1 - \sum_{i=1}^{2} p_i^2$$
$$= 1 - \left(\frac{5}{10}\right)^2 - \left(\frac{5}{10}\right)^2$$

$$Gini(S_1) = 0.5$$

Jour	Temps	Température	Humidité	Vent	Jouer
1	Ensoleillé	chaude	Élevée	faible	non
2	Ensoleillé	chaude	Élevée	fort	non
4	pluie	douce	Élevée	faible	oui
5	pluie	fraiche	normale	faible	oui
6	pluie	fraiche	normale	fort	non
8	Ensoleillé	douce	Élevée	faible	non
9	Ensoleillé	fraiche	normale	faible	oui
10	Pluie	douce	normale	faible	oui
11	Ensoleillé	douce	normale	fort	oui
14	pluie	douce	Élevée	fort	non

Exemple de construction de l'arbre de décision avec l'indice de Gini

2- Sélection du meilleur attribut de S_1

a_j=temps

	S_1= {Ensoleillé} S_2= {pluie}
Gini(S_1)	$1-\left(\frac{2}{5}\right)^2-\left(\frac{3}{5}\right)^2=0.48$
Gini(S_{12})	$1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2=0.48$
$\boldsymbol{Gini_{temps}}$ (S_1,S_2)	$\frac{5}{10}*0.48+\frac{5}{10}*0,48=0.48$

a_j=humidité

	S_1= {élevée} S_2= {normale}
Gini(S_1)	$1-\left(\frac{1}{5}\right)^2-\left(\frac{4}{5}\right)^2=0.32$
Gini(S_2)	$1-\left(\frac{4}{5}\right)^2-\left(\frac{1}{5}\right)^2=0.32$
$\boldsymbol{Gini_{a_j}}$ (S_1,S_2)	$\frac{5}{10}*0.32+\frac{5}{10}*0,32=0.32$

a_j=vent

	S_1= {fort} S_2= {faible}
Gini(S_1)	$1-\left(\frac{1}{4}\right)^2-\left(\frac{3}{4}\right)^2=0.375$
Gini(S_2)	$1-\left(\frac{4}{6}\right)^2-\left(\frac{2}{6}\right)^2=0.444$
$\boldsymbol{Gini_{a_j}}$ (S_1,S_2)	$\frac{4}{10}*0.375+\frac{6}{10}*0,444=0.416$

a_j=température

	S_1= {Chaude, fraiche} S_2= {douce}	S_1= {Chaude, douce} S_2= {fraiche}	S_1= {Chaude} S_2= {fraiche, douce}
Gini(S_1)	$1-\left(\frac{2}{5}\right)^2-\left(\frac{3}{5}\right)^2=0.48$	$1-\left(\frac{3}{7}\right)^2-\left(\frac{4}{7}\right)^2=0.489$	$1-\left(\frac{0}{2}\right)^2-\left(\frac{2}{2}\right)^2=0$
Gini(S_2)	$1-\left(\frac{3}{5}\right)^2-\left(\frac{2}{5}\right)^2=0.48$	$1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2=0.444$	$1-\left(\frac{5}{8}\right)^2-\left(\frac{3}{8}\right)^2=0.468$
$\boldsymbol{Gini_{a_j}}$ (S_1,S_2)	$\frac{5}{10}*0.48+\frac{5}{10}*0,48=0.48$	$\frac{7}{10}*0.489+\frac{3}{10}*0.444=0.471$	$\frac{2}{10}*0+\frac{8}{10}*0,468=0.375$

Exemple de construction de l'arbre de décision avec l'indice de Gini

2- Sélection du meilleur attribut de S_1

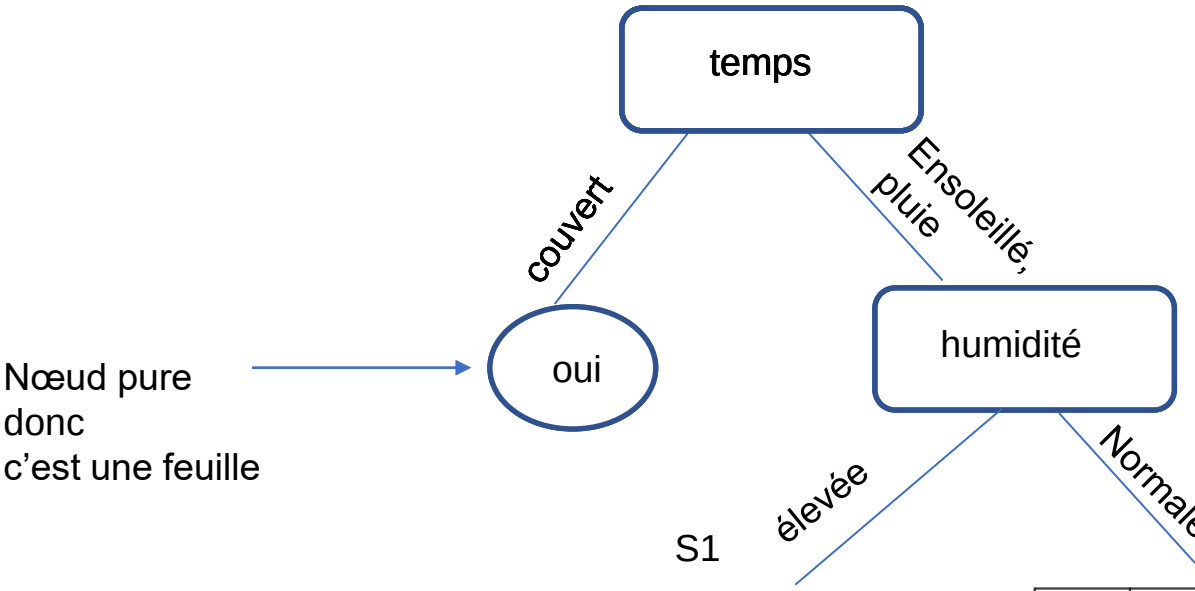
On cherche a_j tel que $\Delta Gini = \mathbf{Gini}(S) - Gini_{a_j}(S_1, S_2)$ est maximum

Attribut	(S_1, S_2)	$Gini_{a_j}(S_1, S_2)$	$\Delta Gini$
temps	S_1= {Ensoleillé} S_2= {pluie}	0.48	0.5-0.48 = 0,02
temperature	S_1= {Chaude} S_2= {fraiche, douce}	0.375	0.5-0,375 = 0.125
humidité	**S_1= {élevée}** **S_2= {normale}**	**0.32**	**0.5-0,32 = 0.18**
vent	S_1= {fort} S_2= {faible}	0.416	0.5-0,416 = 0.084

L'attribut humidité est sélectionné.

Exemple de construction de l'arbre de décision avec l'indice de Gini

3- division de l'ensemble S_1 selon l'attribut temps



Jour	Temps	Température	Humidité	Vent	Jouer
1	Ensoleillé	chaude	Élevée	faible	non
2	Ensoleillé	chaude	Élevée	fort	non
4	pluie	douce	Élevée	faible	oui
8	Ensoleillé	douce	Élevée	faible	non
14	pluie	douce	Élevée	fort	non

Jour	Temps	Température	Humidité	Vent	Jouer
5	pluie	fraiche	normale	faible	oui
6	pluie	fraiche	normale	fort	non
9	Ensoleillé	fraiche	normale	faible	oui
10	Pluie	douce	normale	faible	oui
11	Ensoleillé	douce	normale	fort	oui

Exemple de construction de l'arbre de décision avec l'indice de Gini

1- calcul de L'indice de Gini de l'ensemble S_1

Jour	Temps	Température	Humidité	Vent	Jouer
1	Ensoleillé	chaude	Élevée	faible	non
2	Ensoleillé	chaude	Élevée	fort	non
4	pluie	douce	Élevée	faible	oui
8	Ensoleillé	douce	Élevée	faible	non
14	pluie	douce	Élevée	fort	non

On a 5 données , 1 oui et 4 non

$$\text{Gini}(S_1) = 1 - \sum_{i=1}^{2} p_i^2 = 1 - \left(\frac{1}{5}\right)^2 - \left(\frac{4}{5}\right)^2 = 0.32$$

Exemple de construction de l'arbre de décision avec l'indice de Gini

2- sélection de l'attribut

a_j=temps	S_{11}= {Ensoleillé} S_{12}= {pluie}
Gini(S_{11})	$1-\left(\frac{0}{3}\right)^2-\left(\frac{3}{3}\right)^2=0$
Gini(S_{12})	$1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2=0.5$
$\boldsymbol{Gini_{temps}}$ (S_{11},S_{12})	$\frac{3}{5}*0+\frac{2}{5}*0,5=0.2$

a_j=vent	S_1= {fort} S_2= {faible}
Gini(S_1)	$1-\left(\frac{0}{2}\right)^2-\left(\frac{2}{2}\right)^2=0$
Gini(S_2)	$1-\left(\frac{1}{3}\right)^2-\left(\frac{2}{3}\right)^2=0.444$
$\boldsymbol{Gini_{a_j}}$ (S_1,S_2)	$\frac{2}{5}*0+\frac{3}{5}*0,444=0.266$

a_j=température	S_1= {Chaude} S_2= {douce}
Gini(S_1)	$1-\left(\frac{0}{2}\right)^2-\left(\frac{2}{2}\right)^2=0$
Gini(S_2)	$1-\left(\frac{1}{3}\right)^2-\left(\frac{2}{3}\right)^2=0.444$
$\boldsymbol{Gini_{a_j}}$ (S_1,S_2)	$\frac{2}{5}*0+\frac{3}{5}*0,444$ =0.266

Exemple de construction de l'arbre de décision avec l'indice de Gini

2- Sélection du meilleur attribut de S_1

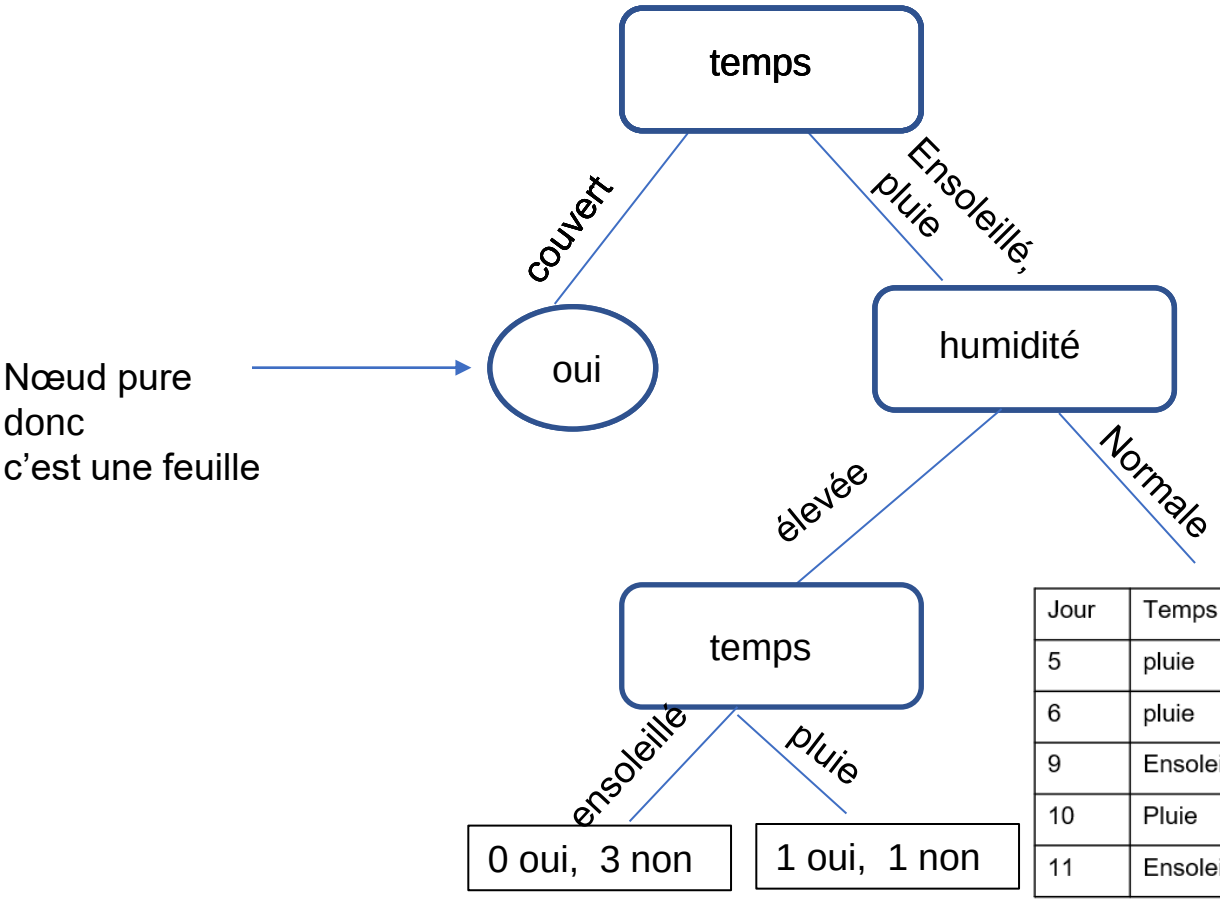
On cherche a_j tel que $\Delta Gini = \mathbf{Gini}(S) - Gini_{a_j}(S_1, S_2)$ est maximum

Attribut	(S_1, S_2)	$Gini_{a_j}(S_1, S_2)$	$\Delta Gini$
temps	**S_1= {Ensoleillé}** **S_2= {pluie}**	**0.2**	**0.32-0.2 = 0.12**
temperature	S_1= {Chaude} S_2= {douce}	0.266	0.32-0,266 = 0.054
vent	S_1= {fort} S_2= {faible}	0.266	0.32-0,266 = 0.054

L'attribut temps est sélectionné.

Exemple de construction de l'arbre de décision avec l'indice de Gini

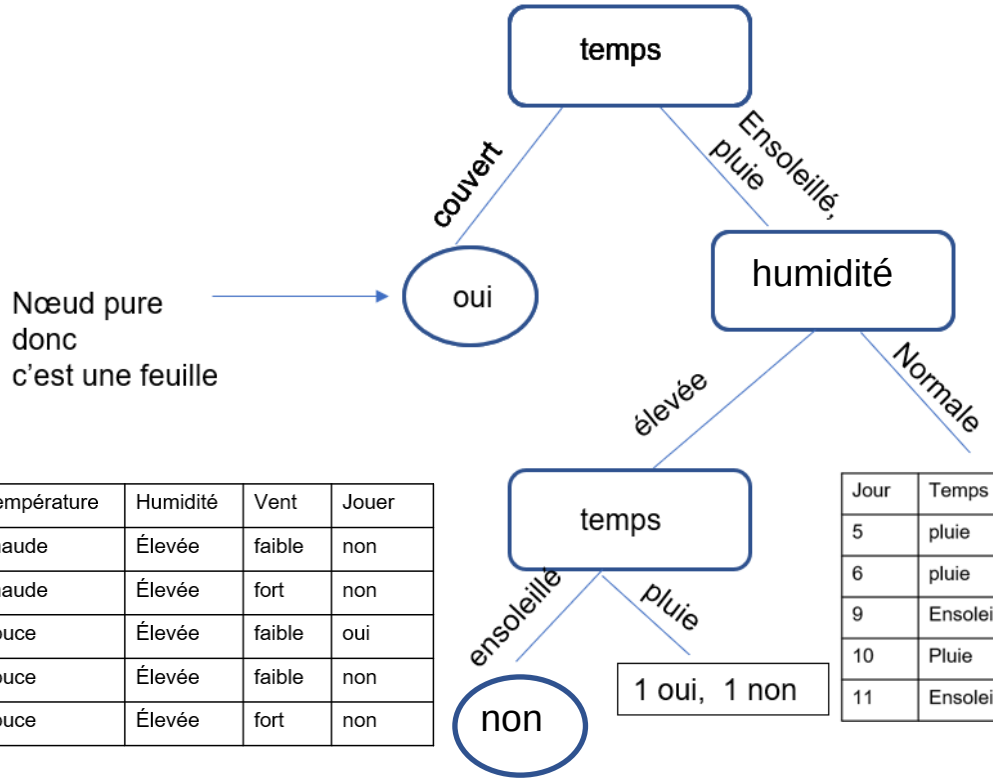
3- division de l'ensemble S_1 selon l'attribut temps



Jour	Temps	Température	Humidité	Vent	Jouer
5	pluie	fraiche	normale	faible	oui
6	pluie	fraiche	normale	fort	non
9	Ensoleillé	fraiche	normale	faible	oui
10	Pluie	douce	normale	faible	oui
11	Ensoleillé	douce	normale	fort	oui

Exemple de construction de l'arbre de décision avec l'indice de Gini

3- division de l'ensemble S_1 selon l'attribut temps

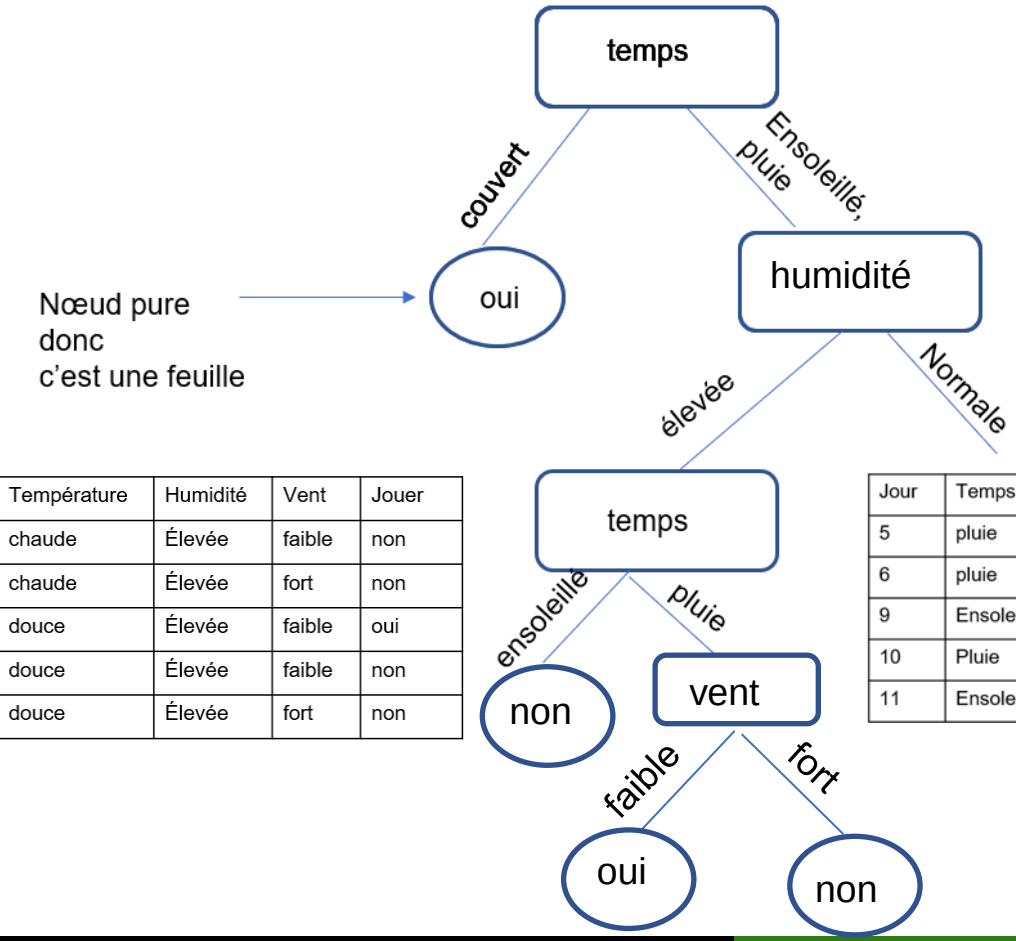


Jour	Temps	Température	Humidité	Vent	Jouer
1	Ensoleillé	chaude	Élevée	faible	non
2	Ensoleillé	chaude	Élevée	fort	non
4	pluie	douce	Élevée	faible	oui
8	Ensoleillé	douce	Élevée	faible	non
14	pluie	douce	Élevée	fort	non

Jour	Temps	Température	Humidité	Vent	Jouer
5	pluie	fraiche	normale	faible	oui
6	pluie	fraiche	normale	fort	non
9	Ensoleillé	fraiche	normale	faible	oui
10	Pluie	douce	normale	faible	oui
11	Ensoleillé	douce	normale	fort	oui

Exemple de construction de l'arbre de décision avec l'indice de Gini

3- division de l'ensemble S_1 selon l'attribut temps



Jour	Temps	Température	Humidité	Vent	Jouer
1	Ensoleillé	chaude	Élevée	faible	non
2	Ensoleillé	chaude	Élevée	fort	non
4	pluie	douce	Élevée	faible	oui
8	Ensoleillé	douce	Élevée	faible	non
14	pluie	douce	Élevée	fort	non

Jour	Temps	Température	Humidité	Vent	Jouer
5	pluie	fraiche	normale	faible	oui
6	pluie	fraiche	normale	fort	non
9	Ensoleillé	fraiche	normale	faible	oui
10	Pluie	douce	normale	faible	oui
11	Ensoleillé	douce	normale	fort	oui

Exemple de construction de l'arbre de décision avec l'indice de Gini

1- calcul de L'indice de Gini de l'ensemble S_2

Jour	Temps	Température	Humidité	Vent	Jouer
5	pluie	fraiche	normale	faible	oui
6	pluie	fraiche	normale	fort	non
9	Ensoleillé	fraiche	normale	faible	oui
10	Pluie	douce	normale	faible	oui
11	Ensoleillé	douce	normale	fort	oui

On a 5 données , 4 oui et 1 non

$$\text{Gini}(S_2) = 1 - \sum_{i=1}^{2} p_i^2 = 1 - \left(\frac{4}{5}\right)^2 - \left(\frac{1}{5}\right)^2 = 0.32$$

Exemple de construction de l'arbre de décision avec l'indice de Gini

2- sélection de l'attribut

temps	S_{11}= {Ensoleillé} S_{12}= {pluie}
Gini(S_{11})	$1-\left(\frac{2}{2}\right)^2-\left(\frac{0}{2}\right)^2=0$
Gini(S_{12})	$1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2=0.444$
$\boldsymbol{Gini_{temps}}$ (S_{11},S_{12})	$\frac{2}{5}*0+\frac{3}{5}*0,444=0.266$

vent	S_1= {fort} S_2= {faible}
Gini(S_1)	$1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2=0.5$
Gini(S_2)	$1-\left(\frac{3}{3}\right)^2-\left(\frac{0}{3}\right)^2=0$
$\boldsymbol{Gini_{a_j}}$ (S_1,S_2)	$\frac{2}{5}*0.5+\frac{3}{5}*0=0.2$

température	S_1= {fraiche} S_2= {douce}
Gini(S_1)	$1-\left(\frac{2}{3}\right)^2-\left(\frac{1}{3}\right)^2=0.444$
Gini(S_2)	$1-\left(\frac{2}{2}\right)^2-\left(\frac{0}{2}\right)^2=0$
$\boldsymbol{Gini_{a_j}}$ (S_1,S_2)	$\frac{3}{5}*0.444+\frac{2}{5}*0$ =0.266

Exemple de construction de l'arbre de décision avec l'indice de Gini

2- Sélection du meilleur attribut de S_2

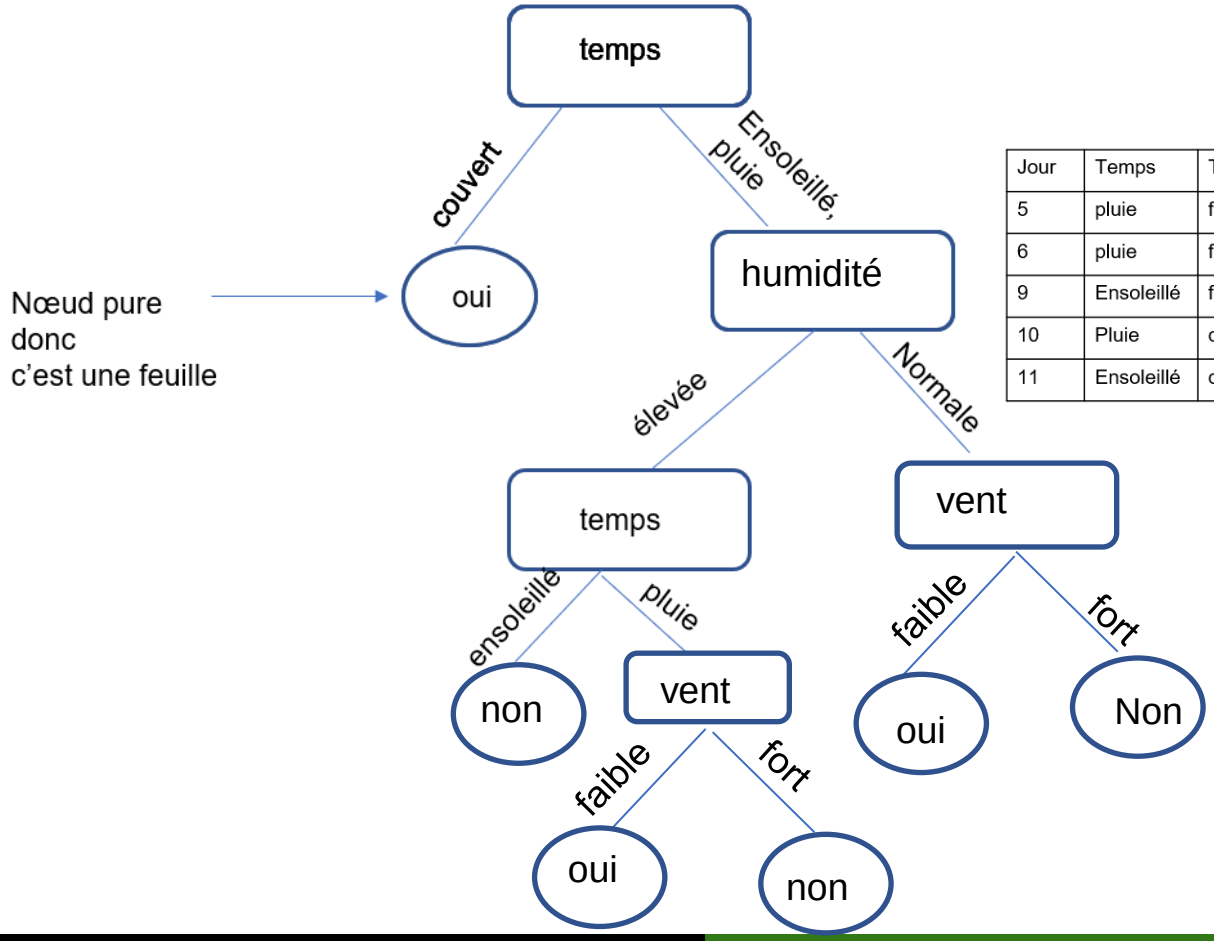
On cherche a_j tel que $\Delta Gini = \mathbf{Gini}(S) - Gini_{a_j}(S_1, S_2)$ est maximum

Attribut	(S_1, S_2)	$Gini_{a_j}(S_1, S_2)$	$\Delta Gini$
temps	S_1= {Ensoleillé} S_2= {pluie}	0.266	0.32-0.266 = 0.054
temperature	S_1= {fraiche} S_2= {douce}	0.266	0.32-0,266 = 0.054
vent	**S_1= {fort}** **S_2= {faible}**	**0.2**	**0.32-0,2 = 0.12**

L'attribut vent est sélectionné.

Exemple de construction de l'arbre de décision avec l'indice de Gini

3- division de l'ensemble S_2 selon l'attribut temps

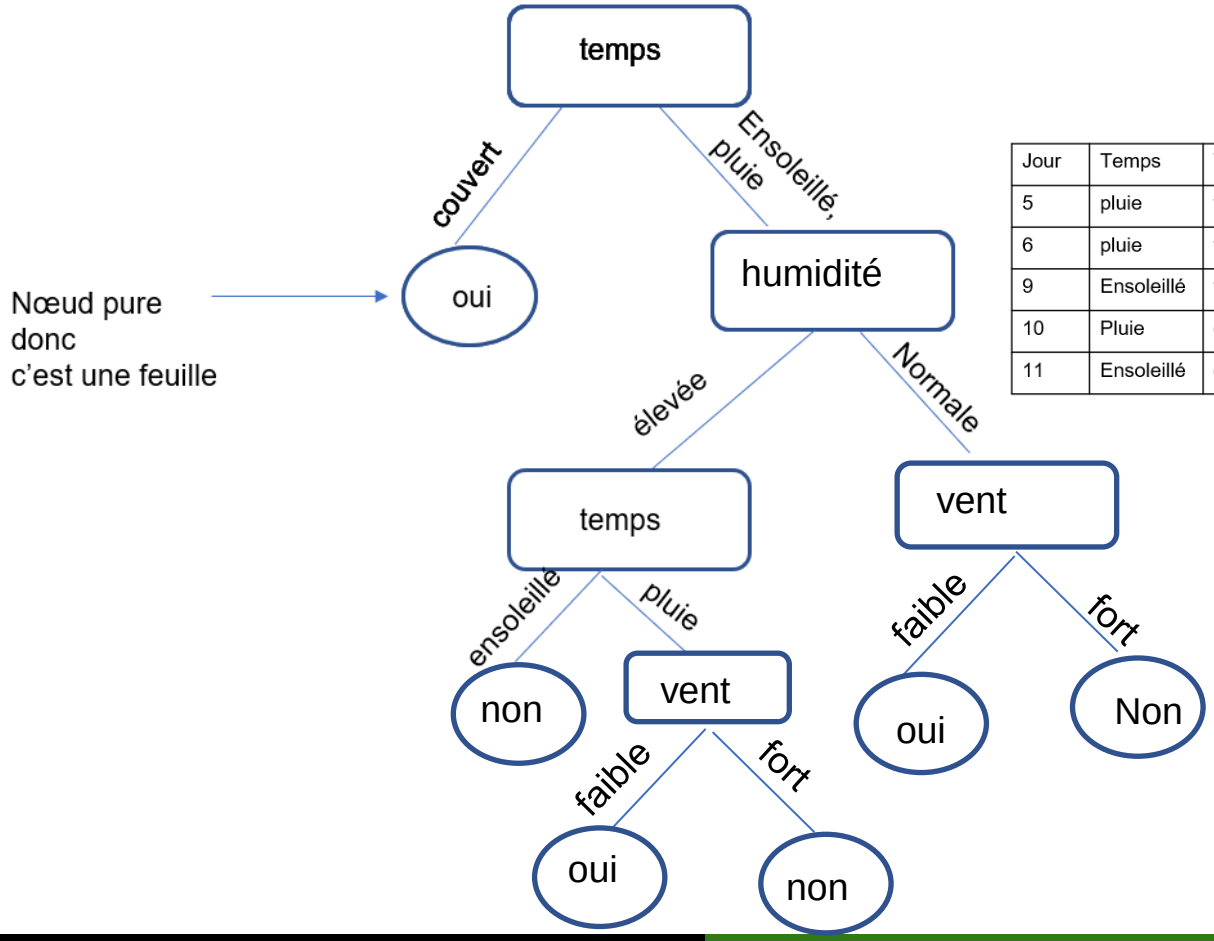


Jour	Temps	Température	Humidité	Vent	Jouer
5	pluie	fraiche	normale	faible	oui
6	pluie	fraiche	normale	fort	non
9	Ensoleillé	fraiche	normale	faible	oui
10	Pluie	douce	normale	faible	oui
11	Ensoleillé	douce	normale	fort	oui

Exemple de construction de l'arbre de décision avec l'indice de Gini

4- Prédiction

(Ensoleillé, Fraîche, Élevée, Fort) est classée comme « non » ;
(Ensoleillé, Fraîche, Normale, Fort) est classée comme « oui » ;
(Pluie, Chaude, Normale, Faible) est classée comme « oui » ;
(Pluie, Fraîche, Élevée, Fort) est classée comme « non ».



Jour	Temps	Température	Humidité	Vent	Jouer
5	pluie	fraiche	normale	faible	oui
6	pluie	fraiche	normale	fort	non
9	Ensoleillé	fraiche	normale	faible	oui
10	Pluie	douce	normale	faible	oui
11	Ensoleillé	douce	normale	fort	oui

Attributs numériques versus attributs catégoriques

Jour	Temps	Température	humidité	vent	Jouer au tennis ?
1	Ensoleillé	Chaude	Élevée	Faible	Non
2	Ensoleillé	Chaude	Élevée	Fort	Non
3	Couvert	Chaude	Élevée	Faible	Oui
4	Pluie	Tiède	Élevée	Faible	Oui
5	Pluie	Fraîche	Normale	Faible	Oui
6	Pluie	Fraîche	Normale	Fort	Non
7	Couvert	Fraîche	Normale	Fort	Oui
8	Ensoleillé	Tiède	Élevée	Faible	Non
9	Ensoleillé	Fraîche	Normale	Faible	Oui
10	Pluie	Tiède	Normale	Faible	Oui
11	Ensoleillé	Tiède	Normale	Fort	Oui
12	Couvert	Tiède	Élevée	Fort	Oui
13	Couvert	Chaud	Normale	Faible	Oui
14	Pluie	Tiède	Élevée	Fort	Non

Valeurs catégoriques

Jour	Temps	Température	humidité	vent	Jouer au tennis ?
1	Ensoleillé	27,5	Élevée	Faible	Non
2	Ensoleillé	25	Élevée	Fort	Non
3	Couvert	26,5	Élevée	Faible	Oui
4	Pluie	20	Élevée	Faible	Oui
5	Pluie	19	Normale	Faible	Oui
6	Pluie	17,5	Normale	Fort	Non
7	Couvert	17	Normale	Fort	Oui
8	Ensoleillé	21	Élevée	Faible	Non
9	Ensoleillé	19,5	Normale	Faible	Oui
10	Pluie	22,5	Normale	Faible	Oui
11	Ensoleillé	22,5	Normale	Fort	Oui
12	Couvert	21	Élevée	Fort	Oui
13	Couvert	25,5	Normale	Faible	Oui
14	Pluie	20,5	Élevée	Fort	Non

Valeurs numériques

Les arbres de décision

Attributs numériques versus attributs catégoriques

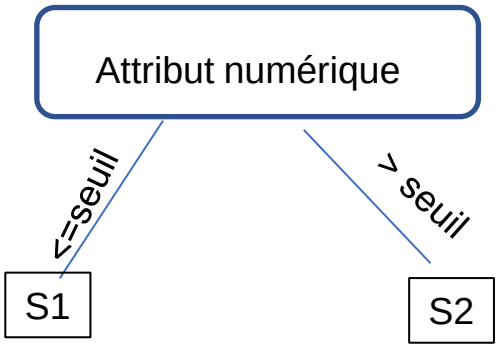
En présence d'attributs numériques:

1- comment sélectionner un attribut s'il est numérique ?

2- Comment subdiviser l'ensemble de données S en des sous-ensemble par rapport à cet attribut numérique

Attributs numériques versus attributs catégoriques

Pour prendre en compte les attributs numériques, C4.5 et CART introduisent un seuil pour la division de l'ensemble en deux sous-ensembles.



Comment choisir le seuil?

Attributs numériques versus attributs catégoriques

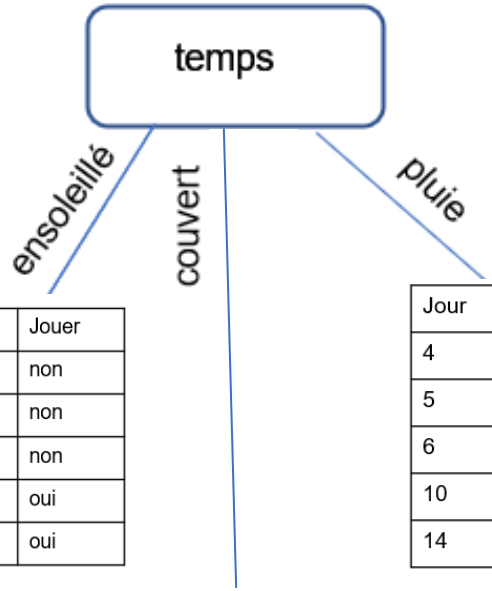
Pour choisir le seuil de séparation, on effectue les étapes suivantes:

1. trier les exemples selon l'ordre croissant de l'attribut quantitatif
2. détecter le changement de classes entre deux exemples consécutifs
3. Si on coupe entre deux valeurs v et w (v<w) alors le seuil est fixé à v (ou bien (v+w)/2).
4. Calculer le gain d'information ou l'index de Gini pour chaque seuil
5. Choisir le seuil qui maximise la mesure d'homogéneité (gain d'information ou indice de Geni)

Exemple de construction de l'arbre de décision avec le gain d'information

On suppose que l'attribut racine est le temps

Jour	Temps	Température	humidité	vent	Jouer au tennis ?
1	Ensoleillé	27,5	Élevée	Faible	Non
2	Ensoleillé	25	Élevée	Fort	Non
3	Couvert	26,5	Élevée	Faible	Oui
4	Pluie	20	Élevée	Faible	Oui
5	Pluie	19	Normale	Faible	Oui
6	Pluie	17,5	Normale	Fort	Non
7	Couvert	17	Normale	Fort	Oui
8	Ensoleillé	21	Élevée	Faible	Non
9	Ensoleillé	19,5	Normale	Faible	Oui
10	Pluie	22,5	Normale	Faible	Oui
11	Ensoleillé	22,5	Normale	Fort	Oui
12	Couvert	21	Élevée	Fort	Oui
13	Couvert	25,5	Normale	Faible	Oui
14	Pluie	20,5	Élevée	Fort	Non



Jour	Temps	Température	Humidité	Vent	Jouer
1	Ensoleillé	27.5	Élevée	faible	non
2	Ensoleillé	25	Élevée	fort	non
8	Ensoleillé	21	Élevée	faible	non
9	Ensoleillé	19.5	normale	faible	oui
11	Ensoleillé	22.5	normale	fort	oui

S1

Jour	Temps	Température	Humidité	Vent	Jouer
4	pluie	20	Élevée	faible	oui
5	pluie	19	normale	faible	oui
6	pluie	17.5	normale	fort	non
10	Pluie	22.5	normale	faible	oui
14	pluie	20.5	Élevée	fort	non

S2

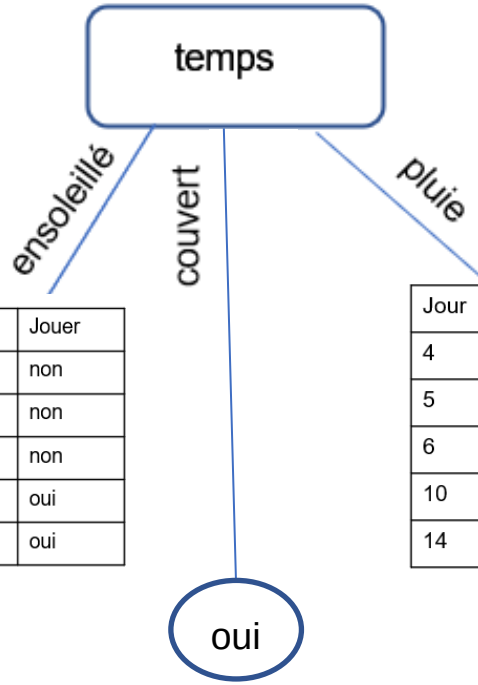
Jour	Temps	Température	Humidité	Vent	Jouer
3	Couvert	26.5	Élevée	faible	Oui
7	Couvert	17	normale	fort	Oui
12	Couvert	21	Élevée	fort	Oui
13	Couvert	25.5	normale	faible	Oui

S3

Exemple de construction de l'arbre de décision avec le gain d'information

Jour	Temps	Température	humidité	vent	Jouer au tennis ?
1	Ensoleillé	27,5	Élevée	Faible	Non
2	Ensoleillé	25	Élevée	Fort	Non
3	Couvert	26,5	Élevée	Faible	Oui
4	Pluie	20	Élevée	Faible	Oui
5	Pluie	19	Normale	Faible	Oui
6	Pluie	17,5	Normale	Fort	Non
7	Couvert	17	Normale	Fort	Oui
8	Ensoleillé	21	Élevée	Faible	Non
9	Ensoleillé	19,5	Normale	Faible	Oui
10	Pluie	22,5	Normale	Faible	Oui
11	Ensoleillé	22,5	Normale	Fort	Oui
12	Couvert	21	Élevée	Fort	Oui
13	Couvert	25,5	Normale	Faible	Oui
14	Pluie	20,5	Élevée	Fort	Non

On suppose que l'attribut racine est le temps



Jour	Temps	Température	Humidité	Vent	Jouer
1	Ensoleillé	27.5	Élevée	faible	non
2	Ensoleillé	25	Élevée	fort	non
8	Ensoleillé	21	Élevée	faible	non
9	Ensoleillé	19.5	normale	faible	oui
11	Ensoleillé	22.5	normale	fort	oui

S1

Jour	Temps	Température	Humidité	Vent	Jouer
4	pluie	20	Élevée	faible	oui
5	pluie	19	normale	faible	oui
6	pluie	17.5	normale	fort	non
10	Pluie	22.5	normale	faible	oui
14	pluie	20.5	Élevée	fort	non

S2

Exemple de construction de l'arbre de décision avec le gain d'information

1- Calcul de l'entropie de l'ensemble S1

$$E(S_1) = -\sum_{i=1}^{2} p_i log_2(p_i) = -\frac{2}{5} log_2 \frac{2}{5} - \frac{3}{5} log_2 \frac{3}{5}$$

$$E(S_1) = 0.971$$

Jour	Temps	Température	Humidité	Vent	Jouer
1	Ensoleillé	27.5	Élevée	faible	non
2	Ensoleillé	25	Élevée	fort	non
8	Ensoleillé	21	Élevée	faible	non
9	Ensoleillé	19.5	normale	faible	oui
11	Ensoleillé	22.5	normale	fort	oui

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut pour diviser S1

1- Si on choisi l'attribut humidité

temps	humidité	Oui	Non	Total
Ensoleillé	Elevée	0	3	3
	Normale	2	0	2

$$IG(S_1, humidité) = 0.971$$

2- Si on choisi l'attribut vent

temps	vent	Oui	Non	Total
Ensoleillé	Fort	1	1	2
	Faible	1	2	3

$$IG(S_1, vent) = 0.020$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut pour diviser S1

3- Si on choisi l'attribut température

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	27.5	Élevée	faible	non
	25	Élevée	fort	non
	21	Élevée	faible	non
	19.5	normale	faible	oui
	22.5	normale	fort	oui

a) Tri des valeurs de l'attribut température par ordre croissant

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	19.5	normale	faible	oui
	21	Élevée	faible	non
	22.5	normale	fort	oui
	25	Élevée	fort	non
	27.5	Élevée	faible	non

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut pour diviser S1

3- Si on choisi l'attribut température

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	19.5	normale	faible	oui
	21	Élevée	faible	non
	22.5	normale	fort	oui
	25	Élevée	fort	non
	27.5	Élevée	faible	non

b) A chaque changement de classe, on considère un seuil:

seuil= 19.5

ou seuil= 21

Ou seuil= 22.5

c) On calcul le gain d'information pour chaque seuil choisi

$IG(S_1, température = 19.5)$

$IG(S_1, température = 21)$

$IG(S_1, température = 22.5)$

d) Choisir le seuil qui maximise le gain d'information

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut pour diviser S1

3- Si on choisi l'attribut température

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	19.5	normale	faible	oui
	21	Élevée	faible	non
	22.5	normale	fort	oui
	25	Élevée	fort	non
	27.5	Élevée	faible	non

c) On calcule le gain d'information pour s=19.5

$$IG(S_1, température = 19.5) = E(S_1) - [\, p(S_{1\ température \leq 19.5}) * E(S_{1\ température \leq 19.5}) + p(S_{1\ température > 19.5}) * E(S_{1\ température > 19.5})]$$

$$p(S_{1\ température \leq 19.5}) * E(S_{1\ température \leq 19.5}) = \frac{1}{5} * \left(-\frac{1}{1} log_2 \frac{1}{1} - \frac{0}{1} log_2 \frac{0}{1}\right)$$

$$p(S_{1\ température > 19.5}) * E(S_{1\ température > 19.5}) = \frac{4}{5} * \left(-\frac{1}{4} log_2 \frac{1}{4} - \frac{3}{4} log_2 \frac{3}{4}\right)$$

$$IG(S_1, température = 19.5) = 0.971 - 0.649 = 0.322$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut pour diviser S1

3- Si on choisi l'attribut température

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	19.5	normale	faible	oui
	21	Élevée	faible	non
	22.5	normale	fort	oui
	25	Élevée	fort	non
	27.5	Élevée	faible	non

c) On calcule le gain d'information pour s=21

$$IG(S_1, température = 21) = E(S_1) - [\, p(S_{1\ température \leq 21}) * E(S_{1\ température \leq 21}) + p(S_{1\ température > 21}) * E(S_{1\ température > 21})]$$

$$p(S_{1\ température \leq 21}) * E(S_{1\ température \leq 21}) = \frac{2}{5} * \left(-\frac{1}{2} log_2 \frac{1}{2} - \frac{1}{2} log_2 \frac{1}{2}\right)$$

$$p(S_{1\ température > 21}) * E(S_{1\ température > 21}) = \frac{3}{5} * \left(-\frac{1}{3} log_2 \frac{1}{3} - \frac{2}{3} log_2 \frac{2}{3}\right)$$

$$IG(S_1, température = 21) = 0.971 - 0.95 = 0.021$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut pour diviser S1

3- Si on choisi l'attribut température

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	19.5	normale	faible	oui
	21	Élevée	faible	non
	22.5	normale	fort	oui
	25	Élevée	fort	non
	27.5	Élevée	faible	non

c) On calcule le gain d'information pour s=22.5

$$IG(S_1, température = 22.5) = E(S_1) - [\, p(S_{1\ température \leq 22.5}) * E(S_{1\ température \leq 22.5}) + p(S_{1\ température > 22.5}) * E(S_{1\ température > 22.5})]$$

$$p(S_{1\ température \leq 22.5}) * E(S_{1\ température \leq 22.5}) = \frac{3}{5} * \left(-\frac{2}{3} log_2 \frac{2}{3} - \frac{1}{3} log_2 \frac{1}{3}\right)$$

$$p(S_{1\ température > 22.5}) * E(S_{1\ température > 22.5}) = \frac{2}{5} * \left(-\frac{0}{2} log_2 \frac{0}{2} - \frac{2}{2} log_2 \frac{2}{2}\right)$$

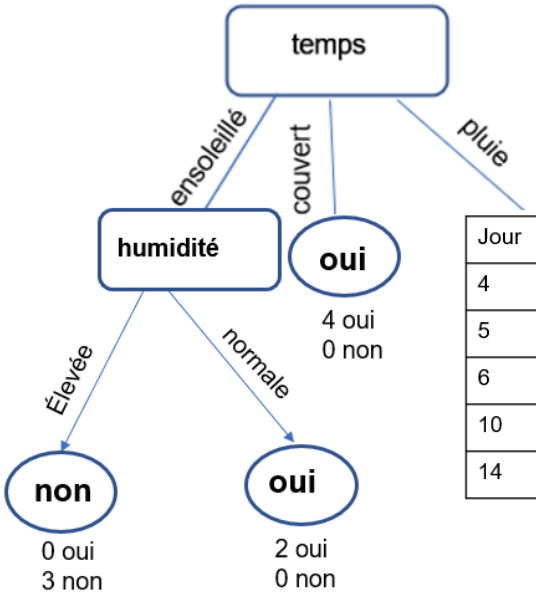
$$IG(S_1, température = 22.5) = 0.971 - 0.55 = 0.420$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Division de l'ensemble S1 selon l'attribut humidité

Attribut	Gain d'information
Température=22.5	0.420
humidité	**0.971**
vent	0.020

Temps	Température	Humidité	Vent	Jouer
Ensoleillé	19.5	normale	faible	oui
	21	Élevée	faible	non
	22.5	normale	fort	oui
	25	Élevée	fort	non
	27.5	Élevée	faible	non



Jour	Temps	Température	Humidité	Vent	Jouer
4	pluie	20	Élevée	faible	oui
5	pluie	19	normale	faible	oui
6	pluie	17.5	normale	fort	non
10	Pluie	22.5	normale	faible	oui
14	pluie	20.5	Élevée	fort	non

S2

Exemple de construction de l'arbre de décision avec le gain d'information

1- Calcul de l'entropie de l'ensemble S2

Jour	Temps	Température	Humidité	Vent	Jouer
4	pluie	20	Élevée	faible	oui
5	pluie	19	normale	faible	oui
6	pluie	17.5	normale	fort	non
10	Pluie	22.5	normale	faible	oui
14	pluie	20.5	Élevée	fort	non

$$E(S_2) = -\sum_{i=1}^{2} p_i log_2(p_i) = -\frac{3}{5} log_2 \frac{2}{5} - \frac{2}{5} log_2 \frac{2}{5}$$

$$E(S_2) = 0.971$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut pour diviser S1

1- Si on choisi l'attribut humidité

temps	humidité	Oui	Non	Total
Pluie	Elevée	1	1	2
	Normale	2	1	3

$$IG(S_1, humidité) = 0.020$$

2- Si on choisi l'attribut vent

temps	vent	Oui	Non	Total
Pluie	Fort	0	2	2
	Faible	3	0	3

$$IG(S_1, vent) = 0.971$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut pour diviser S2

3- Si on choisi l'attribut température

Temps	Température	Humidité	Vent	Jouer
pluie	20	Élevée	faible	oui
	19	normale	faible	oui
	17.5	normale	fort	non
	22.5	normale	faible	oui
	20.5	Élevée	fort	non

a) Tri des valeurs de l'attribut température par ordre croissant

Temps	Température	Humidité	Vent	Jouer
pluie	17.5	normale	fort	non
	19	normale	faible	oui
	20	Élevée	faible	oui
	20.5	Élevée	fort	non
	22.5	normale	faible	oui

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut pour diviser S2

3- Si on choisi l'attribut température

Temps	Température	Humidité	Vent	Jouer
pluie	17.5	normale	fort	non
	19	normale	faible	oui
	20	Élevée	faible	oui
	20.5	Élevée	fort	non
	22.5	normale	faible	oui

b) A chaque changement de classe, on considère un seuil:

seuil= 17.5

ou seuil= 20

Ou seuil= 20.5

c) On calcul le gain d'information pour chaque seuil choisi

$IG(S_1, température = 17.5)$

$IG(S_1, température = 20)$

$IG(S_1, température = 20.5)$

d) Choisir le seuil qui maximise le gain d'information

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut pour diviser S2

Temps	Température	Humidité	Vent	Jouer
pluie	17.5	normale	fort	non
	19	normale	faible	oui
	20	Élevée	faible	oui
	20.5	Élevée	fort	non
	22.5	normale	faible	oui

3- Si on choisi l'attribut température

c) On calcule le gain d'information pour s=17.5

$$IG(S_2, température = 17.5) = E(S_1) - [\, p(S_{2\ température \leq 19.5}) * E(S_{2\ température \leq 19.5}) + p(S_{2\ température > 19.5}) * E(S_{2\ température > 19.5})]$$

$$p(S_{2\ température \leq 17.5}) * E(S_{2\ température \leq 17.5}) = \frac{1}{5} * \left(-\frac{0}{1} log_2 \frac{0}{1} - \frac{1}{1} log_2 \frac{1}{1}\right)$$

$$p(S_{2\ température > 17.5}) * E(S_{2\ température > 17.5}) = \frac{4}{5} * \left(-\frac{3}{4} log_2 \frac{3}{4} - \frac{1}{4} log_2 \frac{1}{4}\right)$$

$$IG(S_2, température = 17.5) = 0.971 - 0.649 = 0.322$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut pour diviser S2

Temps	Température	Humidité	Vent	Jouer
pluie	17.5	normale	fort	non
	19	normale	faible	oui
	20	Élevée	faible	oui
	20.5	Élevée	fort	non
	22.5	normale	faible	oui

3- Si on choisi l'attribut température

c) On calcule le gain d'information pour s=20

$$IG(S_2, température = 20) = E(S_2) - [\, p(S_{2\ température \leq 20}) * E(S_{2\ température \leq 20}) + p(S_{2\ température > 20}) * E(S_{2\ température > 20})]$$

$$p(S_{2\ température \leq 20}) * E(S_{2\ température \leq 20}) = \frac{3}{5} * \left(-\frac{2}{3} log_2 \frac{1}{3} - \frac{1}{3} log_2 \frac{1}{3}\right)$$

$$p(S_{2\ température > 20}) * E(S_{2\ température > 20}) = \frac{2}{5} * \left(-\frac{1}{2} log_2 \frac{1}{2} - \frac{1}{2} log_2 \frac{1}{2}\right)$$

$$IG(S_2, température = 20) = 0.971 - 0.751 = 0.020$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Sélection du meilleur attribut pour diviser S2

Temps	Température	Humidité	Vent	Jouer
pluie	17.5	normale	fort	non
	19	normale	faible	oui
	20	Élevée	faible	oui
	20.5	Élevée	fort	non
	22.5	normale	faible	oui

3- Si on choisi l'attribut température

c) On calcule le gain d'information pour s=20.5

$$IG(S_2, température = 20.5) = E(S_2) - [\, p(S_{2\ température \leq 20.5}) * E(S_{2\ température \leq 20.5}) + p(S_{2\ température > 20.5}) * E(S_{2\ température > 20.5})]$$

$$p(S_{2\ température \leq 20.5}) * E(S_{2\ température \leq 20.5}) = \frac{4}{5} * \left(-\frac{2}{4} log_2 \frac{2}{4} - \frac{2}{4} log_2 \frac{2}{4}\right)$$

$$p(S_{2\ température > 20.5}) * E(S_{2\ température > 20.5}) = \frac{1}{5} * \left(-\frac{1}{1} log_2 \frac{1}{1} - \frac{0}{1} log_2 \frac{0}{1}\right)$$

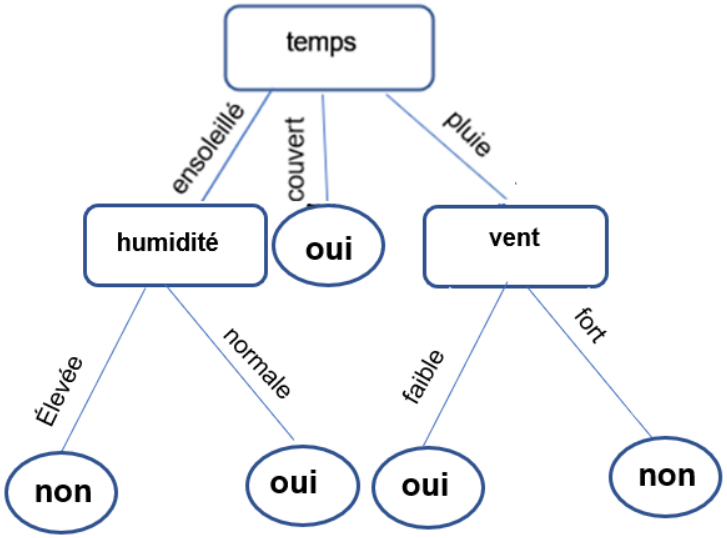
$$IG(S_2, température = 20.5) = 0.971 - 0.8 = 0.17$$

Exemple de construction de l'arbre de décision avec le gain d'information

2- Division de l'ensemble S2 selon l'attribut vent

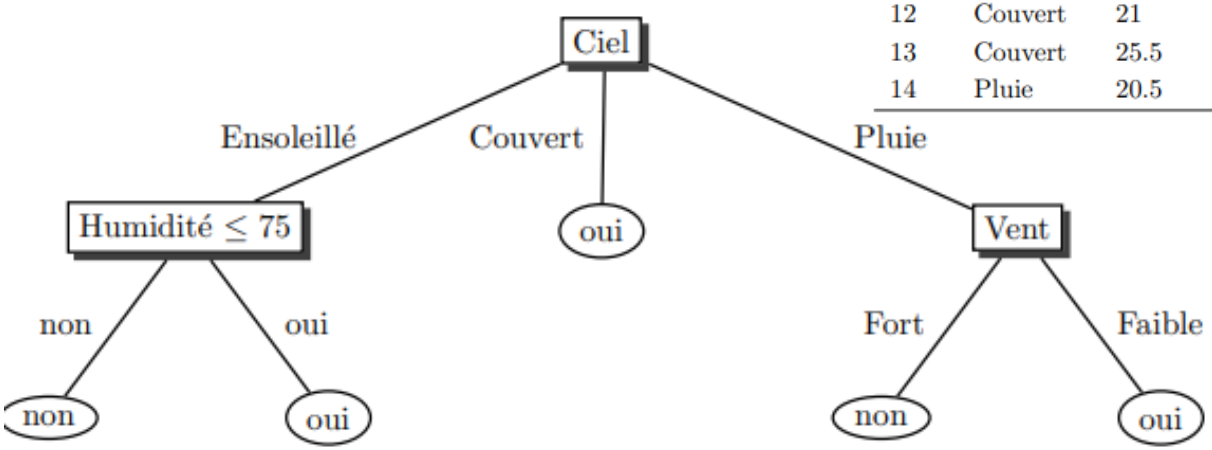
Attribut	Gain d'information
Température=17.5	0.322
humidité	0.020
vent	**0.971**

Temps	Température	Humidité	Vent	Jouer
pluie	20	Élevée	faible	oui
	19	normale	faible	oui
	17.5	normale	fort	non
	22.5	normale	faible	oui
	20.5	Élevée	fort	non



Exemple de construction de l'arbre de décision

Jour	Temps	Température	humidité	vent	Jouer au tennis
1	Ensoleillé	27.5	85	Faible	Non
2	Ensoleillé	25	90	Fort	Non
3	Couvert	26.5	86	Faible	Oui
4	Pluie	20	96	Faible	Oui
5	Pluie	19	80	Faible	Oui
6	Pluie	17.5	70	Fort	Non
7	Couvert	17	65	Fort	Oui
8	Ensoleillé	21	95	Faible	Non
9	Ensoleillé	19.5	70	Faible	Oui
10	Pluie	22.5	80	Faible	Oui
11	Ensoleillé	22.5	70	Fort	Oui
12	Couvert	21	90	Fort	Oui
13	Couvert	25.5	75	Faible	Oui
14	Pluie	20.5	91	Fort	Non



Elagage (optimisation) de l'arbre de décision

L ' élagage consiste à simplifier l'arbre de décision en coupant des branches.

On a deux types :

Pré-élagage: avant la construction de l'arbre
La division d'un sous-ensemble n'est plus nécessaire lorsqu'il est constitué d'un nombre réduit de données ou bien quand la pureté (homogéneité) d'un nœud a atteint un niveau suffisant.

Post-élagage: après la construction de l'arbre
Une fois l'arbre construit, on coupe les branches de l'arbre qui n 'améliorent pas la classification de nouvelles données. Dans ce cas on remplace ces branches par un nœud feuille auquel on associe la classe majoritaire.