Comprendre le principe des arbres de décision et performances

① a- Le meilleur test à choisir à la racine est **t1**, car il va maximiser le gain en information au regard des deux autres tests.

En effet, pour toutes les branches issues de ce test, on arrive à des noeuds dans lesquels les exemples se répartissent très majoritairement dans une seule classe, ce qui mène à un indice de Gini (ou d'entropie) le plus faible.

Le test t2 n'aide en rien la répartition des exemples dans des classes uniques ; le test t3 ne mène que partiellement à une telle discrimination en classes.

b- Choisissons t1.

$A(t_1=1) = \frac{9+}{10} = \oplus$

$A(t_1=2) = \frac{1+4-}{15} = \ominus$

$A(t_1=3) = \frac{5+}{5} = \oplus$

} On choisit la classe majoritaire des exemples arrivant à chaque feuille

Seules 2 erreurs sont à déplorer : l'exemple négatif classé en ⊕ car son $t_1=1$, et l'exemple positif classé en ⊖ car son $t_1=2$.

erreur = $\frac{2}{30} = 0.07$ sur ens. d'apprentissage

② a+b

$t_1(x_1) = +$	$y_1 = +$	OK
$t_1(x_2) = +$	$y_2 = +$	OK
$t_1(x_3) = -$	$y_3 = +$	erreur
$t_1(x_4) = +$	$y_4 = +$	OK
$t_1(x_5) = +$	$y_5 = -$	erreur
$t_1(x_6) = -$	$y_6 = -$	OK
$t_1(x_7) = -$	$y_7 = -$	OK
$t_1(x_8) = +$	$y_8 = +$	OK

Nous comptons 2 erreurs sur 8 exemples

donc erreur = 0.25 sur ens. test.

La matrice de confusion est :

vraie classe \ classe pred	+	−
+	4 TP	1 FN
−	1 FP	2 TN

les erreurs sont hors diagonales, nous vérifions

$e = \frac{2}{8} = 0.25$

c $\underline{\text{rappel}(+) = \frac{TP}{TP + FN} = \frac{4}{5} = 0.8}$

Construction d'un arbre de décision

Construisons un stump : il s'agit de choisir le meilleur test à la racine, ie celui qui maximise le gain en information.

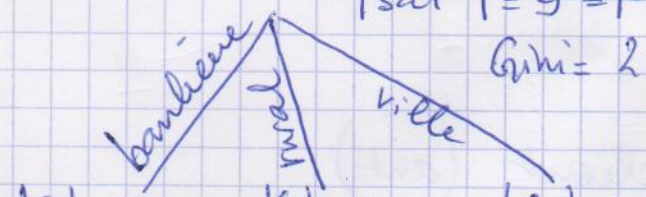
Calculons ce gain pour chaque test basique possible.

insat = +
sat = −

1. Test sur emplacement (epl)

empl $|S| = 14$
$|\text{insat}| = 5 = |+|$
$|\text{sat}| = 9 = |-|$

$\text{Gini} = 2 \times \frac{5 \times 9}{14^2} = 0.46 = \text{Gini}(\mathcal{E})$



banlieue	rural	ville
$\lvert S_1 \rvert = 5$	$\lvert S_2 \rvert = 4$	$\lvert S_3 \rvert = 5$
$\lvert + \rvert = 3$	$\lvert + \rvert = 0$	$\lvert + \rvert = 2$
$\lvert - \rvert = 2$	$\lvert - \rvert = 4$	$\lvert - \rvert = 3$
$\text{Gini} = \frac{2 \times 3 \times 2}{5^2} = 0.48$	$\text{Gini} = \frac{2 \times 0 \times 4}{4^2} = 0$	$\text{Gini} = \frac{2 \times 2 \times 3}{5^2} = 0.48$

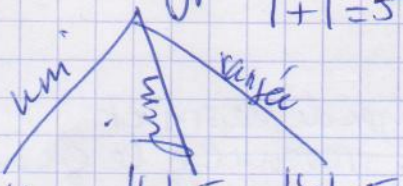
$\text{Gain}(\mathcal{E}, \text{epl}) = \text{Gini}(\mathcal{E}) - \sum_i P_i \, \text{Gini}(S_i)$

$= 0.46 - \left(\frac{5}{14} \cdot 0{,}48 + \frac{4}{14} \cdot 0 + \frac{5}{14} 0.48 \right)$

$= 0{,}117$

2. Test sur type de maison (typ)

typ $|S| = 14$
$|+| = 5 \quad |-| = 9$



villa	semi	maison
$\lvert S_1 \rvert = 4$	$\lvert S_2 \rvert = 5$	$\lvert S_3 \rvert = 5$
$\lvert + \rvert = 2$	$\lvert + \rvert = 1$	$\lvert + \rvert = 2$
$\lvert - \rvert = 2$	$\lvert - \rvert = 4$	$\lvert - \rvert = 3$
$\text{Gini} = \frac{2 \times 2 \times 2}{4^2} = 0.5$	$\text{Gini} = \frac{1 \times 4 \times 2}{5^2} = 0.32$	$\text{Gini} = \frac{2 \times 3 \times 2}{5^2} = 0.48$

$\text{Gain}(\mathcal{E}, \text{typ}) = \text{Gini}(\mathcal{E}) - \sum_i P_i \, \text{Gini}(S_i)$

$= 0.46 - \left(\frac{4}{14} 0.5 + \frac{5}{14} 0.32 + \frac{5}{14} 0.4 \right)$

$= 0{,}031$

3. Test sur revenu (rev)

$|S| = 14$
$|+| = 5$
$|-| = 9$

$Gini(\mathcal{E}) = 0.46$

rev
élevé / bas

$|S_1| = 7$ $\quad$ $|S_2| = 7$
$|+| = 4$ $\quad$ $|+| = 1$
$|-| = 3$ $\quad$ $|-| = 6$

$Gini = \frac{2 \times 4 \times 3}{7^2}$ $\quad$ $Gini = \frac{1 \times 6 \times 2}{7^2}$

$= 0.49$ $\quad$ $= 0.24$

$Gain(\mathcal{E}, rev) = 0.46 - \left(\frac{7}{14} \cdot 0.49 + \frac{7}{14} \cdot 0.24\right)$

$= 0,095$

4. Test sur antériorité du client (ant)

$|S| = 14$
$|+| = 5$
$|-| = 9$

$gini(\mathcal{E}) = 0.46$

ant
oui / non

$|S_1| = 6$ $\quad$ $|S_2| = 8$
$|+| = 3$ $\quad$ $|+| = 2$
$|-| = 3$ $\quad$ $|-| = 6$

$Gini = \frac{3 \times 3 \times 2}{6^2}$ $\quad$ $Gini = \frac{2 \times 6 \times 2}{8^2}$

$= 0.5$ $\quad$ $= 0.375$

$Gain(\mathcal{E}, ant) = 0.46 - \left(\frac{6}{14} \cdot 0,5 + \frac{8}{14} \cdot 0.375\right)$

$= 0,031$

Conclusion Le gain en information est maximal pour le test sur l'emplacement, que nous choisissons donc à la racine.

Nous attribuons la classe majoritaire pour la décision en chaque feuille.