Rappel de Cours n°3

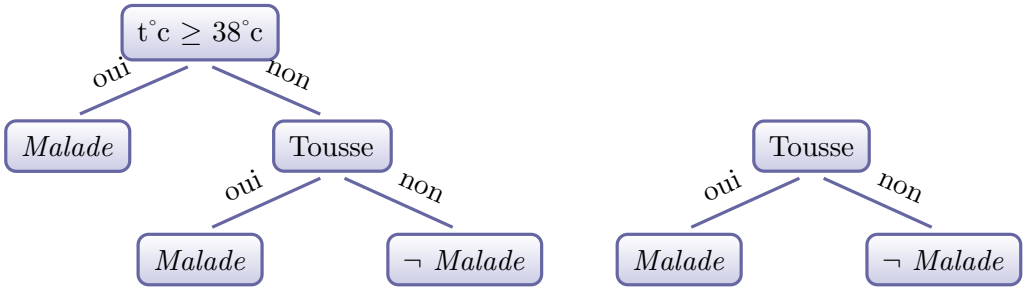
Arbres de Décision

Problème : Disposer de procédures de classification aisément interprétables par des non experts. Les arbres de décision, de par leur représentation graphique des règles de décision, sont une solution à ce problème (parmi d'autre).

Exemple : Supposons que nous sommes médecin. Nous avons 3 patients que nous connaissons et que nous décrivons suivant l'apparition ou non de 2 symptômes (tousse et t°c corporelle). :

	Tousse	t°c ≤ 38	Classe
Patient 1	oui	oui	malade
Patient 2	oui	non	malade
Patient 3	non	non	sein

L'objectif est de représenter ces malades par une structure de donnée permettant de bien les classifier et de diagnostiquer correctement (on l'espère) de nouveaux patients arrivant à l'hôpital.



Ces 2 arbres permettent de classer correctement les 3 patients (l'ensemble d'apprentissage). On se rend déja compte que l'ordre dans lequel sont utilisé les descripteurs influe sur la taille de l'arbre.

Si une nouvelle personne, qui ne tousse pas mais possède une température supérieure à 38°c, se présente à l'hôpital, on va utiliser l'arbre construit pour établir un diagnostic. Dans le cas présent, l'arbre de gauche va considérer qu'elle est malade en se basant uniquement sur sa température. L'arbre de droite va quant-à-lui décider que le patient est sein en se basant uniquement sur le fait qu'il ne tousse pas ! !.

On voit donc que la forme de l'arbre influe non seulement sur la structure mais également sur la décision.

Definition 0.1 ***Arbre parfait*** *Un arbre parfait est un arbre de décision tel que tout les exemples de l'ensemble d'apprentissage soient bien classés.*

L'objectif est d'obtenir l'arbre le plus petit possible (facilitant la recherche) tout en établissant un compromis entre les taux d'erreur sur l'ensemble d'apprentissage et sur l'ensemble de test afin de pouvoir bien généraliser [1]. Pour ce faire, on intervient à 2 niveaux :

1. On sélectionne les attributs qui minimisent la taille de l'arbre tout en classant correctement les exemples de l'ensemble d'apprentissage.
2. On élague certaines branches de manière à garder un pouvoir de généralisation (quitte à faire augmenter l'erreur sur l'ensemble d'apprentissage). Cet élagage peut se faire pendant la construction de l'arbre (pré-élagage) ou après (post-élagage).

1 Sélection des attributs discriminants

Idée générale : Pour la construction de l'arbre de décision, on mesure le "désordre" initial de l'ensemble d'apprentissage. On mesure ensuite le désordre de ce même ensemble après avoir trié celui-ci avec les différents attributs disponible. On compare ensuite ces différentes mesures de "désordre" et on sélectionne l'attribut qui permet de minimiser celui-ci. Dans la pratique, les 2 méthodes les plus connues sont :

1. L'Entropie de l'information (Shannon) [algos ID3 et C4.5 proposés par Quinlan]
   $Entropie(X) = -E[log_2(P(C_i))] = -\sum_{i=1}^{n} P(C_i) * log_2(P(C_i))$
2. Indice de Gini [algo CART, proposé par Breiman]
   $Gini(X) = 1 - \sum_{i=1}^{n} P(C_i)^2 = \sum_{i=1}^{n} P(C_i) * (1 - P(C_i))$

avec $P(C_i) = |C_i|/|X|$ ou X correspond à l'ensemble des exemples et C_i aux différentes classes [2].

Si on prend l'exemple d'un problème où il n'y a que 2 classes, + et -, et si l'on considère que notre ensemble de départ X est constitué de p exemples de la classe + et de n exemples de la classe -, alors :

$E(X) = E(p, n) = -[\frac{p}{p+n} * log(\frac{p}{p+n}) + \frac{n}{p+n} * log(\frac{n}{p+n})]$

$Gini(X) = 1 - [(\frac{p}{p+n})^2 + (\frac{n}{p+n})^2]$

Ont peu également considérer ces 2 grandeurs comme des mesures de la quantité d'information nécessaire pour pouvoir "ranger correctement" X. Dans tout les cas, plus ces valeurs sont importantes, et moins les éléments sont ordonnés.

Le gain d'information apporté par chaque attribut (A), calculé afin de choisir celui dont le gain est le plus élevé, est obtenu comme suit :

$$\text{Gain(A)= F(X) - Reste(A) . avec : } Reste(A) = \sum_{i=1}^{v} \frac{n_i+p_i}{p+n} * F(n_i, p_i)$$

où `v` est le nombre de valuations de A et où `Reste(A)` est la quantité de "désordre" qu'il reste à traiter après avoir utilisé l'attribut A et où `F` est la fonction considérée pour discriminer les attributs (Entropie ou Gini).

Des exercices corrigés associés à l'utilisation de ces 2 grandeurs pour la construction d'un AD sont disponibles à l'adresse suivante : `http://www-desir.lip6.fr/~herpsonc/ia1.htm`

1. Cf le rappel n°2 sur l'apprentissage supervisé pour faire la distinction les 2 types d'erreur
2. On trouve également cette formule avec $P(C_i|X)$ à la place de $P(C_i)$. $P(C_i|X) = P(C_i \cap X)/P(X)$ or $C_i \subset X$ et $X = \Omega$, on retombe donc sur $P(C_i)$