Arbres de décision

Fabien Torre

Université de Lille

Mercredi 30 septembre 2009

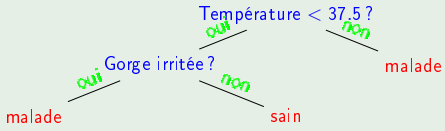
Le formalisme des arbres de décision

Définition : arbre de décision

Arbre binaire, chaque nœud interne porte un test booléen sur un attribut, chaque branche correspond à un résultat du test (vrai ou faux), chaque feuille est étiquetée par une classe.

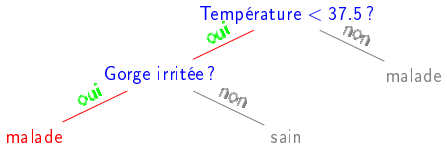
Illustration

Des exemples (des patients), des attributs (`Température` et `Gorge irritée`), des classes (*malade* ou *sain*).



Classification et règles

- Permet de classer un nouvel exemple : `(37.2,oui)`,



- peut être traduit en un système de règles :
  - `si (Temp` $<$ `37.5) et (GorgeIrritée) alors malade`
  - `si (Temp` $<$ `37.5) et non(GorgeIrritée) alors sain`
  - `si (Temp` $\geq$ `37.5) alors malade`

Inférence d'arbres de décision

Objectif : inférer un arbre de décision à partir d'exemples.

- Répartition de la population de patients dans l'arbre ;
- définition d'une méthode d'inférence :
  - comment sélectionner le test à effectuer à un nœud ?
  - comment décider si un nœud est terminal ?
  - quelle classe associée à une feuille ?

Pour le premier point, il est intéressant de savoir mesurer le degré de mélange d'une population.

Mélange

Objectif : mesurer le mélange de deux classes c_1 et c_2 dans un ensemble d'exemples A.

- Utiliser les proportions $p(c_1)$ et $p(c_2)$ dans A ;
- trouver une fonction qui est minimale lorsque $p(c_1) = 0$ ou $p(c_2) = 0$, et maximale lorsque $p(c_1) = p(c_2) = 0.5$;
- fonction de Gini : $1 - p(c_1)^2 - p(c_2)^2$;
- entropie : $-p(c_1).log(p(c_1)) - p(c_2).log(p(c_2))$.

Gain

- Population courante notée p, de n individus ;
- un test divise p en deux sous-ensembles : p_1 de taille n_1 et p_2 de taille n_2 ;
- le gain amené par ce test est donné par :

$$\text{Mélange}(p) - \frac{n_1}{n}.\text{Mélange}(p_1) - \frac{n_2}{n}.\text{Mélange}(p_2)$$

Critère de sélection : choisir le test qui maximise le *gain* du test.

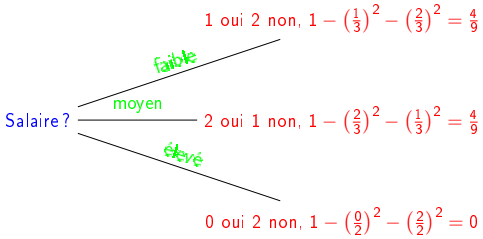
Un jeu de données

id	salaire	âge	résidence	études	internet
1	moyen	moyen	village	oui	oui
2	élevé	moyen	bourg	non	non
3	faible	âgé	bourg	non	non
4	faible	moyen	bourg	oui	oui
5	moyen	jeune	ville	oui	oui
6	élevé	âgé	ville	oui	non
7	moyen	âgé	ville	oui	non
8	faible	moyen	village	non	non

8 clients, 3 ont internet, 5 non, mélange initial (selon Gini) :

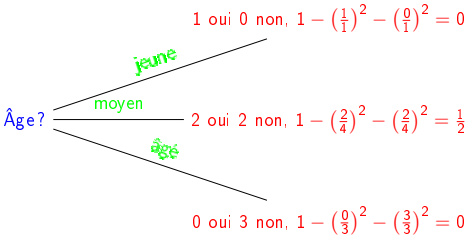
$$1 - \left(\frac{3}{8}\right)^2 - \left(\frac{5}{8}\right)^2 = \frac{15}{32} = 0.46875$$

Tests candidats à la racine : *salaire*



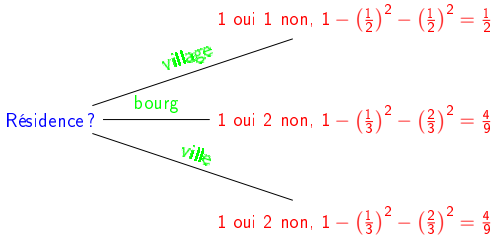
$$\text{Gain} = \frac{15}{32} - \frac{3}{8}.\frac{4}{9} - \frac{3}{8}.\frac{4}{9} - \frac{2}{8}.0 = \frac{13}{96} = 0.135$$

Tests candidats à la racine : *âge*



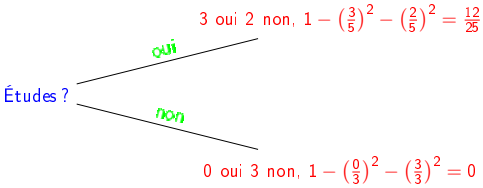
$$\text{Gain} = \frac{15}{32} - \frac{1}{8}.0 - \frac{4}{8}.\frac{1}{2} - \frac{3}{8}.0 = \frac{7}{32} = 0.219$$

Tests candidats à la racine : *résidence*



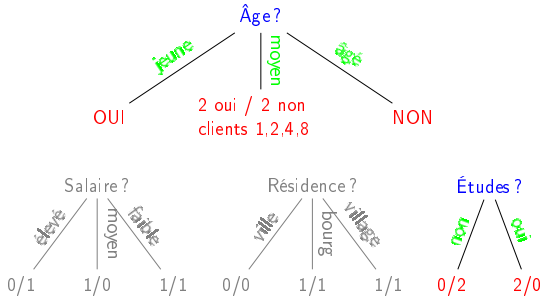
$$\text{Gain} = \frac{15}{32} - \frac{2}{8}.\frac{1}{2} - \frac{3}{8}.\frac{4}{9} - \frac{3}{8}.\frac{4}{9} = \frac{1}{96} = 0.010$$

Tests candidats à la racine : *études*

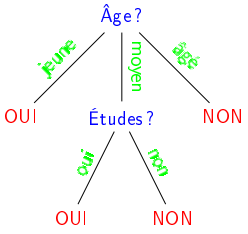


$$\text{Gain} = \frac{15}{32} - \frac{5}{8}.\frac{12}{25} - \frac{3}{8}.0 = \frac{27}{160} = 0.169$$

Premier niveau de l'arbre appris



Arbre finalement appris



Faiblesses

- Algorithme glouton, sans backtrack ;
- transposables en règles avec des règles avec attributs communs... en particulier l'attribut utilisé à la racine !
- difficulté avec les concepts disjonctifs (cf. *agaricus-lepiota*) ;
- faiblesse du codage attributs-valeurs (classification de molécules ?).

Où sont les biais ?

- Biais de langage : arbres/règles ;
- biais de recherche : descendant, ajouter le test qui maximise la diminution du mélange ;
- biais de validation : feuille pure.

arbres et NFL

Mais le *no free lunch theorem* [Wolpert and Macready, 1995] nous dit que tout biais en vaut un autre... et si l'on inversait le biais de recherche [Murphy and Pazzani, 1994, Webb, 1996] ?

Conclusions sur les arbres de décision

- Critère entropique à chaque nœud : arbre petit, classifieur compréhensible ;
- *pourrait-on calculer l'arbre le plus petit ?*
- division du training set à chaque étape : apprentissage rapide ;
- mais rien d'explicite ni de théorique pour minimiser l'erreur en généralisation.
- optimisation possible : élagage de l'arbre appris ;
- on peut créer des attributs (*loto*) ;
- plusieurs implémentations : ID3, C4.5 [Quinlan, 1993], CART ;
- dimension de Vapnik-Chervonenkis ?

Bibliographie I

- Murphy, P. and Pazzani, M. J. (1994).
  Exploring the decision forest.
  In *Computational Learning and Natural Language Workshop, Provincetown*, pages 257–275.
- Quinlan, J. R. (1993).
  C4.5 : programs for machine learning.
  Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.
- Webb, G. I. (1996).
  Further experimental evidence against the utility of occam's razor.
  Journal of Artificial Intelligence Research, 4 :397–417.

Bibliographie II

- Wolpert, D. H. and Macready, W. G. (1995).
  No free lunch theorems for search.
  Technical Report SFI-TR-95-02-010, The Santa Fe Institute.