Arbres de décision

Notes de démonstration, Ift3330

François Paradis

Décembre 2005

Exemple (adapté de http://www.decisiontrees.net/tutorial/1_intro.html)

Emplacement	*Type de maison*	*Revenu*	*Client antérieur?*	*Résultat*
banlieue	Unifamiliale	élevé	non	Insatisfait
banlieue	Unifamiliale	élevé	oui	Insatisfait
rural	Unifamiliale	élevé	non	Satisfait
ville	Jumelée	élevé	non	Satisfait
ville	Jumelée	bas	non	Satisfait
ville	Jumelée	bas	oui	Insatisfait
rural	Jumelée	bas	oui	Satisfait
banlieue	Rangée	élevé	non	Insatisfait
banlieue	Jumelée	bas	non	Satisfait
ville	Rangée	bas	non	Satisfait
banlieue	Rangée	bas	oui	Satisfait
rural	Rangée	élevé	oui	Satisfait
rural	Unifamiliale	bas	non	Satisfait
ville	Rangée	élevé	oui	Insatisfait

Entropie

Quantité moyenne d'information pour classifier un objet:

$$I = -\sum_c p(c) \log_2 p(c)$$

l'entropie est nulle si pour une classe p(c)=1 car les autres classes sont forcément 0 (voir p(rural) ci-dessous)

Dans notre exemple:
$I = -9/14 \log_2 9/14 - 5/14 \log_2 5/14 = 0.947$

Entropie résiduelle pour l'attribut A:

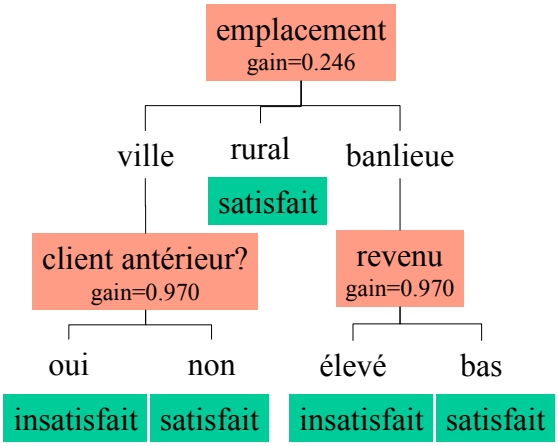
$$I_{res}(A) = -\sum_v p(v) \sum_c p(c \mid v) \log_2 p(c \mid v)$$

où v sont les valeurs possibles de l'attribut A

Dans notre exemple:
I_{res}(emplacement) = p(ville).0.97095+p(banlieue).0.97095+ p(rural).0 = 0.6935

Construction d'un arbre

- à chaque noeud, choisir l'attribut de gain (i.e I-Ires) maximal
- arrêter quand l'entropie est nulle



Construisez cet arbre dynamiquement sur le site de Michael Nashvili:

http://www.decisiontrees.net/tutorial/3_exercise1.html

De meilleures mesures d'impuretés

- Le gain favorise les attributs qui peuvent prendre plusieurs valeurs
  e.g. attribut *date* avec 14 valeurs :
  gain = 0.9403
- Solutions:
  - *binariser* les attributs
  - autres mesures: $gainRatio(A) = \frac{I - I_{res}(A)}{I(A)}$

$$I(A) = -\sum_{v} p(v) \log_2 p(v)$$

De meilleures mesures d'impureté

$$gini(A) = \sum_{v} p(v) \sum_{i \neq j} p(i \mid v) p(j \mid v)$$

pour notre exemple originel + 14 dates différentes:

gini(date) =
p(date1)*(p(satisfait|date1)*p(~satisfait|date1))
+p(date2)*...= 0

gainRatio(date)=0.247

Autre exemple

tiré de Bratko, exercice 18.2

Une maladie *M* est présente dans 25% des cas. Le symptôme *S* est observé chez 75% des patients qui souffrent de *M*, et chez 1/6 des autres patients. Soit les classes *m* et ~*m*. Calculez gainRatio pour l'attribut *S*.

$$I_{res}(S) = -p(s)I(M \mid s) - p(\bar{s})I(M \mid \bar{s})$$

$$I(M \mid S) = p(m \mid S)\log_2 p(m \mid S) + p(\bar{m} \mid S)\log_2 p(\bar{m} \mid S)$$

I=0.8113; p(s) = 0.3125; p(m|s)=0.6; p(~m|s)=0.4
p(m|~s)=0.0909; Ires=0.6056; **gainRatio=0.2296**

Élagage d'arbres

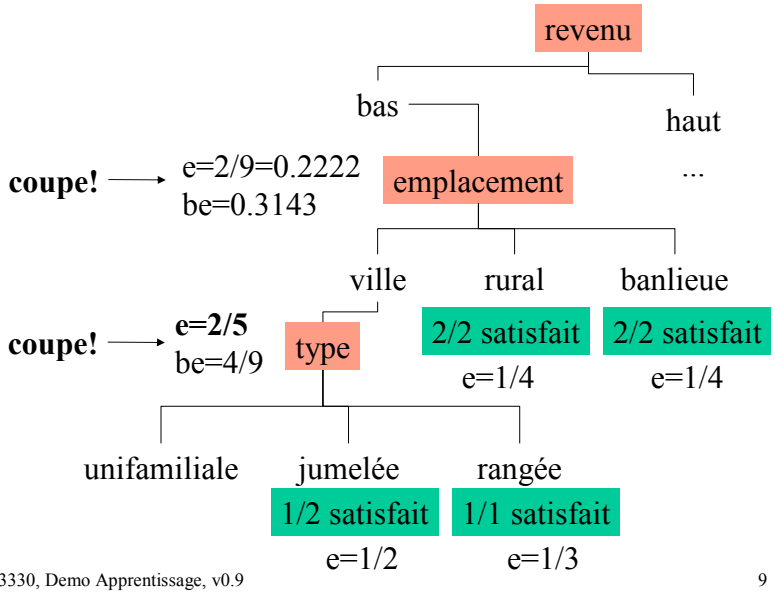
- Problème de surentraînement
- On coupe une branche à un noeud *s* si l'erreur propagée (be) est plus grande que l'erreur statique (e)

$$e(s) = p(classe \neq C \mid s) \approx 1 - \frac{n+1}{N+k}$$

où k est le nb de classes (estimateur de Laplace)

$$be(s) = \sum_i p_i E(s_i) \qquad E(s) = \min(e(s), be(s))$$

Exemple d'élagage



Comparaison

- Couper une branche revient à *généraliser* un arbre.
- Dans l'exemple précédent, le cas de classe insatisfait [revenu=bas, emplacement=ville, type=jumelée] est considéré marginal.
- En coupant une branche on espère diminuer l'erreur (ou augmenter la justesse) de classification. On peut mesurer l'erreur d'un arbre de décision sur un ensemble-test donné:

  erreur = nerr/N, où nerr est le nombre de classes incorrectes produites par l'arbre de décision, et N est la taille de l'échantillon.

Autres approches d'apprentissage

- Supervisées: arbres de décision, naïve Bayes, SVM, NN, boosting/bagging, réseaux de neurones, programmation logique inductive, etc.
- Non supervisées: *clustering*, *principal component analysis*
- Semi-supervisée: *Expectation-maximization*
- Autres: *Reinforcement*, algorithmes génétiques

Logiciels

- **Weka** Data mining software suite
  University of Waikato
  http://www.cs.waikato.ac.nz/~ml/weka/index.html
- **SVMLight** Classification SVM
  Thorsten Joachims, Cornell
  http://www.cs.cornell.edu/People/tj/svm_light/
- **C4.5** Arbres de décision
  Ross Quinlan
  http://www.rulequest.com/Personal/