Arbres d'induction

dans un contexte non classificatoire

Fabio Losa Pau Origoni
Office statistique du Canton du Tessin

Gilbert Ritschard
Dept Econométrie, Université de Genève

EGC, janvier 2005

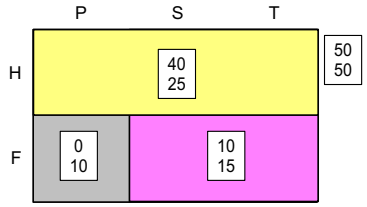
Plan

http://mephisto.unige.ch

1 Principe des arbres de classification

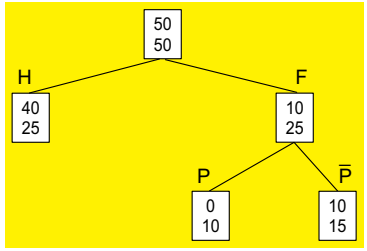
- Arbre de classification
- Arbre de décision
- Arbre d'induction
- Arbre de segmentation

But : Partitionner les données de façon que la distribution de la variable réponse (statut sur marché de l'emploi) diffère le plus possible d'une classe (feuille) à l'autre.



Comment : Par éclatements successifs des nœuds.

- En partant du nœud initial, chercher l'attribut qui génère le meilleur éclatement (le plus discriminant).
- Répéter à chaque nœud
  ⇒ satisfaction d'un critère d'arrêt (gain minimal d'info, taille minimale des nœuds, ...)



Principaux algorithmes

CHAID (Kass, 1980), degré de signification du Khi-2
⇒ éclate selon la variable qui a le plus fort lien statistique avec la variable à prédire

CART (Breiman et al., 1984), indice de Gini
Cherche à obtenir les distributions les plus pures, du type $\begin{pmatrix}1 & 0 & 0 & 0\end{pmatrix}$
Arbres binaires (éclatement en deux seulement)

C4.5 (Quinlan, 1993), gain ratio

Nous avons utilisés la procédure CART de Answer Tree 3.1 (SPSS, 2001)

2 Validation des arbres

Arbres surtout utilisés pour classification (arbres de classification).

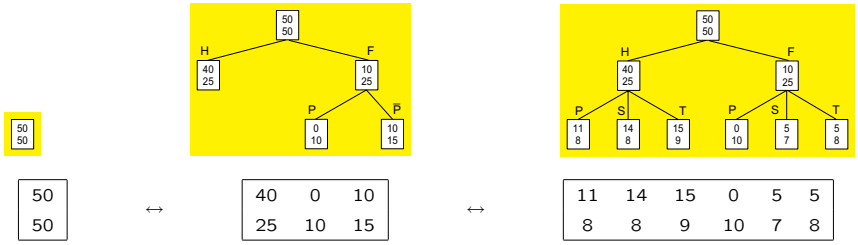
⇒ qualité de l'arbre jugée selon taux d'erreur de classification (en généralisation).

Notre objectif est la description (non la classification)

⇒ Erreur de classification non pertinente pour juger de la qualité descriptive

⇒ besoin de critères mieux adaptés (Ritschard and Zighed, 2004)

2.1 Déviance



Principe de construction de la table prédite

Table prédite $\hat{T}$

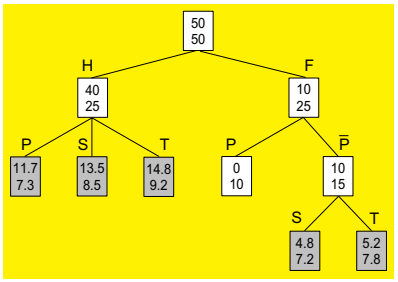
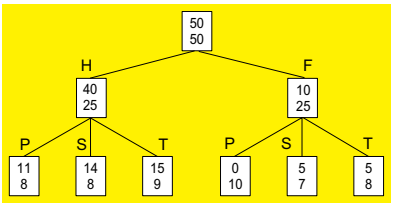


Table cible T



$\hat{T} =$

11.7	13.5	14.8	0	4.8	5.2
7.3	8.5	9.2	10	7.2	7.8

$T =$

11	14	15	0	5	5
8	8	9	10	7	8

2.2 Calcul de la déviance

$T = (n_{ij})$ tableau $\ell \times c$ cible :

ℓ lignes = catégories de la variable à prédire

c colonnes = profils différents en termes des prédicteurs

$\hat{T} = (\hat{n}_{ij})$ tableau $\ell \times c$ prédit par l'arbre

Total de chaque colonne réparti selon distribution de la feuille contenant le profil correspondant.

$$D(m) = -2\sum_{i=1}^{\ell}\sum_{j=1}^{c} n_{ij} \ln\left(\frac{\hat{n}_{ij}}{n_{ij}}\right)$$

Difficulté : construction des tableaux T et $\hat{T}$ car c peut être très grand

Déviance et rapport de vraisemblance

$D(m_0|m)$ = statistique du khi-2 du rapport de vraisemblance
pour test indépendance sur tableau associé à l'arbre induit.

$D(m_0)$ = statistique du khi-2 du rapport de vraisemblance
pour test indépendance sur tableau cible.

Ces deux valeurs s'obtiennent avec les logiciels statistiques (SPSS, SAS, ...)
On obtient la déviance de l'arbre m par différence

$$D(m) \quad = \quad D(m_0) \; - \; D(m_0|m)$$

2.3 Indicateurs dérivés de la déviance

Indicateurs dérivés de la déviance :

- BIC = déviance pénalisée pour la complexité (nbre de paramètres)
  défini à une constante additive près ⇒ seules variations sont pertinentes
- pseudo $R^2 = 1 - D(m)/D(m_0)$,
  pas pertinent avec déviance partielle
- u Theil, taux de réduction de l'entropie de Shannon

$$u = \frac{D(m_0|m)}{-2\sum_i n_{i.} \ln(n_{i./n})}$$

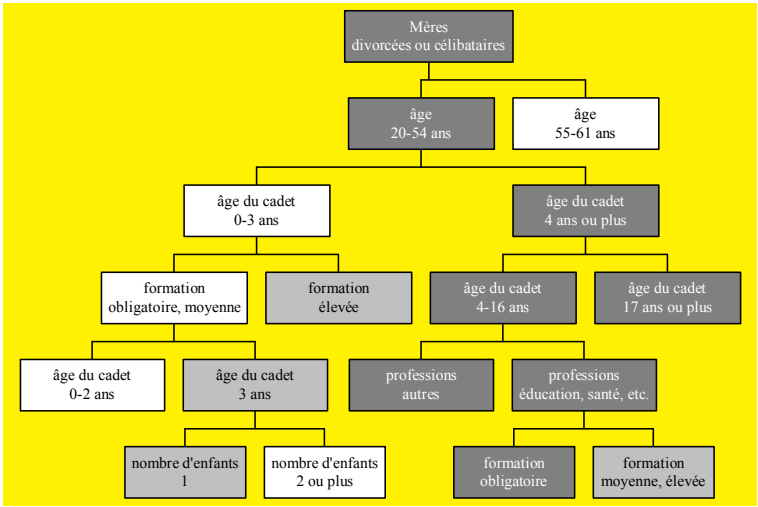
  évolution quadratique ⇒ $\sqrt{u}$ plus pertinent

2.4 Applications aux arbres obtenus

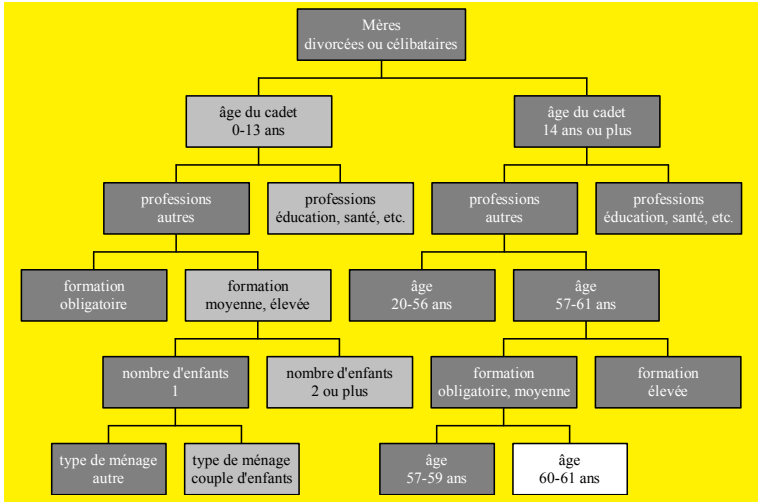
Rappel des arbres obtenus

Qualité des arbres : quelques indicateurs

Statut emploi, mères célibataires ou divorcées, Suisse italienne



Statut emploi, mères célibataires ou divorcées, Suisse romande



Qualité des arbres

Quelques indicateurs

	q	c^*	p	n	$D(m_0\|m)$	d	sig.
CHI	12	263	299	5770	822.2	33	.00
CHF	10	644	674	35239	4293.3	27	.00
CHG	11	684	717	99641	16258.6	30	.00

	$\Delta\text{BIC}(m_0, m)$	$\Delta\text{BIC}(m, m_{T^*})$	u Theil	$\sqrt{u}$
CHI	536.4	3235.7	.056	.237
CHF	4010.7	4160.0	.052	.227
CHG	15913.3	-17504.3	.064	.253

Références

Breiman, L., J. H. Friedman, R. A. Olshen, and C. J. Stone (1984). *Classification And Regression Trees*. New York: Chapman and Hall.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics 29*(2), 119–127.

Losa, F. B. and P. Origoni (2004). Partecipazione e non partecipazione femminile al mercato del lavoro. Modelli socioculturali a confronto. Il caso della svizzera italiana nel contesto nazionale. Aspetti statistici, Ufficio cantonale di statistica, Bellinzona.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo: Morgan Kaufmann.

Ritschard, G. and D. A. Zighed (2004). Qualité d'ajustement d'arbres d'induction. *Revue des nouvelles technologies de l'information E-1*, 45–67.

SPSS (Ed.) (2001). *Answer Tree 3.0 User's Guide*. Chicago: SPSS Inc.