Goodness-of-Fit Measures for Induction Trees

Gilbert Ritschard, Department of Econometrics, University of Geneva
Djamel A. Zighed, ERIC, University of Lyon 2

ISMIS 2003, Maebashi, August 2003

Table of Content

http://mephisto.unige.ch

1 Motivation

Study of Students Enroled at the ESS Faculty in 1998

Response variable:

- Situation in October 1999 (eliminated, repeating 1st year, passed)

Predictors:

- Age
- Registration Date
- Selected Core Curriculum (Business and Economics, Social Sciences)
- Type of Secondary Diploma Obtained
- Place of Obtention of Secondary Diploma
- Age at Obtention of Secondary Diploma
- Nationality
- Mother's Living Place

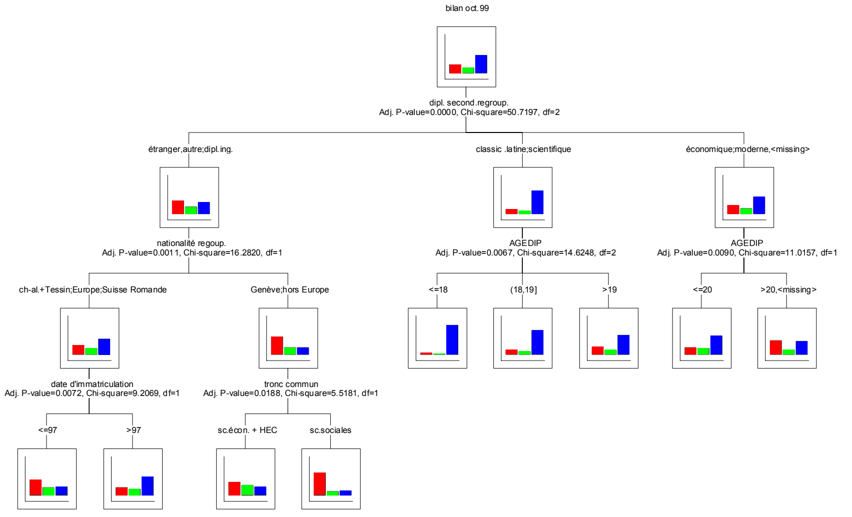
Categorical Data (Multiway Contingency Table)

Sociologists used to

- analyse the structure of association
  ⇒ log-linear models
- study effects on a (categorical) response variable
  ⇒ logistic regression (binary, multinomial)

This kind of data can also be described with trees

or other machine learning methods



2 Induction trees and target table

Induction Trees: supervised learning
(Kass (1980), Breiman et al. (1984), Quinlan (1993), Zighed and Rakotomalala (2000), Hastie et al. (2001))

⇒ 1 categorical response variable y (marital status)

predictors, categorical or quantitative attributes $\mathbf{x} = (x_1, \ldots, x_p)$
(gender, activity sector)

(metric response variable ⇒ regression trees)

2.1 Target Table

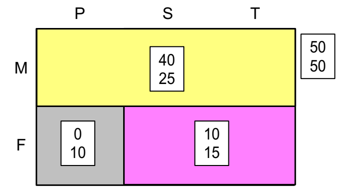
When all variables are categorical, the data can be organized into a contingency table that cross-tabulates the response variable with the composite variable defined by the crossing of all predictors.

Table 1: Example of a target contingency table $\mathbf{T}$

	male			female			
married	primary	secondary	tertiary	primary	secondary	tertiary	total
no	11	14	15	0	5	5	50
yes	8	8	9	10	7	8	50
total	19	22	24	10	12	13	100

An induction tree builds $f(\mathbf{x})$ in two steps:

1. Find a partition of the possible profiles $\mathbf{x}$ such that the distribution p_y of the response Y differs as much as possible from one class to the other.



2. The rule $f(\mathbf{x})$ consists then in giving to each case the value of y that is the most frequent in its class.

$$\hat{y} = f(\mathbf{x}) = \arg\max_i \hat{p}_i(\mathbf{x})$$

2.2 Induction trees: principle

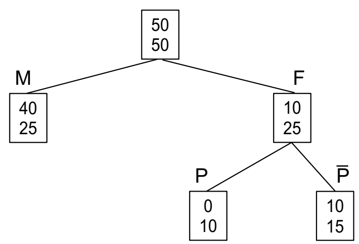


Figure 1: Induced tree

Induction trees determine the partition by successively splitting nodes. Starting with the root node, they seek the attribute that generates the best split according to a given criterion. This operation is then repeated at each new node until some stopping criterion, a minimal node size for instance, is met.

2.3 The criteria

Criteria from

information theory : entropies (uncertainty) of the distribution

Shannon's entropy: $h_S(p) = -\sum_{i=1}^{c} p_i \log_2 p_i$

Quadratic entropy (Gini): $h_Q(p) = \sum_{i=1}^{c} p_i(1-p_i) = 1 - \sum_{i=1}^{c} p_i^2$

⇒ maximize the reduction in entropy (or standardized entropy)

For example, C4.5 maximizes the Gain Ratio $\left(\frac{h_S(p_y) - h_S(p_y|x)}{h_S(p_x)}\right)$

statistical association Pearson Chi-square, measures of association

⇒ maximize the association, minimize the p-value of the no association test.

2.4 Classical validation criteria

The quality of a tree (graph) is evaluated by

- Classification performance (error rates)
- Complexity (number of nodes, number of levels, ...)
- Quality of the partition (entropy, purity, degree of association with response, ...)

Question: Can we transpose the way we evaluate statistical models, log-linear models for instance, to trees? Can we test hypotheses with trees?

independence	fitted model	saturated model
root node	induced tree	saturated tree

R^2 like indicators measure how better we do than the naive model. We can compute percent reduction in error rates or in entropy.

Quid of the quality of reproduction of the target table (distance between predictions and observed table)?

Is there a way to test statistically the effects described by a tree?

3 Fitting the target table

Goodness-of-fit: capacity of the model to reproduce the data.

Two kinds of fit

1. Fit of individual data y_α
2. Fit of the synthetic representation (target table $\mathbf{T}$)

In supervised learning, the objective is generally classification.
⇒ fitting individual data ⇒ quality of the rule $f(\mathbf{x})$.

In social sciences, we are primarily interested in the mechanisms, i.e. in how the predictors influence the response variable.
⇒ examine the effects of $\mathbf{x}$ on the distribution of Y
⇒ fitting the contingency table ⇒ quality of the descriptive model $\mathbf{p}(\mathbf{x})$.

3.1 Table generated by the induced tree

$\mathbf{T}^a$ table crossing the response variable with the partition defined by the tree.

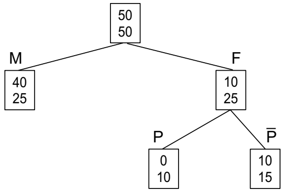


Table 2: Contingency table $\hat{\mathbf{T}}^a$ generated by the tree

married	male	female		total
		primary sector	other sector	
no	40	0	10	50
yes	25	10	15	50
total	65	10	25	100

Saturated tree and target table

Saturated tree: tree that generates exactly the target table $\mathbf{T}$

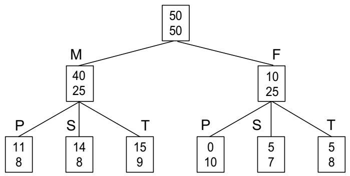


Table 3: Target contingency table $\mathbf{T}$

married	male			female			total
	primary	secondary	tertiary	primary	secondary	tertiary	
no	11	14	15	0	5	5	50
yes	8	8	9	10	7	8	50
total	19	22	24	10	12	13	100

Extended tree and predicted table

Induced tree (white nodes) and its maximal extension

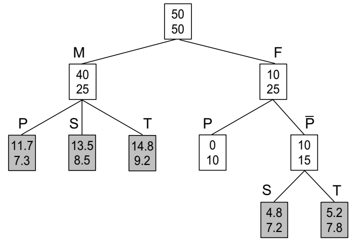


Table 4: Predicted contingency table $\hat{\mathbf{T}}$

married	male			female			total
	primary	secondary	tertiary	primary	secondary	tertiary	
no	11.7	13.5	14.8	0	4.8	5.2	50
yes	7.3	8.5	9.2	10	7.2	7.8	50
total	19	22	24	10	12	13	100

4 Measuring and testing the fit

4.1 The Deviance Chi-square statistic

Fit: distance between $\hat{\mathbf{T}}$ and $\mathbf{T}$

Chi-square divergence measures: for example Likelihood Ratio G^2 statistics (deviance)

$$G^2 = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \ln\left(\frac{n_{ij}}{\hat{n}_{ij}}\right) \qquad (1)$$

When the model is correct, and under some regularity conditions, G^2 has a χ^2 distribution.

What are the degrees of freedom ?

Table rebuilding model and degrees of freedom

We express the table predicted from an induced tree in terms of a parameterized rebuilding model. Letting $\mathbf{T}_j$ stand for the jth column of $\mathbf{T}$, the model is:

$$\hat{\mathbf{T}}_j = n\, a_j \hat{\mathbf{p}}_{|j}, \quad j = 1, \ldots, c \tag{2}$$

$$\text{s.t.} \quad \hat{\mathbf{p}}_{|j} = \mathbf{p}^a_{|k} \quad \text{for all } \mathbf{x}_j \in \mathcal{X}_k \quad k = 1, \ldots, q \tag{3}$$

$\mathcal{X}_k$ is the class of profiles $\mathbf{x}$ defined by the kth leaf of the tree.

The parameters are

- n the total number of cases (learning sample size),
- a_j the proportion of cases in each column $j = 1, \ldots, c$, and
- $\mathbf{p}_{|j}$, the c probability vectors $\mathbf{p}(Y|j)$ of size r that characterize the distribution of Y in each column j of the table.

Degrees of freedom

Number of independent constraints (3)

$$\boxed{d_M = (c - q)(r - 1)}$$

For the independence model: $q = 1$ and hence $d_I = (c-1)(r-1)$.
For the saturated tree: $q = c$ and hence $d_S = 0$.

4.2 Other fit indicators based on the deviance

LR test for comparing two nested trees. If restricted tree M_2 is correct
Agresti (1990)

$$G^2(M_2|M_1) = G^2(M_2) - G^2(M_1) \quad \sim \quad \chi^2_{d_{M_2} - d_{M_1}}$$

Pseudo R^2 : Fit improvement over independence (root node)

$$R^2 = 1 - \frac{G^2(M)}{G^2(I)} \qquad R^2_{\text{adj}} = 1 - \frac{G^2(M)/d_M}{G^2(I)/d_I}$$

Information criteria Deviance penalized for complexity (# free parameters)
Akaike (1973), Schwarz (1978), Raftery (1995), Kass and Raftery (1995)

$$\text{AIC}(M) = G^2(M) + 2(qr - q + c)$$
$$\text{BIC}(M) = G^2(M) + (qr - q + c)\log(n)$$

5 Illustration: ESS98 first year students

Attributes and value grouping selected by CHAID ⇒ 88 target columns

Table 5: ESS 98: Goodness-of-fit of a selection of models

Model	q	d	G^2	sig(G^2)	pseudo R^2_{adj}	AIC	BIC
Saturated	88	0	0	1	1	528	1751.9
Best AIC	14	148	17.4	1	.941	249.4	787.2
CHAID	9	158	177.9	0.133	.336	390.0	881.3
CHAID2	8	160	187.4	0.068	.309	395.4	877.5
CHAID3	7	162	195.2	0.038	.289	399.2	872.1
Best BIC	6	164	75.2	1	.745	275.2	738.8
Independence	1	174	295.1	0.000	0	475.8	892.3

CHAID2 : CHAID without split *datimma* at node 4 (*nationa*≠ GE, non Europe)

CHAID3 : CHAID2 without split *troncom* at node 5 (*nationa*= GE, non Europe)

6 Conclusion and further developments

- "Trees" well suited method for describing a contingency table that cross-tabulates a response variable with a set of predictors.
- Classical statistical tools can be used for assessing the relevance of the tree (indeed of the table predicted by the tree.)
- Effects of predictors can be tested individually or simultaneously.
- Effects can be tested locally at some node or globally.

Further developments

- Continuous predictors (how can we take account of the endogenous discretization?)
- Use goodness-of-fit criteria at the tree growing stage (e.g. algorithm seeking the BIC-optimal tree.)

References

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrox and F. Caski (Eds.), *Second International Symposium on Information Theory*, pp. 267. Budapest: Akademiai Kiado.

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

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. New York: Springer.

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

Kass, R. E. and A. E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association 90*(430), 773–795.

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

Raftery, A. E. (1995). Bayesian model selection in social research. In P. Marsden (Ed.), *Sociological Methodology*, pp. 111–163. Washington, DC: The American Sociological Association.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*, 461–464.

Zighed, D. A. and R. Rakotomalala (2000). *Graphes d'induction: apprentissage et data mining*. Paris: Hermes Science Publications.