Testing Hypotheses with Induction Trees

Gilbert Ritschard
Department of Econometrics, University of Geneva
40, boulevard Carl-Vogt
CH-1211 Geneva 4, Switzerland
gilbert.ritschard@themes.unige.ch

1. Introduction

Data mining has brought a new philosophy in data analysis that is primarily driven by computational efficiency and predictive performance. In this paper we attempt to show that while this new philosophy opens new horizons, these new techniques may themselves gain much when coupled with traditional statistical reasoning.

For example, I recently introduced some sociologists of the University of Geneva to induction trees. These social scientists were impressed by the ease with which such tools allowed them to extract valuable knowledge from their datasets. However, since they were used to fit statistical models like linear or logistic regressions or even multinomial log-linear models, they naturally wanted to know how well the induced trees fit their data. They also wanted to test the significance of specific branch expansions and compare them with alternatives that they found more meaningful. The classification error rates let them unsatisfied. Being primarily interested in how the predictors jointly affect the distribution of the response variable rather than in classification, they expected indeed some divergence Chi-square statistics and inferential tools for comparing alternative structures. Unfortunately they did not found such information in the software outcomes.

It is indeed characteristic of data mining and especially of machine learning to focus on the usefulness and predictive performance of the induced rules and to neglect somehow their descriptive content. Thus, the rules are most often used as black boxes. They provide, however, as our sociologists discovered it, also very useful descriptive knowledge about the phenomenon under study. It makes then sense to statistically validate the description provided. Only few attention has been given so far to this aspect. Textbooks, like Han and Kamber (2001) for example, don't mention it, and, as far as prediction rules are concerned, statistical learning (see Hastie et al., 2001, chap. 7) concentrates on the statistical properties of the classification error rate.

This lack of inferential tools for the descriptive content of classification rules motivated this paper. Focusing on induction trees with categorical variables, we propose a simple trick that permits to apply to them the inferential tools used for instance in the statistical log-modeling of multinomial cross tables.

The paper is organized as follows. Section 2 discusses the fit issue and introduces the trick that renders induced trees conformable with the requirements of Chi-square statistics. Section 4 is devoted to tree comparison and shows how tests of hypotheses about the tree structure can be carried out with the deviance statistic. Section 5 provides concluding remarks.

2. Goodness of Fit of Induced Trees

The goodness of fit of a statistical model refers to its capacity to reproduce the data. In supervised learning, hence with induction trees, the goal is usually to get an individual prediction or classification $\hat{y}_\alpha$ for each case. Then, the $\hat{y}_\alpha$ have to fit the observed values y_α of the response variable for $\alpha = 1, \cdots, n$, n being the number of observed cases. The quality of the fit is in these settings measured by the prediction or classification error rate or some function of it.

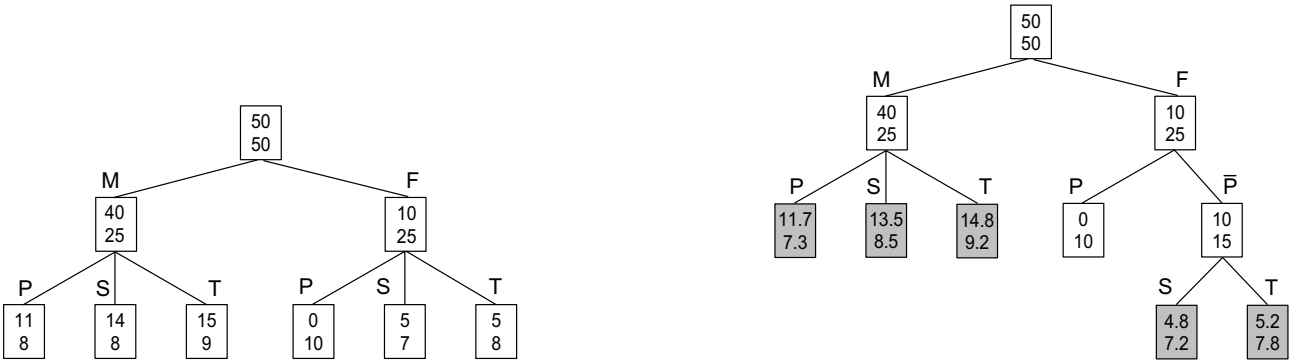

Figure 1. *(left) saturated tree and (right) an induced tree (white nodes) together with its maximal extension (white + grey nodes). The predictors are the gender (M, F) and the sector of activity (P=primary, S=secondary, T=tertiary) and the response variable is the marital status (yes, no). Observe that the distribution in the grayed nodes is the same as in their parental white node.*

Sometimes nevertheless, and particularly in the social sciences, supervised learning tools like induction trees are used to explore the relationships between potential predictors x_s, $s = 1, \ldots, p$ and the response variable Y rather than for classification. The focus is then on how the distribution of Y varies with changes in the predictor profile $\mathbf{x} = (x_1, \ldots, x_p)$ and the target to fit becomes the conditional distributions $\mathbf{p}(\mathbf{x}_j) = \big(p(Y = y_1|\mathbf{x}_j), \ldots, p(Y = y_r|\mathbf{x}_j)\big)$ for $j = 1, \ldots, c$, and we do no longer care about the individual values y_α. Note that if each predictor x_s, $s = 1, \ldots, p$ has c_s different values, the number of possible different profiles, and hence conditional distributions, is $c = \prod_{s=1}^{p} c_s$.

When all variables are discrete, the empirical counterpart of the conditional distributions $\mathbf{p}(\mathbf{x})$ can be derived from the $r \times c$ contingency table $\mathbf{T}$ that cross classifies the r values of Y with the c profiles. Letting n_{ij} denote an element of table $\mathbf{T}$ and $n_{\cdot j}$ the column j total, the maximum likelihood estimation of $\mathbf{p}_{|j} = \mathbf{p}(\mathbf{x}_j)$ is indeed the vector of the observed frequencies $n_{ij}/n_{\cdot j}, i = 1, \ldots, r$. Each column of the table $\mathbf{T}$ corresponds to the terminal node of a so called *saturated tree*, i.e. the tree that exhausts all splits to generate the finest partition for the retained predictors (see Figure 1, left.)

As will be shown, a induced tree provides a prediction $\hat{\mathbf{T}}$ of $\mathbf{T}$. Measuring the (descriptive) goodness of fit of the tree consists then in measuring how $\hat{\mathbf{T}}$ fits $\mathbf{T}$ with for example Pearson or log-likelihood ratio Chi-squares. To explain how we get $\hat{\mathbf{T}}$ from an induced tree, we consider the following rebuilding model where $\hat{\mathbf{T}}_j$ stands for the j-th column of $\hat{\mathbf{T}}$

$$\text{(1)} \qquad \hat{\mathbf{T}}_j \;=\; n\, a_j \hat{\mathbf{p}}_{|j}, \quad j = 1, \ldots, c$$

The parameters are the total number of cases n, the proportions a_j of cases in column (terminal node) $j = 1, \ldots, c$ and the c column distribution vectors $\mathbf{p}_{|j}$. The a_j's are naturally estimated by $n_{\cdot j}/n$. The only trick required concerns the estimation of the $\mathbf{p}_{|j}$'s. Indeed, the induced tree has generally $q < c$ terminal nodes which generate a $r \times q$ table $\mathbf{T}^a$ not conformable with $\mathbf{T}$. To render table $\mathbf{T}^a$ conformable, we have to extend it or equivalently the induced tree.

Definition 1 *The* maximal extension of an induced tree *is obtained by maximally further splitting each terminal node $k = 1, \ldots, q$ of the tree and by distributing the cases in each new node according to the distribution $\mathbf{p}^a_{|k}$ of its parent terminal node of the induced tree. (See Figure 1, right.)*

Formally, letting $\mathcal{X}_k$ denote the subset of profiles that belong to the group defined by the terminal node k, the maximally extended tree leads to the following estimations of $\mathbf{p}_{|j}$

$$\text{(2)} \qquad \hat{\mathbf{p}}_{|j} = \mathbf{p}^a_{|k} \quad \text{for all } \mathbf{x}_j \in \mathcal{X}_k \qquad k = 1, \ldots, q$$

3. Goodness-of-Fit Measures for Induction Trees

Having defined the target table $\mathbf{T}$ and the one $\hat{\mathbf{T}}$ predicted by the induced tree, we can now apply the machinery of statistical tests and goodness indicators used in the statistical modeling of cross tables.

The most popular divergence Chi-square statistics are the Pearson X^2 and the deviance G^2 statistics. Under some regularity conditions (see for instance Bishop et al., 1975, chap. 14) these statistics have, when the induced tree is correct, an asymptotical Chi-square distribution. In our case, the deviance G^2 is

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

and has $d = (r-1)(c-q)$ degrees of freedom, which corresponds to the number of independent constraints (2).

4. Testing Hypotheses about the tree structure

It is well known that the scope of Chi-square statistics is limited when n becomes very large, the smallest departure from the target becoming statistically significant. Further, the regularity conditions, especially interiority that requires non zero expected frequencies, may not hold when the number of variables becomes large. Nevertheless, the G^2 statistic proves useful for testing hypotheses about the tree structure.

Thanks to an additive property, G^2 permits to test the difference between nested models. Let M_2 be a restricted form of model M_1. Then, the deviance between the two models is (see Agresti, 1990, p. 211)

$$G^2(M_2|M_1) = G^2(M_2) - G^2(M_1)$$

which, if M_2 is correct, has an asymptotic Chi-square distribution with $d_2 - d_1$ degrees of freedom.

For induction trees, the deviance between nested trees, i.e. between a tree and the same tree after the subtree of interest has been removed, provides a natural way to test the statistical relevance of the subtree. This way of testing a whole part of the tree clearly complements the information provided by the criteria locally optimized at each split. In the example of Figure 1, we could for instance test if the activity sector has a significant role, by comparing the induced tree with the tree that includes only the split by gender.

Information criteria, like the AIC from Akaike (1973) and the Bayesian BIC (Schwarz, 1978; Kass and Raftery, 1995) are useful for trading off between fit and complexity and hence for model selection. For our settings, they read

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

They penalize the log-likelihood fit statistics G^2 for the degree of complexity measured by the number $(qr-q+c) = rc-d$ of independent parameters. The tree with the smaller value of the criteria offers the better compromise between fit and complexity. Note that unlike G^2 these criteria are suitable for comparing non nested trees.

5. Conclusion

We have shown how testing the statistical significance of an induced tree can be carried out and have emphasized the kind of additional knowledge it may furnish. Though our discussion was limited to induction trees with discrete attributes, it illustrates how traditional statistical reasoning may improve the scope of data mining tools.

We are convinced that the inferential tools presented here could be extended to other settings, e.g. for trees with quantitative attributes, but also to other supervised learning tools like the mining of frequent or interesting itemsets and association rules. For the case of trees with quantitative attributes, the difficulty lies in the discretization thresholds that are usually dynamically and optimally determined during the tree growing process. The thresholds are then parameters of the model of which we should also take account. For the mining of association rules, the central question will probably be that of the representation of an enumerated set of association rules by a parameterized descriptive model.

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.

Bishop, Y. M. M., S. E. Fienberg, and P. W. Holland (1975). *Discrete Multivariate Analysis*. Cambridge MA: MIT Press.

Han, J. and M. Kamber (2001). *Data Mining: Concept and Techniques*. San Francisco: Morgan Kaufmann.

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

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

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

RÉSUMÉ

Les arbres d'induction sont largement utilisés en data mining tant dans un but exploratoire que comme outil de classification supervisée. Rares sont cependant les outils inférentiels disponibles pour valider statistiquement la description fournie par un arbre induit. Nous proposons ici une astuce qui permet d'appliquer aux arbres les outils inférentiels utilisés par exemple en modélisation log-linéaire de tables de contingence multidimensionnelles. Parmi ces outils, la déviance se prête en particulier à divers test sur la structure de l'arbre induit. De la statistique de vraisemblance, on déduit également les critères d'information AIC et BIC qui permettent d'arbitrer entre la complexité et la qualité d'ajustement d'arbres non emboîtés. Ces outils inférentiels pour arbres d'induction sont particulièrement appréciés par les chercheurs de sciences sociales qui s'intéressent en priorité à comprendre comment les facteurs explicatifs interagissent sur la variable réponse, et ne sont donc guère concernés par le taux d'erreur de classification habituellement utilisé pour juger de la qualité des arbres. La possibilité de tester des hypothèses sur la structure de l'arbre contribue à la connaissance du phénomène étudié. Elle constitue une illustration de ce que le raisonnement statistique traditionnel peut apporter au data mining.