Computing and using the deviance with classification trees

Gilbert Ritschard

Department of Econometrics, University of Geneva, Switzerland
gilbert.ritschard@themes.unige.ch

Summary. The reliability of induced classification trees is most often evaluated by means of the error rate. Whether computed on test data or through cross-validation, this error rate is suited for classification purposes. We claim that it is, however, a partial indicator only of the quality of the knowledge provided by trees and that there is a need for additional indicators. For example, the error rate is not representative of the quality of the description provided. In this paper we focus on this descriptive aspect. We consider the deviance as a goodness-of-fit statistic that attempts to measure how well the tree is at reproducing the conditional distribution of the response variable for each possible profile (rather than the individual response value for each case) and we discuss various statistical tests that can be derived from them. Special attention is devoted to computational aspects.

Key words: Classification tree, Deviance, Goodness-of-fit, Chi-square statistics, BIC.

1 Introduction

Induced decision trees have become, since [BFOS84] and [Qui86], popular multivariate tools for predicting continuous dependent variables and for classifying categorical ones from a set of predictors. They are called *regression trees* when the outcome is quantitative and *classification trees* when it is categorical. Though their primary aim is predicting and classifying, trees can be used for many other relevant purposes: as exploratory methods for partitioning and identifying local structures in data sets, as well as alternatives to statistical descriptive methods like linear or logistic regression, discriminant analysis, and other mathematical modeling approaches [Mur98]. As descriptive tools, their attractiveness lies mainly in the ease with which end users can visualize and interpret a tree structure. This is much more immediate than interpreting for instance the values of the coefficients of a logistic regression. A further aspect that is often put forth is that tree induction is non-parametric

in the sense that it needs no a priori assumption on the form of the data distribution.

As for any statistical model, it is of primary importance to evaluate the reliability of an induced tree. For classification trees, the most often used criterion is the classification error rate. An important concern here is over fitting. This occurs essentially when the optimizing criteria used for tree growing rely to entropy measures that are not statistical in the sense that they are unsensitive to the number of cases. The consequence is then that the induced tree may be too closely tied to the learning sample to have any generalization capacity. To prevent this, the growing step is usually followed by a pruning round which attempts to simplify the tree by resorting for example in CART [BFOS84] to an error rate penalized for the number of leaves (terminal nodes). The tree validation is then done either by computing the error rate on a set of validation data (different from the learning set) or through cross-validation.

The problem with the error rate is that while it is well suited for classification purposes, it is of poor help for validating the descriptive capacity of the tree. Consider for example a split into two groups with say distribution $(.1, .9)$ and $(.45, .55)$ for the outcome variable. Clearly this is valuable knowledge while the gain in terms of error rate over the root node will be null, the most frequent value remaining the same for both groups.

We reconsider in this paper the deviance, which we showed how it can be applied to trees in [RZ03]. The deviance usefully complements the error rate and permits to make some statistical inference with trees. Firstly, we give a new presentation of how the deviance that is abundantly used in statistical modeling can be adapted for induction trees. We recall how it can be used for testing the fit and fit variations. We shortly enumerate also indicators like pseudo R^2's and Akaike (AIC) and Bayesian (BIC) information criteria that are derived from the deviance. Then we focus on computational aspects and provide for instance an SPSS syntax for computing the deviance.

2 Tree induction principle: an illustrative example

We recall in this section the terminology and concepts related to tree induction. We start by introducing an illustrative example data set that will serve all along the paper.

2.1 Illustrative example

We consider a fictional example where we are interested in predicting the civil status (married, single, divorced/widowed) of individuals from their gender (male, female) and sector of activity (primary, secondary, tertiary). The civil status is the outcome or response variable, while gender and activity sector are the predictors. The data set is composed of the 273 cases described by table 1.

Table 1. Example: The data set

Civil status	Gender	Activity sector	Number of cases
married	male	primary	50
married	male	secondary	40
married	male	tertiary	6
married	female	primary	0
married	female	secondary	14
married	female	tertiary	10
single	male	primary	5
single	male	secondary	5
single	male	tertiary	12
single	female	primary	50
single	female	secondary	30
single	female	tertiary	18
divorced/widowed	male	primary	5
divorced/widowed	male	secondary	8
divorced/widowed	male	tertiary	10
divorced/widowed	female	primary	6
divorced/widowed	female	secondary	2
divorced/widowed	female	tertiary	2

2.2 Principle, terminology and notations

Classification trees are grown by seeking, through recursive splits of the learning data set, some optimal partition of the predictor space for predicting the outcome class. Each split is done according to the values of one predictor. The process is greedy. At the first step, it tries all predictors to find the "best" split. Then, the process is repeated at each new node until some stopping rule is reached. This requires a local criterion to determine the "best" split at each node. The choice of the criterion is the main difference between the various tree growing methods that have been proposed in the literature, of which CHAID [Kas80], CART [BFOS84] and C4.5 [Qui93] are perhaps the most popular.

A *leaf* is a terminal node. There are 4 leaves in Figure 1.

In the machine learning community, predictors are also called attributes and the outcome variable the predicted attribute. The values of the outcome variable are called the classes. We prefer using "outcome values" to avoid confusion with the classes of the population partition defined by the leaves.

We call *profile* a vector of predictor values. For instance, (female, tertiary) is a profile in Table 1.

We call *target table* and denote by T the contingency table that cross classifies the outcome values with the set of possible profiles. As shown in Table 2, there are 6 possible profiles for our data.

Notice that the root node contains just the marginal distribution of the outcome variable. It is useful also to point out that the columns of the target

table are just the leaves of a maximally developed tree (see the right side of Figure 2). We call *saturated tree* this maximally developed tree.

The count in cell (i, j) of the target table T is denoted n_{ij}. We designate by $n_{.j}$ and $n_{i.}$ the total of respectively the jth column and ith row.

3 Validating the tree descriptive ability

For the reliability of the description, individual predictions do not matter. Rather, we focus on the posterior distribution of the response variable, i.e. on the distribution conditioned by the values of the predictors. These posterior distributions are the columns of the target table. Our concern is thus to measure how well a tree may predict this target table. This is a goodness-of-fit issue very similar to that encountered in the statistical modeling of multiway cross tables. According to our knowledge, however, it has not been addressed so far for induced trees. Textbooks, like [HK01] or [HMS01] for example, do not mention it, and, as far as this model assessment issue is concerned, statistical learning focuses almost exclusively on the statistical properties of the classification error rate (see for example [HTF01] chap. 7).

In statistical modeling, e.g. linear regression, logistic regression or more generally generalized linear models (GLM), the goodness-of-fit is usually assessed by two kinds of measures. On the one hand, indicators like the coefficient of determination R^2 or pseudo R^2's tell us how better the model does than some naive baseline model. On the other hand we measure, usually with

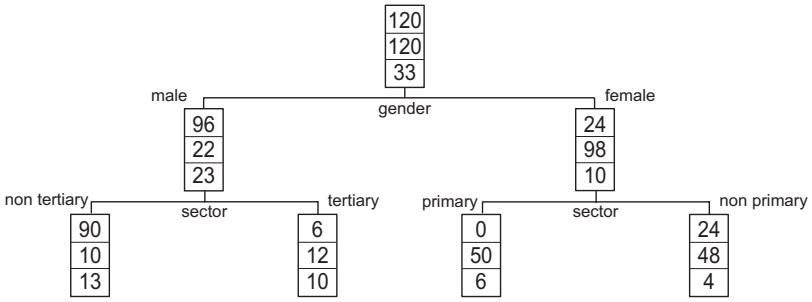


Fig. 1. Example: Induced tree for civil status (married, single, divorced/widowed)

Table 2. Target table

	male			female			
	primary	secondary	tertiary	primary	secondary	tertiary	total
married	50	40	6	0	14	10	120
single	5	5	12	50	30	18	120
div./wid.	5	8	10	6	2	2	33
total	60	53	28	56	46	30	273

divergence Chi-square statistics, how well the model reproduces some target or, in other words, how far we are from the target.

Our contribution is a trick that permits to use this statistical machinery with induced trees. The trick allows us to propose, among others, an adapted form of the Likelihood Ratio deviance statistic with which we can test statistically the significance of any expansion of a tree. Other criteria discussed are R^2 like measures and the powerful model selection AIC and BIC criteria.

3.1 The deviance

Having defined the target table, we propose using the deviance for measuring how far the induced tree is from this target (Figure 2). By comparing with the deviance between the root node and the target, we should also be able to evaluate the overall contribution of the predictors, i.e. what is gained over not using any predictor.

The general idea of the deviance of a statistical model m is to measure how far the model is from the target, or more specifically how far the values predicted by the model are from the target. In general (see for instance [MN89]), this is measured by minus twice the log-likelihood of the model ($-2\text{LogLik}(m)$) and is just the log-likelihood ratio Chi-square in the modeling of multiway contingency tables [Agr90]. For a 2 way $r \times c$ table, it reads for instance

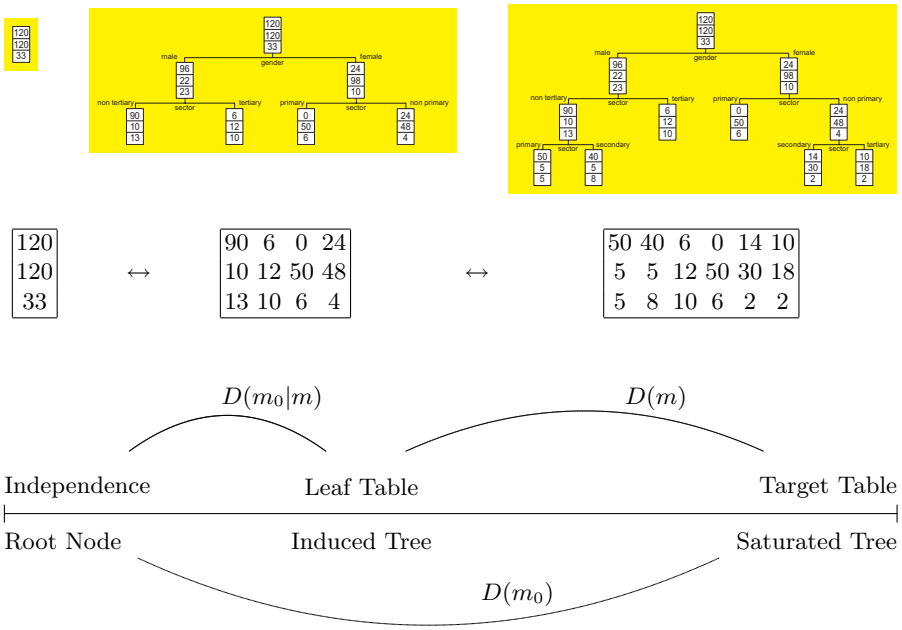
$$D(m) = 2\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} \ln\left(\frac{n_{ij}}{\hat{n}_{ij}}\right) \ , \tag{1}$$

where $\hat{n}_{ij}$ is the estimation of the expected count provided by the model for cell (i, j). The likelihood is obtained assuming simply a multinomial distribution which is by noway restrictive. Under some regularity conditions (see for instance [BFH75] chap. 4), the Log-Likelihood Ratio statistic has an approximate Chi-square distribution when the model is correct. The degrees of freedom d are given by the difference between the number of cells and the number of free parameters of the model.

The advantage of the deviance over for instance the Pearson Chi-square is an additivity property that permits to test the difference between a model m_1 and a restricted version m_2 with the difference $D(m_2|m_1) = D(m_2) - D(m_1)$. This difference has indeed also an approximate Chi-square distribution when the restricted model is correct. Its number of degrees of freedom equals the difference $d_2 - d_1$ in degrees of freedom for each model.

3.2 Deviance for a tree

We have already defined the target table for a classification tree with discrete attributes. Hence, we should be able to compute a deviance for the tree. We face two problems however:



$$\begin{bmatrix} 120 \\ 120 \\ 33 \end{bmatrix} \leftrightarrow \begin{bmatrix} 90 & 6 & 0 & 24 \\ 10 & 12 & 50 & 48 \\ 13 & 10 & 6 & 4 \end{bmatrix} \leftrightarrow \begin{bmatrix} 50 & 40 & 6 & 0 & 14 & 10 \\ 5 & 5 & 12 & 50 & 30 & 18 \\ 5 & 8 & 10 & 6 & 2 & 2 \end{bmatrix}$$

Fig. 2. Deviance

1. How do we compute the predicted counts $\hat{n}_{ij}$ from the induced tree ?
2. What are the degrees of freedom ?

To answer these questions we postulate a (non restrictive) multinomial distribution of the outcome variable for each profile. More specifically, we assume a discrete distribution

$$\mathbf{p}_j = (p_{1|j}, \ldots, p_{r|j}) \ ,$$

where $p_{i|j}$ is the probability to be in state i of the outcome variable for a case with profile $\mathbf{x}_j$.

A tree with $q \leq c$ leaves can be seen as a model of the target table. It states that the probability $p_{i|j}$ of being in the ith value of the outcome variable is equal for all profiles j belonging to a same leaf k, i.e.

$$p_{i|j} = p^*_{i|k}, \quad \text{for all } \mathbf{x}_j \in \mathcal{X}_k, k = 1, \ldots, q \ ,$$

where $\mathcal{X}_k$ stands for the set of profiles of leaf k. The tree parameterizes the rc probabilities $p_{i|j}$ in terms of rq parameters $p^*_{i|k}$, which leaves

$$d = (r-1)(c-q) \text{ degrees of freedom } . \tag{2}$$

The probabilities $p^*_{i|k}$'s are estimated by the observed proportions, i.e $\hat{p}^*_{i|k} = n_{ij}/n_{.j}$. Estimates of the probabilities $p_{i|j}$ are derived from those of the $p^*_{i|k}$'s, i.e. $\hat{p}_{i|j} = \hat{p}^*_{i|k}$ when $\mathbf{x}_j \in \mathcal{X}_k$.

Table 3. Predicted counts

	male			female			
	primary	secondary	tertiary	primary	secondary	tertiary	total
married	47.8	42.2	6	0	14.5	9.5	120
single	5.3	4.7	12	50	29.1	18.9	120
div./wid.	6.9	6.1	10	6	2.4	1.6	33
total	60	53	28	56	46	30	273

For given n_j's and given distributions $\mathbf{p}_j$, the expected counts for a profile $\mathbf{x}_j$ is $n_{.j}p_{i|j}$, for $i = 1, \ldots, r$. Now, replacing the $p_{i|j}$'s by their estimates, we get estimates $\hat{n}_{ij}$ of the expected counts:

$$\hat{n}_{ij} = n_{.j}\hat{p}^*_{i|k} \quad \text{for all } \mathbf{x}_j \in \mathcal{X}_k, k = 1, \ldots, q \ . \tag{3}$$

Table 3 shows the counts predicted this way from the tree in Figure 1.

Considering the counts of the target table and the estimates (3), the deviance $D(m)$ of a tree m can be computed using formula (1). For our example we find $D(m) = 1.69$. The number of degrees of freedom is $d(m) = (3-1)(6-4) = 4$. The obtained deviance being much less than $d(m)$, it is clearly not statistically significant indicating that the induced tree fits well the target T.

3.3 Using the deviance

The approximated Chi-square distribution of the deviance holds when the expected counts per cell are all say greater than 5. This is rarely the case when the number of predictors is large. Hence, the deviance will not be so useful for testing the goodness-of-fit. Note that we have exactly the same problem with, for instance, logistic regression.

Nevertheless, the difference in the deviance for two nested trees will have a Chi-square distribution, even when the deviances themselves do not.

$$D(m_2|m_1) = D(m_2) - D(m_1) \sim \chi^2 \text{ with } d_2 - d_1 \text{ degrees of freedom} \quad .$$

Thus, the main interest of the deviance is to test differences between nested trees. A special case is testing the difference with the root node with $D(m_0|m)$, which is the equivalent of the usual Likelihood Ratio Chi-square statistic used in logistic regression.

For our example, we have $D(m_0|m) = 167.77$ for 6 degrees of freedom. This is clearly significant and demonstrates that the tree describes the outcome significantly better than independence (root node). The predictors bring significant information.

As a further illustration, let us test if pruning the branches below "female" in the tree of Figure 1 implies a significant change. The reduced tree m_1 has a deviance $D(m_1) = 32.4$ for 6 degrees of freedom. This is statistically

significant, indicating that the reduced tree does not fit the target correctly. The difference with the induced tree m is $D(m_1|m) = 32.4 - 1.7 = 30.7$ for 2 degrees of freedom. This is also significant and demonstrates that pruning the branch deteriorates significantly the deviance.

3.4 Further deviance based indicators

It is very convenient to measure the gain in information in relative terms. Pseudo R^2's, for instance, represent the proportion of reduction in the root node deviance that can be achieved with the tree. Such pseudo R^2's come in different flavors. [McF74] proposed simply $\big(D(m_0) - D(m)\big)/D(m_0)$. A better choice is the improvement of [CS89]'s proposition suggested by [Nag91]:

$$R^2_{\text{Nagelkerke}} = \frac{1 - \exp\{\frac{2}{n}\big(D(m_0) - D(m)\big)\}}{1 - \exp\{\frac{2}{n}D(m_0)\}} \ .$$

The McFadden pseudo R^2 is 0.99, and with Nagelkerke formula we get 0.98.

We may also consider the percent reduction in uncertainty of the outcome distribution for the tree as compared with the root node. The uncertainty coefficient u of [The70], which reads $u = D(m_0|m)/(-2\sum_i n_{i.} \ln(n_{i.}/n))$ in terms of the deviance, and the association measure τ of [GK54] are two such measures. The first is the proportion of reduction in Shannon's entropy and the second in quadratic entropy. These two indexes produce always very close values. They evolve almost in a quadratic way from no association to perfect association [OR95]. Their square root is therefore more representative of the position between these two extreme situations. For our induced tree, we have $\sqrt{u} = 0.56$, and $\sqrt{\tau} = 0.60$, indicating that we are a bit more than half way to full association. For the reduced tree m_1 (pruning branch below female), these values are smaller $\sqrt{u} = 0.51$, and $\sqrt{\tau} = 0.57$ indicating that the pruned branch bears some useful information about the distribution.

From the deviance, we can derive AIC and BIC information criteria. For instance, the BIC value for a tree m is

$$\text{BIC}(m) = D(m) - d\ \ln(n) + \text{constant} \ ,$$

where n is the number of cases and d the degrees of freedom in the tree m. The constant is arbitrary, which means that only differences in BIC values matter. Recall, that according to Raftery [Raf95], a difference in BIC values greater than 10 provides strong evidence for the superiority of the model with the smaller BIC, in terms of trade-off between fit and complexity.

4 Computational aspects

Though the deviance could easily be obtained on our simple example, its practical use on real life data raises two major issues.

1. Existing softwares for growing trees do not provide the deviance, nor do they provide in an easily usable form the data needed to compute the target table and the estimates $\hat{p}_{i|j}$.
2. The number of possible profiles, hence the number c of columns of the target table becomes rapidly excessively large when the number of predictors increases. Theoretically, denoting by c_v the number of values of variable $x_v, v = 1, \ldots, V$, the number of profiles may be as large as $\prod_v c_v$, which may become untractable.

Regarding the *first point*, we need to compute the "profile" variable, i.e. assign to each case a profile value. The profile variable can be seen as a composite variable x_{prof} with a unique value for each cell of the cross classification of all predictors x_v. Assuming that each variable has less than 10 values, we can compute it, for example, by using successive powers of 10

$$x_{prof} = \prod_{v=1}^{V} 10^{v-1} x_v \ .$$

We need also a "leaf" variable x_{leaf} that indicates to which leaf each case belongs. Here we have to rely on tree growing softwares that either directly produce this variable or, like AnswerTree [SPS01] for instance, generate rules for assigning the leaf number to each case.

The next step is to compute the counts of the target table and those of the leaf table resulting from the cross tabulation of the outcome variable with the leaf variable. This can be done by resorting to softwares that directly produce cross tables. However, since the number of columns of at least the target table may be quite large and the tables very scarce, a more careful coding that would take advantage of the scarcity is a real concern. A solution is to aggregate cases by profiles and outcome values, which is for instance easily done with SPSS. Creating a similar file by aggregating by leaves and outcome values, the resulting files can then be merged together so as to assign the leaf data to each profile. From here, it is straightforward to get the estimated counts with formula (3) and then compute the deviance $D(m)$ with formula (1). Figure 3 shows the SPSS syntax we used for getting the deviance of our example induced tree.

An alternative solution that can be used by those who do not want to write code, is to use the Likelihood Ratio Chi-square statistic that most statistical packages provide for testing the row-column independence in a contingency table. For the target table this statistic is indeed the deviance $D(m_0)$ between the root node m_0 and the target, while for the leaf table it is the deviance $D(m_0|m)$ between the root node and the leaf table associated to the induced tree. The deviance for the model is then just the difference between the two (see Figure 2)

$$D(m) = D(m_0) - D(m_0 \mid m) \ .$$

For our example, we obtain with SPSS $D(m_0) = 169.46$ and $D(m_0|m) = 167.77$, from which we deduce $D(m) = 169.46 - 167.77 = 1.69$. This is indeed

```
GET FILE='civst_gend_sector.sav'.
compute profiles
  = ngender*10^1 + nsect.
**Rules generated by AnswerTree**.
IF (ngender NE 2) AND (nsect NE 3)
  leaf = 3.
IF (ngender NE 2) AND (nsect EQ 3)
  leaf = 4.
IF (ngender EQ 2) AND (nsect EQ 1)
  leaf = 5.
IF (ngender EQ 2) AND (nsect NE 1)
  leaf = 6.
END IF.
**Computing the deviance**.
SORT CASES BY profiles .
AGGREGATE
  /OUTFILE='profiles.sav'
  /PRESORTED
  /BREAK=profiles
  /prof_mar = PIN(ncivstat 1 1)
  /prof_sgl = PIN(ncivstat 2 2)
  /prof_div = PIN(ncivstat 3 3)
  /leaf = first(leaf)
  /nj=N.
SORT CASES BY leaf.
AGGREGATE
  /OUTFILE='leaves.sav'
  /PRESORTED
  /BREAK=leaf
  /leaf_mar = PIN(ncivstat 1 1)
  /leaf_sgl = PIN(ncivstat 2 2)
  /leaf_div = PIN(ncivstat 3 3)
  /nj=N.
GET FILE='profiles.sav'.
SORT CASES BY leaf.
MATCH FILES /FILE=*
 /TABLE='leaves.sav'
 /RENAME (nj = d0)
 /DROP d0
 /BY leaf.
COMPUTE pre_mar=leaf_mar*nj/100.
COMPUTE pre_sgl=leaf_sgl*nj/100.
COMPUTE pre_div=leaf_div*nj/100.
COMPUTE n_mar=prof_mar*nj/100.
COMPUTE n_sgl=prof_sgl*nj/100.
COMPUTE n_div=prof_div*nj/100.
**Restructuring data table**.
VARSTOCASES
 /MAKE count
   FROM n_mar n_sgl n_div
 /MAKE pre
   FROM pre_mar pre_sgl pre_div
 /INDEX= Index1(3)
 /KEEP = profiles leaf
 /NULL = DROP
 /COUNT= nclass .

SELECT IF count > 0.
COMPUTE
  deviance=2*count*ln(count/pre).
SORT CASES BY leaf profiles.
COMPUTE newleaf = 1.
IF (leaf=lag(leaf,1))
    newleaf = 0.
COMPUTE newprof = 1.
IF (profiles=lag(profiles,1))
    newprof = 0.
COMPUTE one = 1.
FORMAT one (F2.0)
  /newleaf newprof (F8.0).
**Results in one row table**.
AGGREGATE
  /OUTFILE='deviance.sav'
  /PRESORTED
  /BREAK=one
  /deviance = sum(deviance)
  /nprof    = sum(newprof)
  /nleaves  = sum(newleaf)
  /nclass   = first(nclass)
  /ncells   = N.
GET FILE='deviance.sav'.
**DF and Significance**.
COMPUTE
  df=(nclass-1)*(nprof-nleaves).
COMPUTE
  sig=CDF.CHISQ(deviance,df).
EXECUTE.
```

Fig. 3. SPSS syntax for computing the deviance of the tree

the value we obtained by applying directly formula (1). Note that this approach is limited by the maximal number of columns (or rows) accepted for cross tables. This is for instance 1000 in SPSS 13, which makes this approach unapplicable when the number of possible profiles exceeds this number.

Let us now turn to the *second issue*, i.e. the possibly excessive number of a priori profiles. The solution we propose is to consider partial deviances. The idea is to define the target table from the mere predictors retained during the growing process. This will reduce the number of variables. We could go even further and group the values of each predictors according to the splits used in the tree. For instance, if the induced tree leads to the 3 leaves "male", "female and primary sector", "female and non primary sector", we would not distinguish between secondary and tertiary sectors. There would thus be 4 profiles — instead of 6 — for the target table, namely "male and primary sector", "male and non primary sector", "female and primary sector", "female and non primary sector".

The resulting target table T^* is clearly somewhat arbitrary. The consequence is that the partial deviance, i.e. the deviance $D(m|m_{T^*})$ between the tree m and T^*, has no real meaning by itself. However, we have $D(m) = D(m|m_{T^*}) + D(m_{T^*})$ thanks to the additivity property of the deviance. It follows that $D(m_2) - D(m_1) = D(m_2|m_{T^*}) - D(m_1|m_{T^*})$. The difference in the partial deviance of two nested trees m_1 and m_2 remains unchanged, whatever target m_{T^*} is used. Thus, all tests based on the comparison of deviances, between the fitted tree and the root node for example, remain applicable.

The partial deviance can also be used for defining AIC and BIC criteria, since only differences in the values of the latter matter. Pseudo R^2's, however, are not very informative when computed from partial deviances, due to the arbitrariness of the target table. It is preferable to consider the percent reduction in uncertainty, which does not depend on the target table, and to look at the square root of Theil's u or Goodman and Kruskal's τ.

5 Conclusion

With the deviance discussed in this article, we focused on the descriptive capacity of the tree, i.e. on its capacity to reproduce the outcome distribution for each possible profile in terms of the predictors. Such insights usefully complement the error rate that exclusively considers the classification performance. For instance, the loss in deviance that results from pruning the branches below "female" in our example has been shown to be statistically significant. This contrasts with the effect on the classification error which is not affected by this change. Though the deviance is not provided by available tree growing softwares, we have shown that it may readily be obtained either from independence Likelihood Ratio statistics provided for cross tables, or through for instance a SPSS syntax.

References

[Agr90] Agresti, A.: *Categorical Data Analysis*. Wiley, New York (1990)

[BFH75] Bishop, Y.M.M., Fienberg, S.E., Holland, P.W.: *Discrete Multivariate Analysis*. MIT Press, Cambridge MA (1975)

[BFOS84] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: *Classification And Regression Trees*. Chapman and Hall, New York (1984)

[CS89] Cox, D.R., Snell, E.J.: *The Analysis of Binary Data*. 2nd edn. Chapman and Hall, London (1989)

[GK54] Goodman, L.A., Kruskal, W.H.: Measures of association for cross classifications. *Journal of the American Statistical Association* **49** (1954) 732–764

[HK01] Han, J., Kamber, M.: *Data Mining: Concept and Techniques*. Morgan Kaufmann, San Francisco (2001)

[HMS01] Hand, D.J., Mannila, H., Smyth, P.: *Principles of Data Mining*. Adaptive Computation and Machine Learning. MIT Press, Cambridge MA (2001)

[HTF01] Hastie, T., Tibshirani, R., Friedman, J.: *The Elements of Statistical Learning*. Springer, New York (2001)

[Kas80] Kass, G.V.: An exploratory technique for investigating large quantities of categorical data. *Applied Statistics* **29** (1980) 119–127

[MN89] McCullagh, P., Nelder, J.A.: *Generalized Linear Models*. Chapman and Hall, London (1989)

[McF74] McFadden, D.: The measurment of urban travel demand. *Journal of Public Economics* **3** (1974) 303–328

[Mur98] Murthy, S.K.: Automatic construction of decision trees from data: A multi-disciplinary survey. *Data Mining and Knowledge Discovery* **2** (1998) 345–389

[Nag91] Nagelkerke, N.J.D.: A note on the general definition of the coefficient of determination. *Biometrika* **78** (1991) 691–692

[OR95] Olszak, M., Ritschard, G.: The behaviour of nominal and ordinal partial association measures. *The Statistician* **44** (1995) 195–212

[Qui86] Quinlan, J.R.: Induction of decision trees. *Machine Learning* **1** (1986) 81–106

[Qui93] Quinlan, J.R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo (1993)

[Raf95] Raftery, A.E.: Bayesian model selection in social research. In Marsden, P., ed.: *Sociological Methodology*. The American Sociological Association, Washington, DC (1995) 111–163

[RZ03] Ritschard, G., Zighed, D.A.: Goodness-of-fit measures for induction trees. In Zhong, N., Ras, Z., Tsumo, S., Suzuki, E., eds.: *Foundations of Intelligent Systems, ISMIS03*. Volume LNAI 2871. Springer, Berlin (2003) 57–64

[SPS01] SPSS, ed.: *Answer Tree 3.0 User's Guide*. SPSS Inc., Chicago (2001)

[The70] Theil, H.: On the estimation of relationships involving qualitative variables. *American Journal of Sociology* **76** (1970) 103–154