

Computing and using the deviance with classification trees

Gilbert Ritschard
 Dept of Econometrics, University of Geneva
 Compstat, Rome, August 2006

Outline

- 1 Introduction
- 2 Motivation
- 3 Deviance for Trees
- 4 Outcome for the mobility tree example
- 5 Computational Issues
- 6 Women's labour participation example
- 7 Conclusion

<http://mephisto.unige.ch>

1 Introduction

- About classification trees
- Descriptive non classificatory usages
- Measuring the quality of the tree (with the deviance)
- Computational issues

COMPSTAT06 toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 1

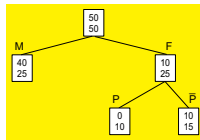
COMPSTAT06 toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 2

Principle of tree induction

Goal: Find a partition of data such that the **distribution** of the outcome variable **differs as much as possible** from one leaf to the other.

How: Proceeds by successively splitting nodes.

- Starting with root node, seek **attribute** that generates the **best split** according to a given **criterion**.
- **Repeat operation at each new node** until some stopping criterion, a minimal node size for instance, is met.



Main algorithms:

- CHAID (Kass, 1980), **significance of Chi-Squares**
- CART (Breiman et al., 1984), **Gini index**, binary trees
- C4.5 (Quinlan, 1993), **gain ratio**

COMPSTAT06 toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 3

2 Motivation

In social sciences, induced trees are most often used for **descriptive** (non classificatory) aims.

Examples:

- **Mobility trees between social statuses of sons, fathers and grandfathers** (data from act of marriage in the 19th century Geneva) (Ritschard and Oris, 2005)
Goal: How do the statuses of the father and grandfather **affect the chances** of the groom to be in a lower, medium or high position?
- **Determinants of women's labor participation** (Swiss census data) (Losa et al., 2006)
Goal: How do age, number of children, education, etc. **affect the chances** of the woman to work at full time, long part time, short part time or not to work at all?

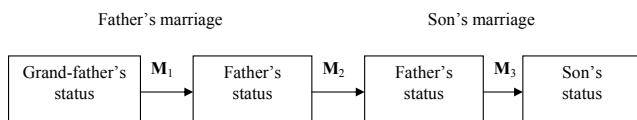
COMPSTAT06 toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 4

Mobility tree

Statuses defined from profession mentioned in marriage acts.

Acts for all men having a name beginning with a "B".

For 572 cases, was possible to match with data from father's marriage
 ⇒ **social mobility over 3 generations**



Groom's status (3 values) is response variable.

Predictors are birthplace and statuses of father and grandfather.

Method: CHAID (sig 5%, minimal child node size = 15, parent node = 30)

COMPSTAT06 toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 5

Mobility tree. Son's Status: Low (workers and craftsmen), Clock Maker, High



COMPSTAT06 toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 6

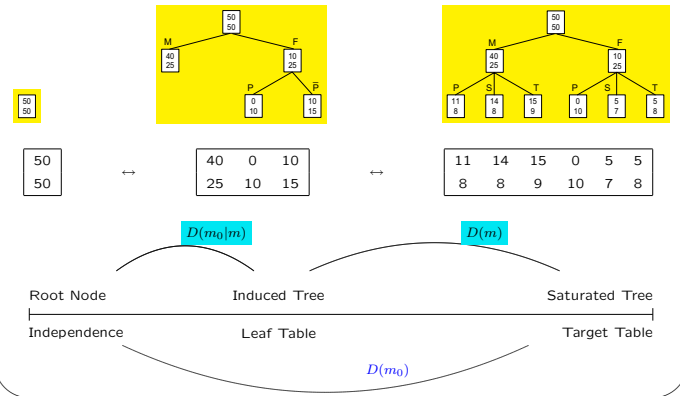
Validating Tree in a Non-classificatory Setting

- Trees are usually validated with the **classification error rate** (on test data or through cross-validation)
 - Claim:** Classification error rate **not suited for non classificatory purposes**
- Example: Split into two groups with distribution

$$\begin{bmatrix} 10\% \\ 90\% \end{bmatrix} \text{ and } \begin{bmatrix} 45\% \\ 55\% \end{bmatrix}$$

- Distributions clearly different (valuable knowledge)
- Split does not improve the error rate (assuming majority rule).
- Our suggestion** (Ritschard and Zighed, 2003): Use the **deviance** for measuring the descriptive power of a tree.

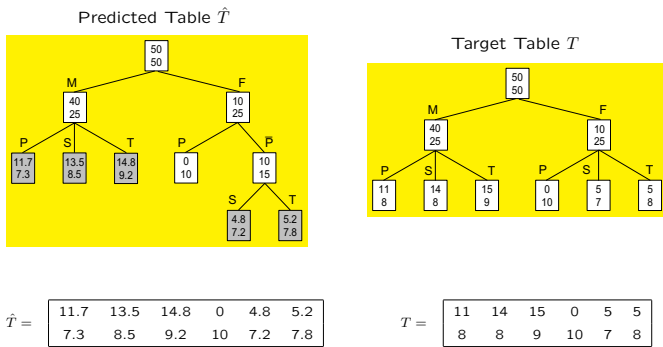
3 Deviance for Trees



COMPSTAT06 toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 7

COMPSTAT06 toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 8

Target and Predicted Tables



COMPSTAT06 toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 9

Deviance: Formal Definition

$T = (n_{ij})$ $r \times c$ target table:
 r rows = categories of the outcome variable
 c columns = different profiles in terms of the predictors

$\hat{T} = (\hat{n}_{ij})$ $r \times c$ table predicted from the tree

Total of each column (profile) distributed according to the distribution in the leaf to which the profile belongs

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

Under regularity conditions (Bishop et al., 1975):

- $D(m) \sim \chi^2$ with $d = (r-1)(c-q)$ degrees of freedom (see Ritschard and Zighed, 2003)
- $D(m_2|m_1) = D(m_2) - D(m_1) \sim \chi^2$ with $d_2 - d_1$ degrees of freedom if m_2 restricted version of m_1

COMPSTAT06 toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 10

Deviance based indicators

BIC: deviance penalized for complexity (nbr of parameters)

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

pseudo R^2 McFadden $R^2 = 1 - D(m)/D(m_0)$,

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

Theil's u (proportion of reduction of Shannon's entropy)

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

Evolves quadratically between independence and full association
 $\Rightarrow \sqrt{u}$ represents position between the 2 extremes.

COMPSTAT06 toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 11

4 Outcome for the mobility tree example

- Error rate: 42.4%, (55.6% at the root node; 10 folds CV: 51.4% error)
- Goodness of fit

Tree m	$D(m)$	df	sig	BIC	AIC	Theil \sqrt{u}
Indep	482.3	324	0.000	2319.6	812.3	0
Level 1	408.2	318	0.000	1493.9	750.2	0.25
Level 2	356.0	310	0.037	1492.5	714.0	0.32
Level 3	327.6	304	0.168	1502.2	697.6	0.36
Fitted	312.5	300	0.298	1512.5	690.5	0.37
Saturated	0	0	1	3104.7	978.0	0.63

COMPSTAT06 toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 12

Between level deviance improvement

$D(\text{row model}) - D(\text{column model})$

	Level 1	Level 2	Level 3	Fitted	Saturated
Indep	74.1*** (6 df)	126.3*** (14 df)	154.7*** (20 df)	169.8*** (24 df)	482.3*** (324 df)
Level 1		52.2*** (8 df)	80.6*** (14 df)	95.7*** (18 df)	408.2*** (318 df)
Level 2			28.4*** (6 df)	43.5*** (10 df)	356** (310 df)
Level 3				15.1*** (4 df)	327.6 (304 df)
Fitted					312.5 (300 df)

*** significant at 1%, ** at 5%, * at 10%

Between level BIC variation

$BIC(\text{row model}) - BIC(\text{column model})$

	Level 1	Level 2	Level 3	Fitted	Saturated
Indep	825.7	827.1	817.4	807.1	-785.1
Level 1	0	1.4	-8.3	-18.6	-1610.8
Level 2		0	-9.7	-20	-1612.2
Level 3			0	-10.3	-1602.5
Fitted				0	-1592.2

From the BIC standpoint, Level 1 and Level 2 models look the most interesting.

5 Computational Issues

- Softwares for growing trees **do not provide**
 - the deviance
 - nor easily usable information for computing the target and predicted tables

Solution: look at LR statistics for cross tables.
- Number of possible profiles (columns) may become **excessively large**.

May be as large as $\prod_{v=1}^V c_v$
with c_v the number of values of the v -th predictor

Solution: partial deviance (distance to a smaller arbitrary target table.)

Deviance and Likelihood Ratio Chi-squares

$D(m_0|m)$ = LR Chi-square statistic for testing independence on **Leaf Table** (crosstabulation of response variable with **leaf variable**).

$D(m_0)$ = LR Chi-square statistic for testing independence on **Target Table** (crosstabulation of response variable with **profile variable**).

These statistics can easily be computed with most statistical package (SPSS, SAS, ...)

Deviance of Tree m is just their difference

$$D(m) = D(m_0) - D(m_0|m)$$

Need just to retrieve for each case:

- leaf number
- profile number

Partial deviance $D(m|m_{T^*})$

Arbitrary $r \times c^*$ target table T^*

defined from the c^* profiles in terms of the mere **predictors** and value groupings **retained by the induced tree**.

Due to arbitrariness of T^*

- Deviance $D(m_{T^*})$ is **no longer** distance to true target.
- Pseudo R^2 's based on $D(m_{T^*})$ are irrelevant.

Differences of deviances between nested trees are **independent of the target**. For example:

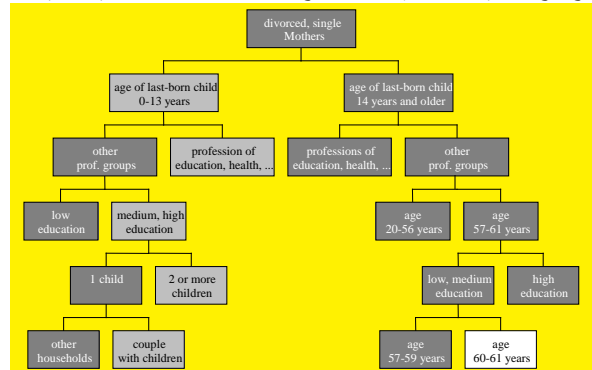
$$D(m_0|m) = D(m_0) - D(m) = D(m_0|m_{T^*}) - D(m|m_{T^*})$$

measures the gain over the root node (as the classical Chi-square used with logistic regression).

BIC and \sqrt{u} can still be used.

6 Women's labour participation example

Tree for participation of divorced or single mothers, French speaking region.



Quality of the trees

	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

	$\Delta\text{BIC}(m_0, m)$	$\Delta\text{BIC}(m_{T^*}, m)$	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

COMPSTAT06 toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 19

7 Conclusion

Summary:

- Deviance may be used with trees.
- Deviance and differences in deviances useful for evaluating the descriptive power of trees.
- Deviance based measures, such as BIC and Theil's u , also useful.
- Computation issues: solutions exist.

Further issues for descriptive trees:

- Using BIC as tree growing criterion.
- Evaluating the stability of induced trees (Dannegger, 2000).

COMPSTAT06 toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 20

THANK YOU

COMPSTAT06 toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 21

References

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

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

Dannegger, F. (2000). Tree stability diagnostics and some remedies for instability. *Statistics In Medicine* 19(4), 475–491.

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

Losa, F. B., P. Orioni, and G. Ritschard (2006). Experiences from a socio-economic application of induction trees. In N. Lavrač, L. Todorovski, and K. P. Jantke (Eds.), *Discovery Science, 9th International Conference, DS 2006, Barcelona, October 7-10, 2006, Proceedings*, Volume LNAI 4265, pp. 316–320. Berlin Heidelberg: Springer.

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

Ritschard, G. and M. Oris (2005). Life course data in demography and social sciences: Statistical and data mining approaches. In P. Ghisletta, J.-M. Le Goff, R. Levy, D. Spini, and E. Widmer (Eds.), *Towards an Interdisciplinary Perspective on the Life Course*, *Advancements in Life Course Research*, Vol. 10, pp. 289–320. Amsterdam: Elsevier.

References toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 22

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

References toc Intro Motiv MobTr Dev Ex1 Comp Ex2 Conc ◀▶▲▼ 26/8/2006gr 23