Classification trees with missing data

An application of Imprecise Probability Methods?

Paul Fink

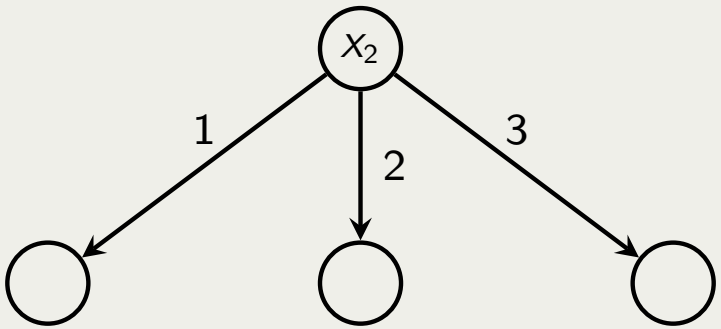
Department of Statistics, LMU Munich

16 December 2013

Introduction

Classification setup:

- Aim: Predicting value(s) of classification variable C based on feature variables $X_1, \ldots X_n$
- Method: Classification tree; splitting data space into disjoint homogeneous subspaces with respect to values of feature variables (k-array splitting)
- Impurity measure: Information Gain (IG) as split criterion



Splitting

Class probabilities within a node are estimated by relative frequencies (as in the classical setup)

$$IG_{X_i} = \sum_{x_i \in X_i} P(X_i = x_i) Ent(C|X_i = x_i)$$

How to deal with missing values?

- dropping
- imputation (most common)
- take into account by model

No assumption on missing process

$\Longrightarrow$ Need to ensure for the worst case!

Missing only in class

Assume we have the following table of observed class counts:

X \ Y	a	b	c	*mis*	Σ
0	n_{0a}	n_{0b}	n_{0c}	n_{0*}	n_0
1	n_{1a}	n_{1b}	n_{1c}	n_{1*}	n_1

for class variable $Y \in \{a, b, c\}$ and feature variable $X \in \{0, 1\}$; column *mis* gives the number of missing values

Of interest is the conditional probability distribution $P(Y|X)$:

If there were no missing values, one would get

$$\hat{P}(Y = a \,|\, X = 0) = \frac{n_{0a}}{n_0}$$

As there are n_{0*} missing values for $X = 0$ we need to take them into account!

$\Longrightarrow$ Rewriting the table by splitting column *mis* into 3 hidden states of Y: a^*, b^* and c^*.

In state a^* those of the missing values are collected whose original state was a before missing process changed it.

X \Y	a	b	c	a^*	b^*	c^*	Σ
0	n_{0a}	n_{0b}	n_{0c}	r	s	$n_{0*} - r - s$	n_0
1	n_{1a}	n_{1b}	n_{1c}	t	u	$n_{1*} - t - u$	n_1

Now conditional probability is given to:

$$\hat{P}(Y = a, Y = a^* \,|\, X = 0) = \frac{n_{0a} + r}{n_0} \quad .$$

Accounting for all configurations of class assignment within the missing values results in set of probability measures

This also holds for all other on X conditioned probabilities.

Bounds of interested quantity:

$$\frac{n_{0a}}{n_0} \leq \hat{P}(Y = a, Y = a^* \mid X = 0) \leq \frac{n_{0a} + n_{0*}}{n_0}$$

Bounds remind on ones obtained when estimating an IDM with $s = n_{0*}$ on complete data only.

$\Longrightarrow$ s as number of pseudo-counts in IDM

Further steps

Missing values do not influence $P(X = x)$ in IG, maximising of entropy part piecewise to maximise IG

When maximising entropy in the next step IDM structure could be exploited.

Table-based probability set is contained in IDM-set, so IDM approximation may be too cautious (upper bound)

$\Longrightarrow$ maximising entropy over true set

Missing only in features

No assumptions on missing process

Conditional Probability is more affected by missing values, as both nominator and denominator depends on actual number of missing values for a given X

Optimization of IG could be formalized as a Partial Identification problem.

$$IG_{X_i}^{\min} = \max_{\sigma_{x_i^*}} \sum_{x_i \in X_i} P(X_i = x_i) Ent(C|X_i = x_i)$$

where $\sigma_{x_i^*}$ means all compatible configurations of full data

Further ideas

Tree-wide accounting for missing values:

- Growing *simple* tree on each compatible data set
- Aggregation of trees
- Assumption of non-MAR still valid?