Embedded Variable Selection in Classification Trees

Servane Gey, Tristan Mary-Huard

Embedded Variable Selection in Classification Trees

Servane Gey[1], Tristan Mary-Huard[2]

[1]MAP5, UMR 8145, Université Paris Descartes, Paris, France
[2]UMR AgroParisTech/INRA 518, Paris, France

Overview

Introduction

★ Binary classification setting

★ Model and variable selection in classification

★ Classification tree

Variable selection for CART

★ Classes of classification trees

★ Theoretical results

★ Comparison with practice.

Binary classification

Prediction of the unknown label Y (0 or 1) of an observation X.

$\Rightarrow$ Use training sample $D = (X_1, Y_1), ..., (X_n, Y_n) \overset{i.i.d}{\sim} \mathbb{P}$ to build a classifier $\widehat{f}$

$$\begin{array}{rccc} \widehat{f}: & \mathscr{X} & \rightarrow & \{0, 1\} \\ & X & \mapsto & \widehat{Y}. \end{array}$$

Quality assessment

$\star$ Classification risk and loss : Quality of the resulting classifier $\widehat{f}$

$$\begin{aligned} L(\widehat{f}) &= \mathbb{P}(\widehat{f}(X) \neq Y | D) \\ \ell(\widehat{f}, f^*) &= L(\widehat{f}) - L(f^*) \end{aligned}$$

$\star$ Average loss : Quality of the classification algorithm

$$\mathbb{E}_D\left[\ell(\widehat{f}, f^*)\right]$$

Remark : All these quantities depend on $\mathbb{P}$ that is unknown.

Basics of Vapnik Theory : structural risk minimization (SRM)

Consider and collection of classes of classifiers $\mathscr{C}_1, ..., \mathscr{C}_M$. Define

$$\bar{f}_m = \arg\min_{f \in \mathscr{C}_m} L(f), \qquad \widehat{f}_m = \arg\min_{f \in \mathscr{C}_m} L_n(f), \qquad \widehat{f} = \arg\min_m \left[L_n(\widehat{f}_m) + \alpha \frac{V_{\mathscr{C}_m}}{n} \right]$$

★ Class complexity

If $\mathscr{C}_1, ..., \mathscr{C}_M$ have finite VC dimensions $V_{\mathscr{C}_1}, ... V_{\mathscr{C}_M}$, then

$$\mathbb{E}_D \left[\ell(\widehat{f}, f^*) \right] \leq C \left\{ \inf_m \left(\ell(\bar{f}_m, f^*) + K \sqrt{\frac{V_{\mathscr{C}_m}}{n}} \right) \right\} + \frac{\lambda}{n} \qquad \text{(Vapnik, 1998).}$$

★ Classification task complexity (Margin Assumption)

If there exists $h \in]0; 0.5[$ such that

$$\mathbb{P}\left(\left| \eta(x) - 1/2 \right| \leq h \right) = 0, \quad \text{with } \eta(x) = \mathbb{P}(Y = 1 | X = x)$$

then

$$\mathbb{E}_D \left[\ell(\widehat{f}, f^*) \right] \leq C \left\{ \inf_m \left(\ell(\bar{f}_m, f^*) + K' \left(\frac{V_{\mathscr{C}_m}}{n} \right) \right) \right\} + \frac{\lambda'}{n} \qquad \text{(Massart \& Nédélec, 2006).}$$

Application to variable selection in classification

Assume that $X \in \mathbb{R}^p$. Define

$$\bar{f}_{m(k)} = \underset{f \in \mathscr{C}_{m(k)}}{\arg\min}\, L(f), \qquad \widehat{f}_{m(k)} = \underset{f \in \mathscr{C}_{m(k)}}{\arg\min}\, L_n(f)$$

★ Variable selection

Choose $\widehat{f}$ such that

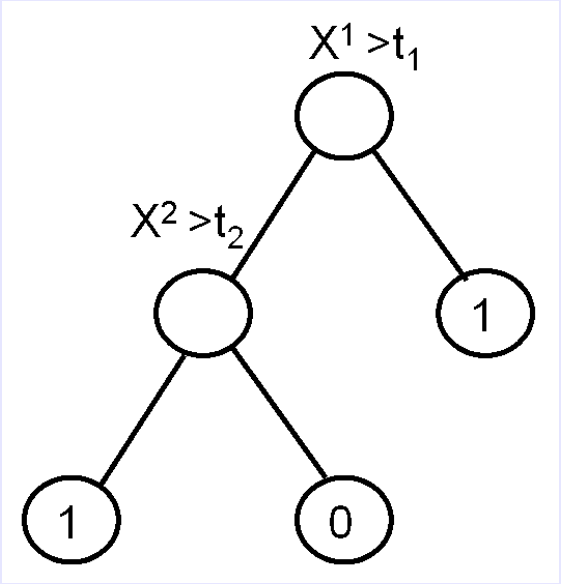
$$\widehat{f} = \underset{m(k)}{\arg\min} \left[L_n\left(\widehat{f}_{m(k)}\right) + \alpha \frac{V_{\mathscr{C}_{m(k)}}}{n} + \alpha' \log\left[\binom{p}{k}\right] \right]$$

Then (under strong margin assumption)

$$\mathbb{E}_D\left[\ell\left(\widehat{f}, f^*\right)\right] \leq C \log(p) \left\{ \inf_{m(k)} \left(\ell\left(\bar{f}_{m(k)}, f^*\right) + K' \left(\frac{V_{\mathscr{C}_{m(k)}}}{n} \right) \right) \right\} + \frac{\lambda}{n}$$

(Massart, 2000, Mary-Huard et al., 2007)

Classification trees



General strategy

Choose

$$\widehat{f} = \arg\min_{T} L_n\left(f_T\right) + \alpha \frac{|T|}{n}$$

Heuristic approach (CART, Breiman, 1984)

★ Find a tree $T_{\max}$ such that $L_n\left(f_{T_{\max}}\right) = 0$,

★ Prune $T_{\max}$ using criterion :

$$\widehat{f} = \arg\min_{T \subseteq T_{max}} L_n\left(f_T\right) + \alpha \frac{|T|}{n}$$

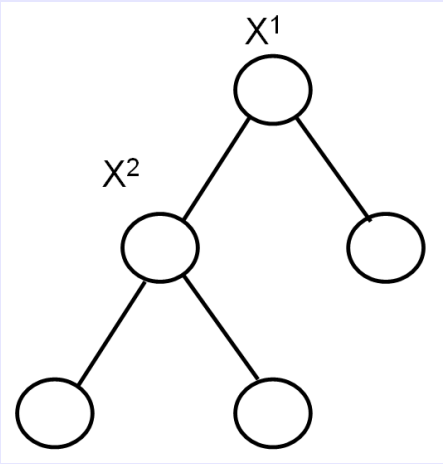
Definitions

Consider a tree $T_{c\ell}$ with

- a given configuration c,

- a given list ℓ of associated variables.

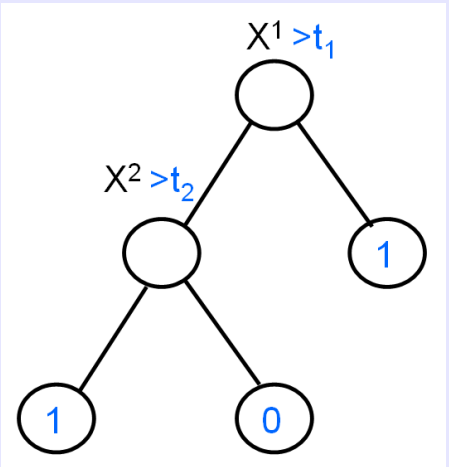
Remark : A same variable may be associated with several nodes.



Class of tree classifiers

Define

$$\begin{aligned}
\mathscr{C}_{c\ell} &= \{f / f \text{ based on } T_{c\ell}\} \ , \\
H_{c\ell} &= \text{VC log-entropy of class } \mathscr{C}_{c\ell}, \\
\bar{f}_{c\ell} &= \arg\min_{f \in \mathscr{C}_{c\ell}} L(f) \ , \\
\widehat{f}_{c\ell} &= \arg\min_{f \in \mathscr{C}_{c\ell}} L_n(f) \ .
\end{aligned}$$



Remark : Two classifiers $f, f' \in \mathscr{C}_{c\ell}$ only differ in their thresholds and labels.

Risk bound for one class

Proposition

Assume that strong margin assumption is satisfied. For all $C > 1$, there exist positive constants K^1 and K^2 depending on C such that

$$\mathbb{E}_D\left[\ell\left(\widehat{f}_{c\ell}, f^*\right)\right] \leq C\left\{\ell\left(\bar{f}_{c\ell}, f^*\right) + K^1\left(\frac{|T_{c\ell}|\log(2n)}{n}\right)\right\} + \frac{K^2}{n} \quad .$$

Idea of proof

★ Show that $E\left[H_{c\ell}\right] \leq |T_{c\ell}|\log(2n)$,

★ Apply general theory (Koltchinskii, 2006).

Combinatorics for variable selection

To take into account variable selection in the penalized criterion, one needs to count the number of classes sharing the same a priori complexity.

★ Parametric case (Logistic regression, LDA,...)

- One parameter per variable,
- 2 classes with classifiers based on k variables have the same a priori complexity,

$\Rightarrow \binom{p}{k}$ classes of a priori complexity k.

★ Classification trees

- One parameter per internal node (threshold),
- 2 classes $\mathscr{C}_{c\ell}$ and $\mathscr{C}_{c'\ell'}$ such that $|T_{c\ell}| = |T_{c'\ell'}|$ have the same a priori complexity

$\Rightarrow$ Count the number of classes based on trees of size k !

Combinatorics for variable selection

A tree $T_{c\ell}$ is defined by

- a configuration,

- a list of variables associated with each node.

★ Number of configurations of size k :

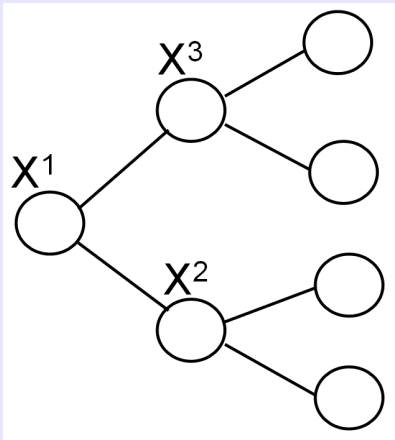
$$N_c^k = \frac{1}{k}\binom{2k-2}{k-1}$$

★ Number of variable lists of size k :

- the list is ordered : $\{1,2,3\} \neq \{2,1,3\}$,

- variables are selected with replacement : $\{1,2,1\}$.

$$\Rightarrow N_\ell^k = p^{k-1} \text{ instead of } \binom{p}{k} !$$



★ Number of classes based on trees of size $|T_{c\ell}| = k$:

$$N^k = N_c^k \times N_\ell^k = \frac{1}{k}\binom{2k-2}{k-1} \times p^{k-1}$$

$$\Rightarrow \log(N^k) \leq \lambda |T_{c\ell}| \log(p)$$

Combinatorics for variable selection

A tree $T_{c\ell}$ is defined by

- a configuration,

- a list of variables associated with each node.

★ Number of configurations of size k :

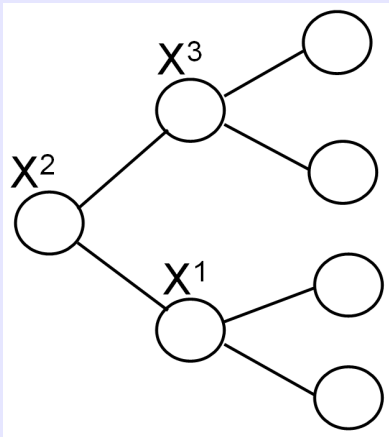
$$N_c^k = \frac{1}{k}\binom{2k-2}{k-1}$$

★ Number of variable lists of size k :

- the list is ordered : $\{1,2,3\} \neq \{2,1,3\}$,

- variables are selected with replacement : $\{1,2,1\}$.

$$\Rightarrow N_\ell^k = p^{k-1} \text{ instead of } \binom{p}{k} !$$



★ Number of classes based on trees of size $|T_{c\ell}| = k$:

$$N^k = N_c^k \times N_\ell^k = \frac{1}{k}\binom{2k-2}{k-1} \times p^{k-1}$$

$$\Rightarrow \log\left(N^k\right) \leq \lambda |T_{c\ell}| \log(p)$$

Combinatorics for variable selection

A tree $T_{c\ell}$ is defined by

- a configuration,

- a list of variables associated with each node.

★ Number of configurations of size k :

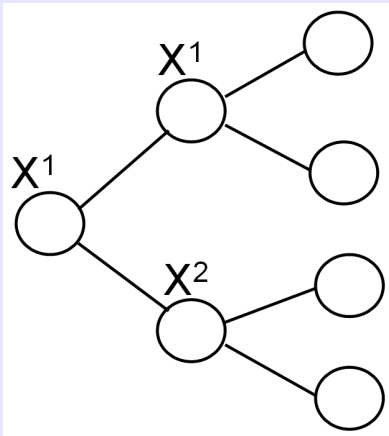
$$N_c^k = \frac{1}{k}\binom{2k-2}{k-1}$$

★ Number of variable lists of size k :

- the list is ordered : $\{1,2,3\} \neq \{2,1,3\}$,

- variables are selected with replacement : $\{1,2,1\}$.

$$\Rightarrow N_\ell^k = p^{k-1} \text{ instead of } \binom{p}{k} !$$



★ Number of classes based on trees of size $|T_{c\ell}| = k$:

$$N^k = N_c^k \times N_\ell^k = \frac{1}{k}\binom{2k-2}{k-1} \times p^{k-1}$$

$$\Rightarrow \log(N^k) \leq \lambda |T_{c\ell}| \log(p)$$

Risk bound for tree classifiers

Proposition

Assume that strong margin assumption is satisfied. If

$$\widehat{f} = \arg\min_{c,\ell} \left(L_n(\widehat{f}_{c\ell}) + pen(c,\ell) \right),$$

$$\text{where } pen(c,\ell) = C_h^1 \frac{|T_{c\ell}| \log(2n)}{n} + C_h^2 \frac{|T_{c\ell}| \log(p)}{n}$$

with constants C_h^1, C_h^2 depending on h appearing in the margin condition, then there exist positive constants C, C', C'' such that

$$\mathbb{E}_D[l(\widehat{f}, f^*)] \leq C \log(p) \left\{ \inf_{c,\ell} \left\{ \ell(\bar{f}_{c\ell}, f^*) + C' \left(\frac{|T_{c\ell}| \log(2n)}{n} \right) \right\} \right\} + \frac{C''}{n}.$$

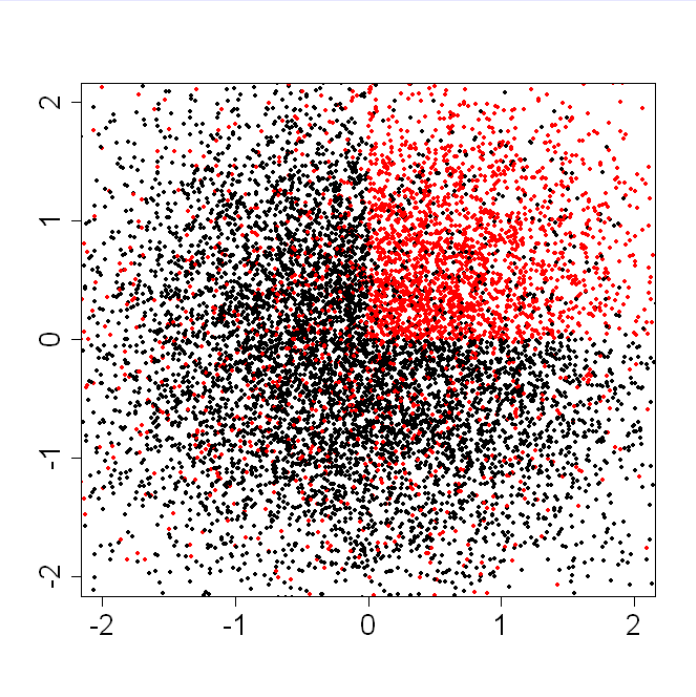
Remark :

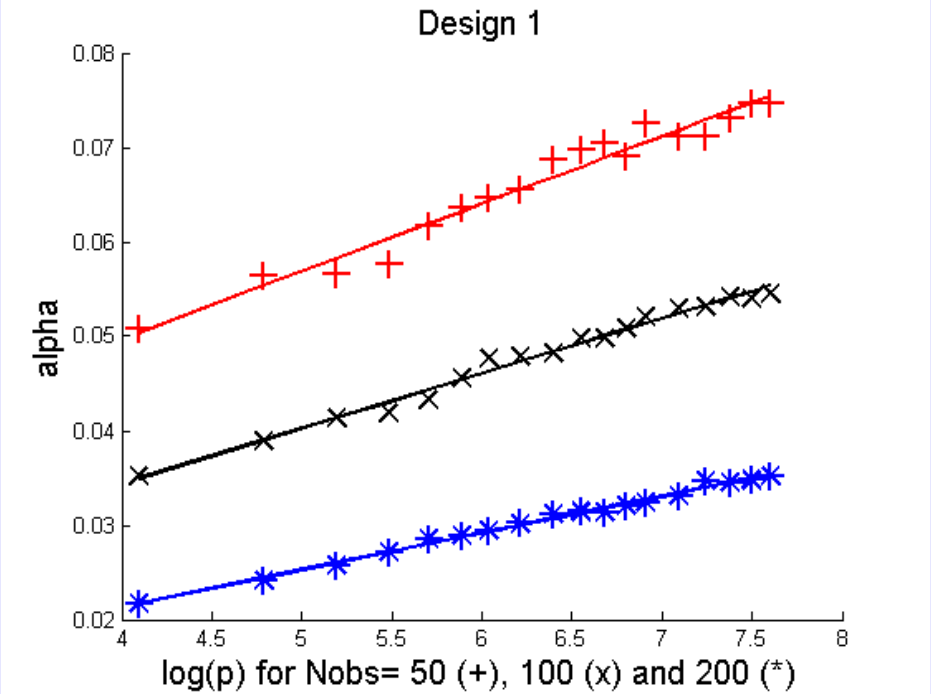
$$\text{Theory : } pen(c,\ell) = (a_n + b_n \log(p)) |T_{c\ell}| = \alpha_{p,n} |T_{c\ell}|$$

$$\text{Practice (CART) : } pen(c,\ell) = \alpha_{CV} |T_{c\ell}|$$

Does α_{CV} match $\alpha_{p,n}$?

Illustration on simulated data (1)



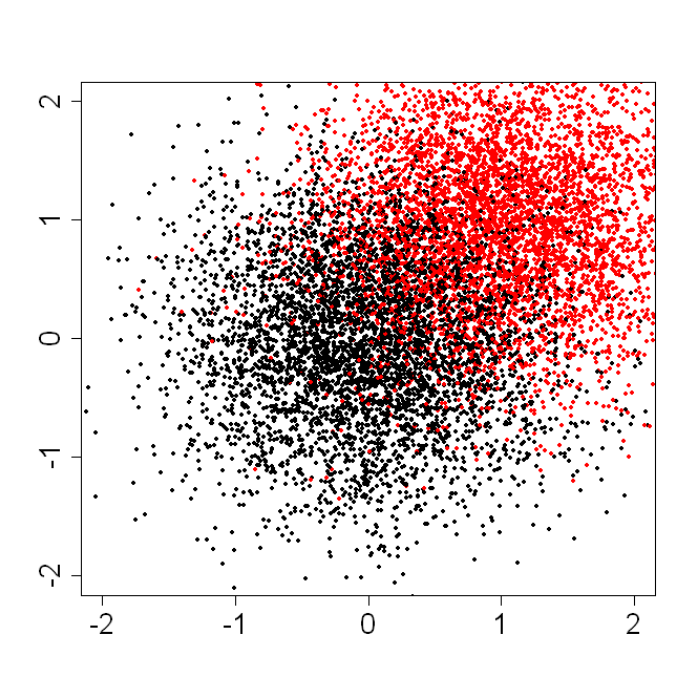


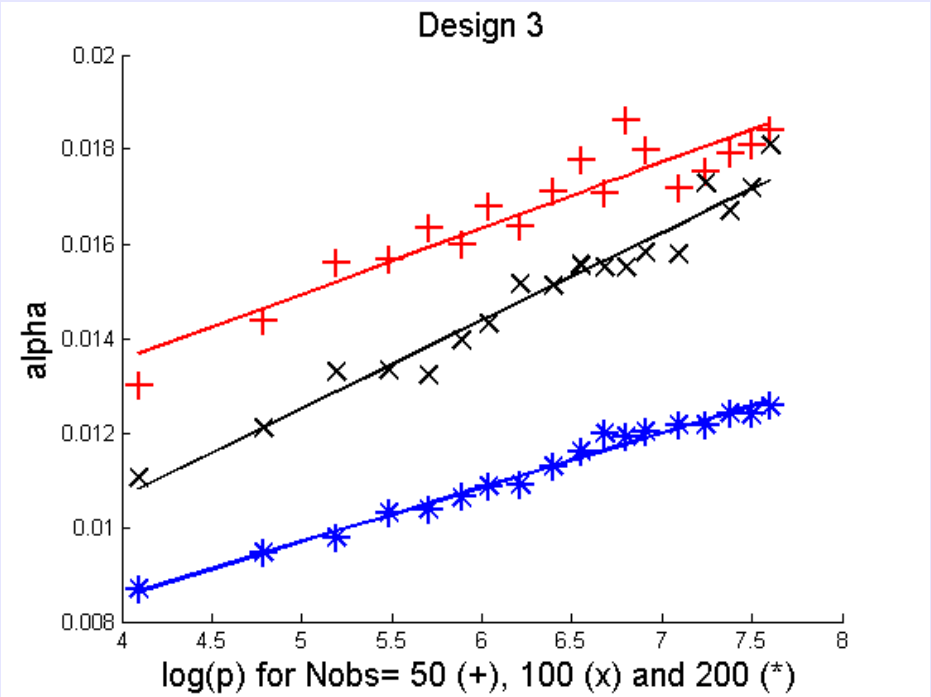
- Variables $X^1, ..., X^p$ are independent,
- If $X^1 > 0$ and $X^2 > 0$ $P(Y = 1) = q$, otherwise $P(Y = 1) = 1 - q$

Remark : Easy case

- The Bayes classifier belongs to the collection of classes,
- Strong margin assumption is satisfied.

Illustration on simulated data (2)





- $P(Y=1)=0.5$
- For $j=1,2$, $X^j|Y=0 \hookrightarrow \mathcal{N}(0,\sigma^2)$ and $X^j|Y=1 \hookrightarrow \mathcal{N}(1,\sigma^2)$,
- Additional variables are independent and non-informative.

Remark : Difficult case

- The Bayes classifier does NOT belong to the collection of classes,
- Strong margin assumption is NOT satisfied.

Conclusion

Model selection for tree classifiers :

- Already investigated (Nobel 02, Gey & Nedelec 06, Gey 10),
- Variable selection not investigated so far.
- Pruning step now validated from this point of view.

Theory vs practice

- Theory : exhaustive search,
- Practice : forward strategy,
- Nonetheless theoretical results are informative !

Extension

- In this talk : strong margin assumption
- Can be extended to less restrictive margin assumption
- Manuscript on arXiv.org :

http ://arxiv.org/abs/1108.0757

Bibliography

Breiman L., Friedman J., Olshen R. & Stone, C. (1984) *Classification And Regression Trees*, Chapman & Hall.

Gey S. & Nédélec E. (2005) *Model selection for CART regression trees*, IEEE Trans. Inform. Theory, 51, 658–670.

Koltchinskii, V. (2006) *Local Rademacher Complexities and Oracle Inequalities in Risk Minimization*, Annals of Statistics, 34, 2593–2656.

Mary-Huard T., Robin S. & Daudin J.-J. (2007) *A penalized criterion for variable selection in classification*, J. of Mult. Anal., 98, 695–705.

Massart P. (2000) *Some applications of concentration inequalities to statistics*, Annales de la Faculté des Sciences de Toulouse.

Massart P. & Nédélec E. (2006) *Risk Bounds for Statistical Learning*, Annals of Statistics, 34, 2326–2366.

Nobel A.B. (2002) *Analysis of a complexity-based pruning scheme for classification trees*, IEEE Trans. Inform. Theory, 48, 2362–2368.