Discussion on Minimal Penalty and the Slope Heuristic: A Survey, by S. Arlot

Servane Gey

Discussion on "*Minimal penalties and the slope heuristics: a survey*" by Sylvain Arlot

Dr. Servane Gey, UMR8145 MAP5, University of Paris, France.
Servane.Gey@parisdescartes.fr

This paper gives a very complete overview of the state of the art on minimal penalties and slope heuristics developed these past 20 years. Sylvain Arlot has to be congratulated for providing a survey which clearly relies theory and heuristics, and which proposes several ways to proceed heuristics in practice. My comments will focus on binary classification with the 0-1 loss, for which, as mentioned several times by Sylvain Arlot, there is still no clear results on slope heuristics.
We propose to illustrate the behavior of two slope heuristics by an empirical simulation study, and on the benchmark well-known *spam* data set. We focus on classification trees model selection with the 0-1 loss, and we use the CART algorithm proposed by Breiman *et al.* in 1984 [3] to build the models. The results of the slope heuristics are compared with the ones obtained by the classical 10-fold cross-validation via the prediction errors of the selected classifiers, and the dimensions of the corresponding classification tree models. Some final comments are given at the end of the discussion.

1 CART and Slope Heurisitics

The pruning procedure used in CART is based on a penalized empirical misclassification rate criterion, with a penalty proportional to the classification trees' number of leaves. It allows to reduce drastically the collection of candidate tree classifiers by providing a collection of nested models $m_K \preceq \ldots \preceq m_1$, associated with an increasing sequence of complexity parameters $\widehat{C}_1 < \cdots < \widehat{C}_K$, with $\widehat{C}_1 = 0$ and $D_{m_K} = 1$. Here the dimension D_m of a classification tree model m is the number of its leaves. Gey *et al.* [4, 5] obtained oracle-type inequalities with the 0-1 loss under the strong margin assumption (denoted by SMA in the following) proposed by Massart and Nédélec [9] (see also the paper by Bartlett *et al.* [2]), which is a particular case of the margin assumptions proposed by Mammen and Tsybakov [8], and later generalized by Koltchinskii [7]. These results show that the penalty term to use in classification tree model selection is of the form $(C/h)(D_m/n)$, where C is a large enough unknown constant, and $h \in (0, 1]$ is the unknown margin parameter of the n observations' common distribution. Hence each complexity parameter $(\widehat{C}_i)_{1 \leqslant i \leqslant K}$ is a data-driven calibration of the unknown penalty constant (C/h). Let us mention that subadditive penalties can be used in the slope heuristics to select a final tree after pruning (see the paper by Scott [10]).
To be able to select automatically a model via the slope heuristics, we use heuristics proposed by Bar-Hen, Gey and Poggi [1], well-adapted to CART classification trees: a first tree model $\hat{m}_{jump}$ is selected via a modified version of Algorithm 5 with $\text{pen}_0 = \text{pen}_1$ (see Appendix); a second tree model $\hat{m}_{plateau}$ is selected by taking $plateau = \text{argmax}_i \left(\widehat{C}_{i+1} - \widehat{C}_i \right)$, corresponding to the elbow selection. A typical example of the behavior of $\widehat{C}_i \mapsto D_{m_i}$ is given in Figure 1.

2 Simulation study

The simulation designs are described in Table 1. Since all the covariates have symmetrical roles, only designs with 2 covariates are considered. For each, a representation of one sample's realization of size $n = 1000$ is given in Figure 2, with fixed global or local margin parameter $h = 0.9$. The represented partitions are the ones of the true underlying conditional distribution. *Checkerbord* and *Crux* designs are easy for CART: the Bayes classifiers are trees, and the SMA is fulfilled. The two other ones are difficult for CART: in the *Line* design, the SMA is fulfilled, but the Bayes classifier's model is very difficult to approximate by trees; in the *Square* design, the Bayes classifier is a tree, but the SMA is not fulfilled.
The CART models are built using the R packages `rpart` for 10-fold cross validation (denoted by *CV*), and `tree` for the two slope heuristics (denoted by *Jump* and *Plateau* respectively). For each design, the methods are compared with respect to the value of the global or local margin parameter h: for each value $h \in [0, 1]$ on a regular grid, the average CART tree classifiers' risks and model's dimensions over 400 samples of size $n = 1000$ are computed. The risk of a tree is computed as its expected misclassification rate (estimated on a large test sample of size 2000), minus the Bayes error. Designs' Bayes errors can be found in Table 1.

The results are represented in Figure 3 for the easy designs, and in Figure 4 for the difficult ones. The calibration parameter α for the *Jump* heuristic is taken as $\alpha = 10\%$ for the *Checkerboard*, *Crux* and *Square* designs, and $\alpha = 5\%$ for the *Line* design. The black lines on the dimensions' graphs represent the dimension to take under the true underlying observations' distribution if it were known.
When h is close to 0, the risk of every selected model is close to 0 since the underlying true labels' distribution depends slightly on the covariates' one, and is Bernoulli with parameter close to 1/2. Nevertheless, one can see that the slope heuristics select more intuitive low dimensional models, while *CV* selects much larger dimensional ones to better separate labels.
As soon as $h \geqslant 0.75$, the three methods give similar results on the easy designs, and recover the true models, what is encouraging for the use of slope heuristics for classification trees. Let us just mention that there are border artefacts on the regions' limits, leading *CV* to select slightly larger dimensional trees. These artefacts seem to be automatically compensated by the slope heuristics, at a small cost on the risk. One can also see that all risks present a maximum. The shapes' decreasing part confirms the penalty term's dependency on $1/h$ as soon as h is sufficiently large. The shapes' increasing part might correspond to margin values for which the penalty term is not a linear function of the dimension (see [9, 5]).
When there is an approximation bias (see *Line* on Figure 4), the slope heuristics differ from *CV* by choosing smaller dimensional models rather than decreasing bias. *CV* and *Jump* heuristic's risks stabilize with increasing margin parameter, while *Plateau* heuristic's risk keeps on increasing, what is characteristic of overpenalization. For the *Square* design (see *Square* on Figure 4), the risks behave the same way as for the easy designs, but the models' dimension differ largely as soon as $h \geqslant 0.25$: the risks are almost equal, but *CV* selects high dimensional models to cope with the region where the SMA is not fulfilled, while the slope heuristics "forget" it and recover the true 2-dimensional intuitive model.

3 Application on the *spam* data set

The *spam* data set consists of information from 4601 email messages, in a study to screen email for "spam" (i.e. junk email). The data are presented in details in [6, p. 301]. The response has

values nonspam or spam, and there are 57 covariates relative to specific words and characters indicators in the email. This data set can be found in the `kernlab` R package, or on the UCI Machine Learning data base: https://archive.ics.uci.edu/ml/datasets/spambase.
The three methods described in Section 2 are applied on the *spam* data set, with a calibration parameter $\alpha = 1\%$ for the *Jump* heuristic. The function $D_{m_i} \mapsto \widehat{C}_i$ is represented in Figure 5, and the CART trees computed on the whole data set are represented in Figure 6. Let us mention that the probability for an email to be a spam is estimated far from 0.5 in every leaf of the corresponding probability tree, what might indicate that the SMA is fulfilled for these data. The methods are compared through Monte-Carlo average prediction errors and models' dimensions computed on 400 random drawing of a learning set to build the trees, and a test set representing 10% of the data to estimate the prediction error. The results are presented in Table 2: *CV* and the *Jump* heuristic select almost always the same 7-dimensional tree, while the *Plateau* heuristic always selects the 2-dimensional tree. The prediction errors indicate that the *Plateau* heuristic overpenalizes too much compared to the two other methods, which have comparable and much better performance on the data.

4 Comments

The empirical results obtained for classification trees are very encouraging for the use of slope heuristics for classification with the 0-1 loss. They show that, even if the slope heuristics do not perform as well as cross-validation, they seem to adapt better to the strong margin assumption. It has to be noticed that one drawback of the *Plateau* elbow heuristic seems to be overpenalization.
Further investigations should be made with models more regular than classification trees, and with model selection methods allowing to visit all models' dimensions. Also, as for classification with the quadratic loss, it could be interesting to investigate theory under margin assumptions to have a better handle of the slope heuristics' performance with respect to margin parameters.

References

[1] Bar-Hen, A., Gey, S., and Poggi, J.-M. Spatial cart classification trees.

[2] Bartlett, P. L., Jordan, M. I., and McAuliffe, J. D. Convexity, classification, and risk bounds. Journal of the American Statistical Association 101, 473 (2006), 138–156.

[3] Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. Classification and regression trees. CRC press, 1984.

[4] Gey, S. Risk bounds for cart classifiers under a margin condition. Pattern Recognition 45 (2012), 3523–3534.

[5] Gey, S., and Mary-Huard, T. Risk bounds for embedded variable selection in classification trees. IEEE Transactions on Information Theory 60, 3 (2014), 1688–1699.

[6] Hastie, T., Tibshirani, R., and Friedman, J. The elements of statistical learning. Springer Series in Statistics. Springer-Verlag, New York, 2001. Data mining, inference, and prediction.

[7] Koltchinskii, V. Local Rademacher complexities and oracle inequalities in risk minimization. Ann. Statist. 34, 6 (2006), 2593–2656.

[8] Mammen, E., and Tsybakov, A. B. Smooth discrimination analysis. Ann. Statist. 27, 6 (1999), 1808–1829.

[9] Massart, P., and Nédélec, É. Risk bounds for statistical learning. Ann. Statist. 34, 5 (2006), 2326–2366.

[10] Scott, C. Tree pruning with subadditive penalties. IEEE Transactions on Signal Processing 53, 14 (2005), 4518–4525.

Appendix

$\hat{m}_{jump}$ selection for CART classification trees The pruning step of the CART algorithm does not visit all possible configurations, nor all possible dimensions. Hence, if there exist negligible plateaus in the graph of $\widehat{C}_{m_i} \mapsto D_{m_i}$, then the corresponding dimensions are considered as artefacts of the pruning procedure. These negligible plateaus, and therefore the corresponding tree dimensions, are then removed, and the final tree model is selected through the classical maximal jump in this new collection of dimensions. In practice, only plateaus occurring before the largest one are considered, and a plateau is set negligible if it represents less than a proportion α of the largest one (where α is chosen with respect to the relative size of the largest plateau).

Tables

Data	Covariates	Response's Distribution	Bayes error
Checkerboard	$X_i \sim \mathcal{U}([0,1])$ $i = 1, 2$	$Y \in \{blue; red\} := \{0; 1\}$ On a regular 3 by 3 *blue* and *red* checkerboard; if (X_1, X_2) belongs to a *blue* square, $Y \sim \mathcal{B}(1, \frac{1-h}{2})$ if (X_1, X_2) belongs to a *red* square, $Y \sim \mathcal{B}(1, \frac{1+h}{2})$	$\frac{1-h}{2}$
Crux	$X_i \sim \mathcal{N}(0,1)$ $i = 1, 2$	$Y \in \{blue; red\} := \{0; 1\}$ if $X_1 > 0$ and $X_2 > 1/2$, or if $X_1 < 0$ and $X_2 < 1/2$, $Y \sim \mathcal{B}(1, \frac{1+h}{2})$ $Y \sim \mathcal{B}(1, \frac{1-h}{2})$ otherwise	$\frac{1-h}{2}$
Line	Same as *Crux*	$Y \in \{blue; red\} := \{0; 1\}$ if $X_1 + X_2 > 0$, $Y \sim \mathcal{B}(1, \frac{1+h}{2})$ $Y \sim \mathcal{B}(1, \frac{1-h}{2})$ otherwise	$\frac{1-h}{2}$
Square	Same as *Checkerboard*	$Y \in \{blue; red\} := \{0; 1\}$ unit square split into 3 parts: one center square U_c of surface $1/2 \times 1/2$, two parts U_a above and U_b below U_c, U_a and U_b delimited by the line $x_2 = 1/2$; if $(X_1, X_2) \in U_a$, $Y \sim \mathcal{B}(1, \frac{1+h}{2})$ if $(X_1, X_2) \in U_b$, $Y \sim \mathcal{B}(1, \frac{1-h}{2})$ if $(X_1, X_2) \in U_c$, $Y \sim \mathcal{B}(1, \frac{1}{2} + \frac{1}{\sqrt{n}})$ with n the number of simulated data	$\frac{1-0.75\times h}{2} - \frac{1}{4\sqrt{n}}$

Table 1: *Simulated data sets, with $h \in [0, 1]$ the global or local margin parameter of the response/covariates conditional distribution.*

	10-fold CV	Jump heuristic	Plateau heuristic
Prediction error	9.7%	11.4%	21.6%
Dimension	7	7.3	2

Table 2: *Average prediction error (in percentage) and tree's dimension over* 400 *learning/test random drawing for the* `spam` *dataset.*

Figures

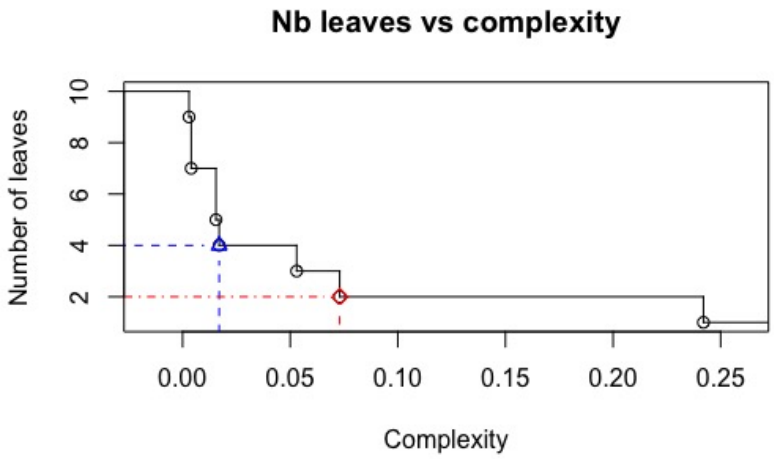

Figure 1: Typical example of the behavior of dimension $(D_{m_i})_{1\leqslant i\leqslant K}$ with respect to complexity parameter $(\widehat{C}_i)_{1\leqslant i\leqslant K}$ for CART classification trees. $\triangle$ represents the tree $\hat{m}_{jump}$, while $\diamond$ represents the tree $\hat{m}_{plateau}$.

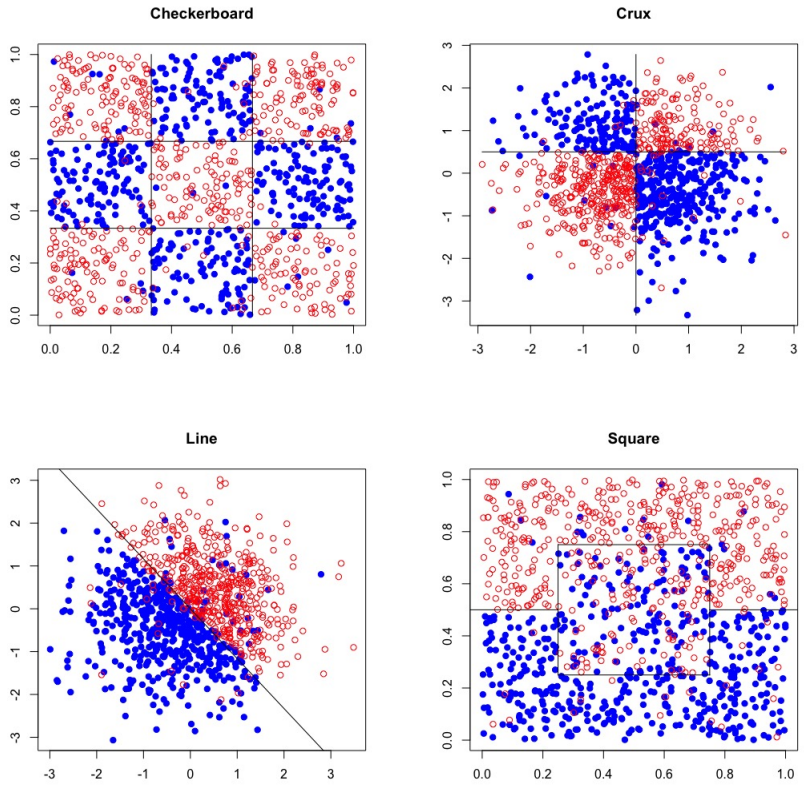

Figure 2: Examples of one sample's realization of size $n = 1000$ for each design, with common global or local margin parameter $h = 0.9$. From left to right, and up to bottom: *Checkerboard*, *Crux*, *Line*, *Square*.

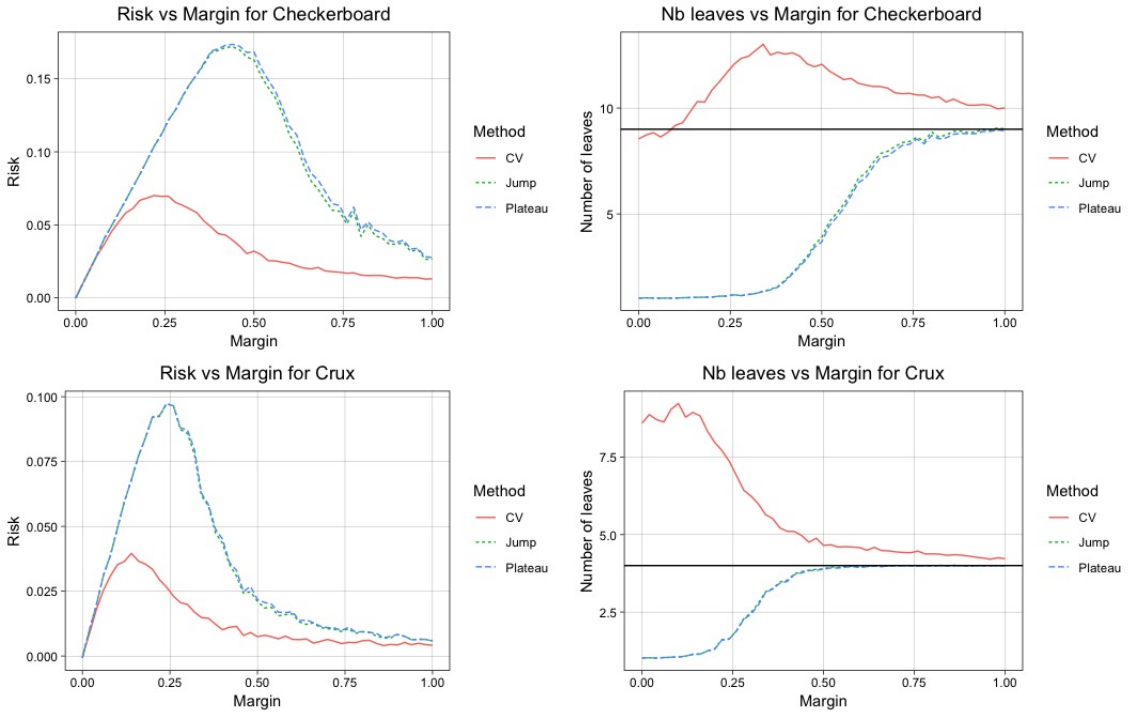

Figure 3: Average risk (*left*) and model's dimension (*right*) obtained by CV, Jump and Plateau heuristics over 400 simulations of the *Checkerboard* (*up*) and *Crux* (*bottom*) simulated data sets. *Black line*: dimension of the Bayes classifier's model.

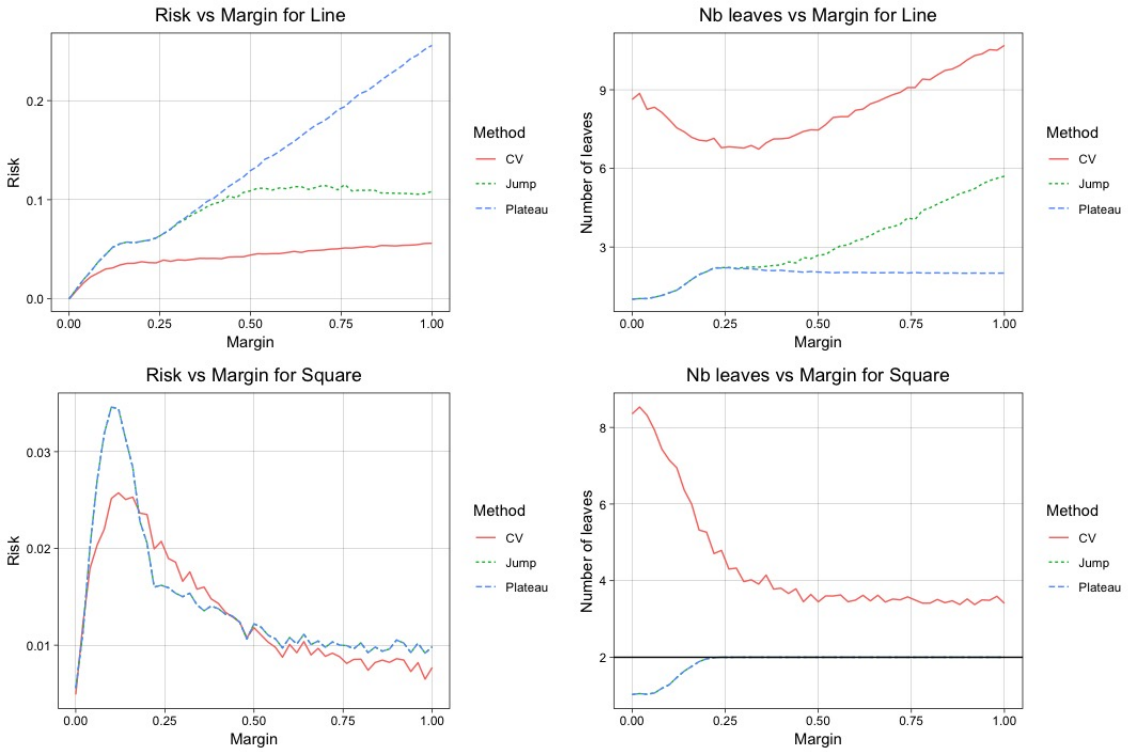

Figure 4: Average risk (*left*) and model's dimension (*right*) obtained by CV, Jump and Plateau heuristics over 400 simulations of the *Line* (*up*) and *Square* (*bottom*) simulated data sets. *Black line*: dimension to take if the true underlying observations' common distribution were known for the *Square* design.

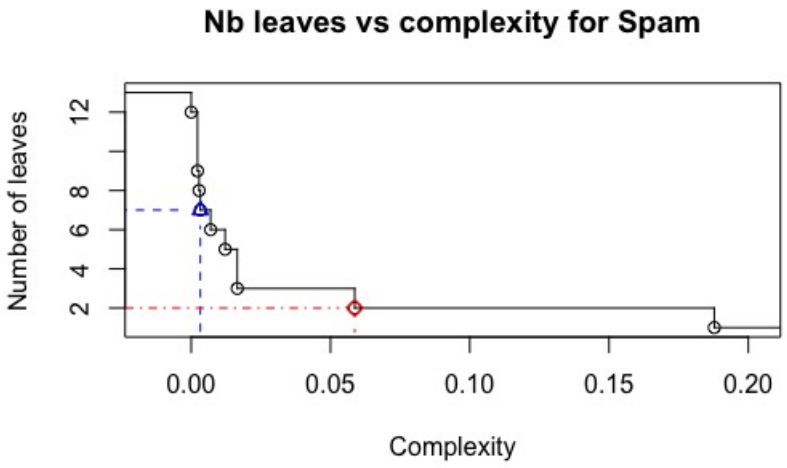

Figure 5: Dimension $(D_{m_i})_{1\leqslant i\leqslant K}$ with respect to complexity parameter $(\widehat{C}_i)_{1\leqslant i\leqslant K}$ for the *spam* data set. $\triangle$ represents the tree $\hat{m}_{jump}$, while $\diamond$ represents the tree $\hat{m}_{plateau}$.

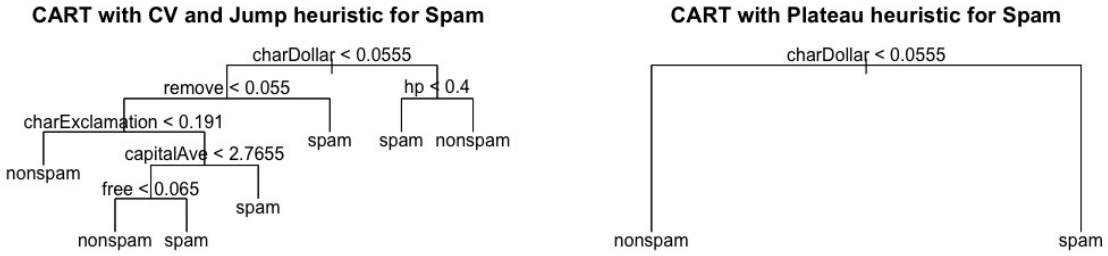

Figure 6: Classification trees for the spam data set. *Left*: tree selected via CV and Jump heuristic; *Right*: tree selected via Plateau heuristic.