STATISTICAL LEARNING SYSTEMS

LECTURE 14: SEMI-SUPERVISED LEARNING. SMALL n LARGE p PROBLEMS

Jacek Koronacki

Institute of Computer Science, Polish Academy of Sciences

Ph. D. Program 2013/2014

The project is co-financed by the European Union within the framework of European Social Fund

Semi-supervised learning: a "must link - cannot link" approach

One semi-supervised learning algorithm (proposed by Davidson and Ravi) for set S of observations:

1. Construct the transitive closure of the ML ("must link") constraints resulting in r connected components $M_1, M_2, \ldots, M_r$.
2. If two points $\{\mathbf{x}, \mathbf{y}\}$ are both a CL ("cannot link") and ML constraint, then output "No Solution" and stop.
3. Let $S_1 = S - (\cup_{i=1}^{r} M_i)$. Let $k_{\max} = r + |S_1|$.
4. Construct an initial feasible clustering with $k_{\max}$ clusters consisting of the r clusters $M_1, \ldots, M_r$ and a singleton cluster for each point in S_1. Set $t = k_{\max}$.
5. **while** (there exists a pair of mergeable clusters) **do**

    (a) Select a pair of clusters C_l and C_m according to the specified distance criterion.

    (b) Merge C_l into C_m and remove C_l. {The result is $Dendrogram_{t-1}$.}

    (c) $t = t - 1$.

endwhile

Transductive SVM

For a learning task $P(\mathbf{x}, y) = P(y|\mathbf{x})P(\mathbf{x})$ the learner L is given a hypothesis space H of functions $h : \mathbf{X} \to \{-1, 1\}$ and an i.i.d sample S_{train} of n training examples

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n).$$

The learner is also given an i.i.d. sample S_{test} of k test examples

$$\mathbf{x}_1^\star, \mathbf{x}_2^\star, \ldots, \mathbf{x}_n^\star$$

from the same distribution. L aims to select a function $h_L = L(S_{train}, S_{test})$ from H so that the expected number of erroneous predictions

$$R(L) = \int \frac{1}{k} \sum_{i=1}^{k} \Theta(h_L(\mathbf{x}_i^\star), y_i^\star) dP(\mathbf{x}_1, y_1) \cdots dp(\mathbf{x}_k^\star, y_k^\star)$$

on the test examples is minimized; $\Theta(a, b) = 0$ if $a = b$ and it is 1 otherwise.

Transductive SVM (TSVM)

Now, for a linear TSVM and confining ourselves to a linearly separable case, we get the optimization task:

Minimize over $y_1^\star, \ldots, y_k^\star, \mathbf{w}, b$

$$\frac{1}{2}\|\mathbf{w}\|^2$$

subject to:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geqslant 1, \quad i = 1, \ldots, n,$$

and

$$y_j^\star(\mathbf{w} \cdot \mathbf{x}_j^\star + b) \geqslant 1, \quad j = 1, \ldots, k.$$

In this way, TSVM performs transductive inference, i.e., unlike in inductive inference, unlabeled examples are also used for learning.

Laplacian SVM (LapSVM)

As of now, it seems an established fact that there are two common assumtions on marginal distributions of unlabeled (as well as labeled) data: the cluster assumption and the manifold assumption (in our short exposition we follow the paper by Melacci and Belkin, *Lalpacian Support Vector Machines Trained in the Primal*, JMLR 12 (2011)).

The 1st assumtion underlies, e.g., TSMVs, the 2nd lies behind many graph based methods (which, however, usually perform only transductive inference), while the LapSVMs provide a natural out-of-sample extension, so taht they can classify data that become available after the training process, without having to retrain the classifier.

Laplacian SVM (LapSVM)

Let $\mathcal{S} = \{x_i, i = 1, \ldots, n\}$ with $x_i \in X \subset R^m$ be the training examples, the first ℓ of them being labeled, with label $y_i \in \{-1, 1\}$, and the remaining u points being unlabeled ($n = \ell + u$ and $\mathcal{S} = \mathcal{L} \cup \mathcal{U}$).

Labeled examples are generated from the distribution P on $X \times R$, whereas unlabeled examples are drawn according to the marginal distribution P_X of P.

Let L be the graph Laplacian associated to $\mathcal{S}$, given by $L = D - W$, where W is the adjacency matrix of the data graph (e.g., with the exponential weights) and D is diagonal with the degree of each node ($d_{ii} = \sum_{j=1}^{n} w_{ij}$).

Laplacian SVM (LapSVM)

Let $K \in R^{n,n}$ be the Gram matrix associated to the n points of $\mathcal{S}$, where the i,j-th entry of the matrix is the evaluation of the kernel function $k(x_i, x_j)$, $k : X \rightarrow R$.

The unknown target function to be estimated from data is denoted as $f : X \rightarrow R$. In classification problem, the decision function then is $y(x) = g(f(x))$.

Laplacian SVM (LapSVM)

Within manifold regularization approach, P_X is assumed to have the structure of a Riemannian manifold $\mathcal{M}$. An intrinsic regularizer $\|f\|_I$, whose role is to enforce the solution to take into account the intrinsic geometry of P_X, is estimated using the graph Laplacian as

$$\|f\|_I^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(f(x_i) - f(x_j))^2 = f'Lf.$$

The rationale for this regularizer is that the labels of two points that are close in the intrinsic geometry of P_X (w.r.t. to geodesic distances on $\mathcal{M}$) should be the same or $P(y|x)$ should change little between two such points.

Laplacian SVM (LapSVM)

Given a kernel function $k(\cdot,\cdot)$, its associated RKHS $\mathcal{H}_k$ of functions $X \longrightarrow R$ with corresponding norm $\|\cdot\|_A$ (index A coming from "ambient space"), we estimate the target function by

$$f^\star = \arg\min_{f\in\mathcal{H}_k} \sum_{i=1}^{\ell} \max(1 - y_i f(x_i), 0) + \gamma_A \|f\|_A^2 + \gamma_I \|f\|_I^2.$$

While the ambient norm enforces smoothness of possible solutions, the intrinsic norm enforces smoothness along the sampled $\mathcal{M}$.

It has been shown that $f^\star$ admits the following expansion:

$$f^\star(x) = \sum_{i=1}^{n} \alpha_i^\star k(x_i, x) + b.$$

Laplacian SVM (LapSVM)

By introducing slack variables ξ_i, and given the form of $f^\star$, the minimization problem can be written as:

$$\min_{\alpha \in R^n, \xi \in R^\ell} \sum_{i=1}^{\ell} \xi_i + \gamma_A \alpha' K \alpha + \gamma_I \alpha' K L K \alpha$$

subject to

$$y_i \left(\sum_{j=1}^{n} \alpha_i k(x_i, x_j) + b \right) \geqslant 1 - \xi_i, \quad i = 1, \ldots . \ell$$

and

$$\xi_i \geqslant 0, \quad i = 1, \ldots, \ell.$$

Examples of applications of LapSVM will be presented during the lecture.

Small n large p problems: Introductory remarks

A major challenge in the analysis of many biological data matrices is due to their sizes: relatively small number of records (samples), often of the order of tens, versus thousands of attributes or features for each record.

An obvious example are microarray gene expression experiments (here, the features are genes or, more precisely, their expression levels). Another, and a very specific one, is that of analyzing molecular interaction networks underlying HIV-1 resistance to reverse transcriptase inhibitors (here, the features are some physicochemical properties of amino acids). In GWAS, while we have thousands observations, each consists of hundreds of thousands of features.

Small n large p problems: Introductory remarks

By far, it is not only in Life Sciences, where problems of this type appear and have to be dealt with.

Indeed, in our own work, we met fascinating problems of commercial origin, including transactional data from a major multinational FMCG (fast-moving consumer goods) company and geological data from oil wells operated by a major American oil company.

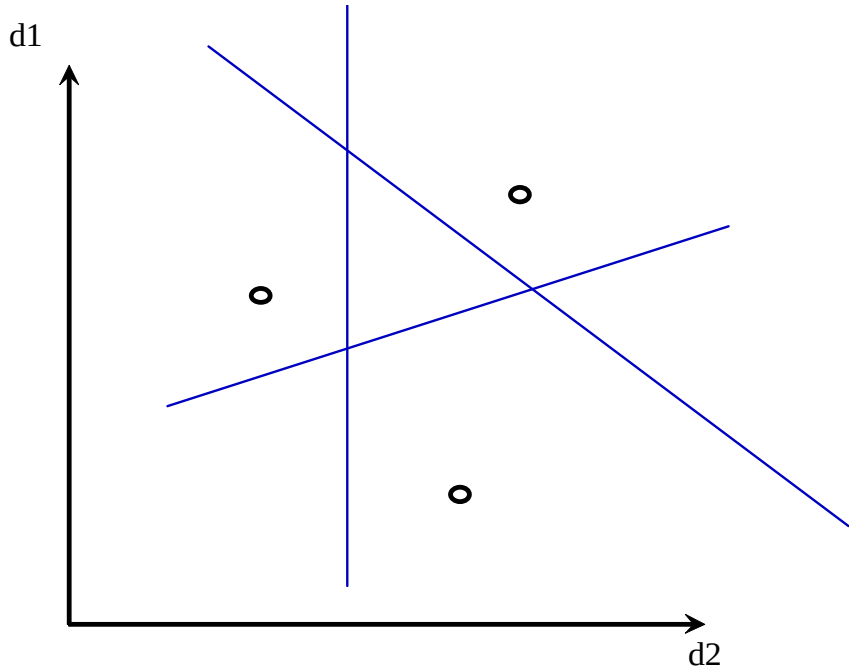
Small n large p problems: Introductory remarks

Such tasks, regardless of whether the data are to explain a quantitative (as in regression) or categorical (as in classification) trait, are quite different from typical data mining problems, in which the number of features is much smaller than the number of samples.

Indeed, in a sense, these are ill-posed problems. It is immediately clear in the case of linear regression fitted by least-squares.

Small n large p problems: Introductory remarks

For two-class classification, at least from the geometrical point of view, the task is trivial, since in a d-dimensional space, as many as $d + 1$ points can be divided into two arbitrary and disjoint subsets by some hyperplane, provided that these points do not lie in a proper subspace of the d-dimensional space.



Small n large p problems: Introductory remarks

It is another matter that the hyperplane (or any other classification rule) found should have the generalization ability.

In any case, whether in classification or in regression, since it is rather a rule than an exception that most features in the data are not informative, it is of utmost importance to select the few ones that are informative and that may form the basis for class prediction or building a proper regression model.

That is, before building a classifier or a regression model, or while building any of them, we would like to find out which features are specifically linked to the problem at hand and should be included in the solution.

Small n large p problems: Introductory remarks

Mathematically, properly formulated sparsity constraints should be included when seeking a solution. As we shall see, this requirement can be fulfilled by randomization or regularization.

Regarding classification one more important issue should be emphasized:

More often than not, rather than obtaining the best possible classifier, the Life Scientist needs to know which features contribute best to classifying observations (samples) into distinct classes and what are the interdependencies between the features which describe the observation.

Multiple hypothesis testing

Univariate approach based on multiple hypothesis testing: while disregarding interactions between features, it is statistically sound and all to well illustrates the intricacy of the problem:

Assume a two-class classification case. For each k-th feature we are interested in testing the null hypothesis H_{0k} of no relationship between the decision attribute (class) and the feature against the alternative that such a relationship does exist.

For each k-th feature, $k = 1, \ldots, d$, a natural test statistic is a t-statistic

$$\frac{\bar{x}_{1k} - \bar{x}_{2k}}{s_{1k} + s_{2k}}$$

although examined without assuming normal distribution of the feature.

A real catch is that we have to perform not one but d such tests!

Multiple hypothesis testing

The battery of tests should have a fixed level of the probability of type one error, e.g.,

$$\mathrm{FWER} \equiv \text{family-wise error rate} = P(FP \geqslant 1) \leqslant \alpha$$

where FP stands for the number of false positives (i.e., type I errors)

or

$$\mathrm{FDR} \equiv \text{false discovery rate} = E(FP/(FP + TP)) \leqslant \alpha$$

as well as a reasonable power of the whole procedure, e.g.,

$$P(TP \geqslant 1)$$

where TP stands for the number of true positives.

Multiple hypothesis testing

Bonferroni's (1936) classical procedure, under which any null hypothesis is rejected at level α/d, controls the FWER,

$$\mathrm{FWER} \equiv \text{family-wise error rate} = P(FP \geqslant 1) \leqslant \alpha,$$

for arbitrary test statistics joint null distributions; that is,

$$P(FP \geqslant 1) \leqslant \sum_{i \in \mathcal{H}_0} P_{H_{0i}}(\text{i-th test rejects}) \leqslant \frac{h}{d}\alpha \leqslant \alpha,$$

where $\mathcal{H}_0$ runs over the indices corresponding to true null hypotheses and $h = |\mathcal{H}|$.

(Under independence of test statistics and complete null hypothesis,

$$\mathrm{FWER} = 1 - (1 - \alpha/d)^d;$$

the FWER is smaller, if they are positively dependent.)

Multiple hypothesis testing

Note that under the Bonferroni procedure any null hypothesis is rejected regardless of the values of test statistics for other hypotheses.

A more sophisticated procedure of Benjamini and Hochberg (1995; see the next slide) controls the FDR,

$$\mathrm{FDR} \equiv \text{false discovery rate} = E(FP/(FP + TP)) \leqslant \alpha,$$

for independent test statistics (or, more generally, for positively regression dependent test statistics).

Multiple hypothesis testing

The Banjamini and Hochberg procedure:

1. Let

$$p_{(1)} \leqslant p_{(2)} \leqslant \cdots \leqslant p_{(d)}$$

denote the observed ordered p-values

2.

$$L = \max\{j : \ p_{(j)} < \alpha \cdot \frac{j}{d}\}$$

3. Reject all hypotheses H_{0j}, such that $p_{(j)} \leqslant p_{(L)}$.

Thus, the p-values must be obtained, but this can be done by a simple resampling procedure.

For this section see S. Dudoit and M. J. van der Laan, *Multiple Testing Procedures with Applications to Genomics*, Springer 2011.

Multiple hypothesis testing - a word on ANOVA

Gene expression microarray technologies, which have been developed since the end of the 20-th century, call for suitable experimental designs for the ANOVA models, such as, e.g.,

$$\log(y_{ijkg}) = \mu + A_i + D_j + G_g + V_k + (AG)_{ig} + (DG)_{jg} + (VG)_{kg} + \varepsilon_{ijkg},$$

where μ represents the general mean, A_i is the array effect for i-th array, D_j is the affect for j-th dye, V_k is the effect for k-th variety, G_g is the effect for g-th gene, while $(AG)_{ig}$, $(DG)_{jg}$ and $(VG)_{kg}$ account for first order interactions.

See, e.g., Kerr, Churchill, Cui and Martin (2000), (2001a), (2001b), (2003).

Broman and Speed (2002): Let

$$y_i = \mu + \sum_{j=1}^{d} \beta_j x_{ij} + \varepsilon_i,$$

where $x_{ij} = 1$ or $x_{ij} = 0$ and the ε_i are i.i.d. and normally distributed, $N(0, \sigma^2)$ (in fact, x_{ij} represents genotype at marker j for individual i). The task is to select a model for which Schwarz's Bayesian Information Criterion (BIC) assumes the minimal value;

$$BIC = n \cdot \log RSS(\beta) + \frac{1}{2} k \log n,$$

where k is the number of parameters β_j in the model. It was observed by Broman i Speed that the BIC tends to overestimate the number of parameters in the model. Accordingly, they proposed the 1st modification of the BIC.

Model selection for linear regression - Bayesian approaches

The Bayesian model selection advocates choosing the model M that maximizes posterior probability of the model given the data, this probability being proportional to

$$L(y|M)\pi(M),$$

where $\pi(M)$ is a prior probability for model M (Schwartz assumed noninformative uniform prior π), and

$$L(y|M) = \int L(y|M,\beta)f(\beta|M)d\beta,$$

$f(\beta|M)$ being some prior distribution on the vector of model parameters; for a wide class of these distributions one gets

$$\log L(y|M) = \log L(y|\beta) - \frac{1}{2}(k+2)\log n.$$

For the family of normal linear regression models, maximization of this last expression is equivalent to minimization of the BIC.

Model selection for linear regression - Bayesian approaches

Bogdan et al. (2004) introduced another modification of BIC (mBIC), assuming binomial prior distribution, $\mathrm{Bin}(d, c/d)$, with some fixed c, for the model size.

Later (2008), it was shown that if $k_n/d_n \to 0$ and $d_n/\sqrt{n} \to \mathrm{const} \in (0, \infty]$, as $n \to \infty$, then the expected number of false positives (false regressors) detected by BIC may go to infinity.

Minimizing the mBIC was also shown to be closely connected to following the Bonferroni procedure and controlling the FWER. Still later (2011), for another modification of the BIC, mBIC2, it was shown that its minimization is tied to the Benjamini and Hochberg procedure and controlling the FDR. Finally, weak consistency of the mBIC and mBIC2 procedures have been proved under appropriate conditions.

Model selection for linear regression - Bayesian approaches

It is easy to extend the outlined approach to include regression models with interactions.

It is also possible to extend it to include generalized linear models (possibly with constraints on the model's parameters).

The outlined approach is by far not the only one possible among Bayesian approaches; e.g., a similar approach is that based on the extended BIC, and a completely different approach, which bears some relationship with support vector machines, is that of relevance vector machines. (See, e.g., Chen et al. (2008) and (2011), and Tipping (2001), Fletcher (2010) and Saarela et al. (2010).)

Model selection for linear regression - ℓ_1 regularization

Back to the The Lasso (Least Absolute Selection Operator) for linear models:

As usual, we are given n observations, each with d explanatory variables (predictors), $(x_{i1}, x_{i2}, \ldots, x_{id})$, and one response variable, y_i,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_d x_{i,d} + \varepsilon_i, \quad i = 1, 2, \ldots, n,$$

where ε_i are i.i.d. random errors with mean 0 and unknown variance σ^2, and $\beta_0, \ldots, \beta_d$ are unknown parameters.

Minimize

$$\{\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2\}$$

subject to

$$\sum_{j=1}^{p} |\beta_j| \leqslant t.$$

Model selection for linear regression - ℓ_1 regularization, and more

The Lasso, in contrast to ridge regression (i.e., ℓ_2 regularization), eliminates for small t some variables from the model. It can thus be used as a feature selection method.

For exhaustive account of the Lasso and related approaches see Peter Bühlmann and Sara van de Geer, *Statistics for High-Dimensional Data*, Springer, 2011. See there also for a different approach which stems from undirected graphical modeling and is based on inferring zero partial correlations for variable selection (the so-called PC-simple algorithm).

A still another and promising approach, which builds on ranking the marginal correlations and is referred to as sure independence screening, has been introduced by Fan and Lv (2008); see also Fan and Song (2010).

Model selection for linear regression - Random Subspace Method (RSM)

Mielniczuk and Teisseyre (2011) and (2013): Let $T_{i,m}$ be a t-statistic for i-th predictor in a linear regressionn model m with $|m|$ predictors. We have:

$$\frac{T_{i,m}^2}{n-|m|} = \frac{\mathrm{RSS}_{m-\{i\}} - \mathrm{RSS}_m}{\mathrm{RSS}_m}$$

It follows that the value of $T_{i,m}^2$ can serve as a measure of, simulatneously, the importance of the i-th predictor in model m and the quality of this very model.

Model selection for linear regression - Random Subspace Method (RSM)

In the RSM, a random subset m of features (predictors), of size $|m|$ smaller than the number of all features d and a number of observations n, is chosen. The model is fitted in the reduced feature space by OLS. Each of the selected features is assigned a weight describing its relevance in the considered submodel.

Random selection of features is repeated many times, corresponding submodels are fitted and the final weights (scores) of all d features are computed on the basis of all submodels.

The final model can then be constructed based on predetermined number of the most significant predictors or using a selection method applied to the nested list of models given by the ordering of predictors.

MCFS-ID Algorithm of Draminski et al.: the Monte Carlo Feature Selection (or MCFS) part

In what follows we begin with a brief description of an effective method for ranking features according to their importance for classification regardless of a classifier to be later used. Our procedure is conceptually very simple, albeit computer-intensive.

We consider a particular feature to be important, or informative, if it is likely to take part in the process of classifying samples into classes "more often than not".

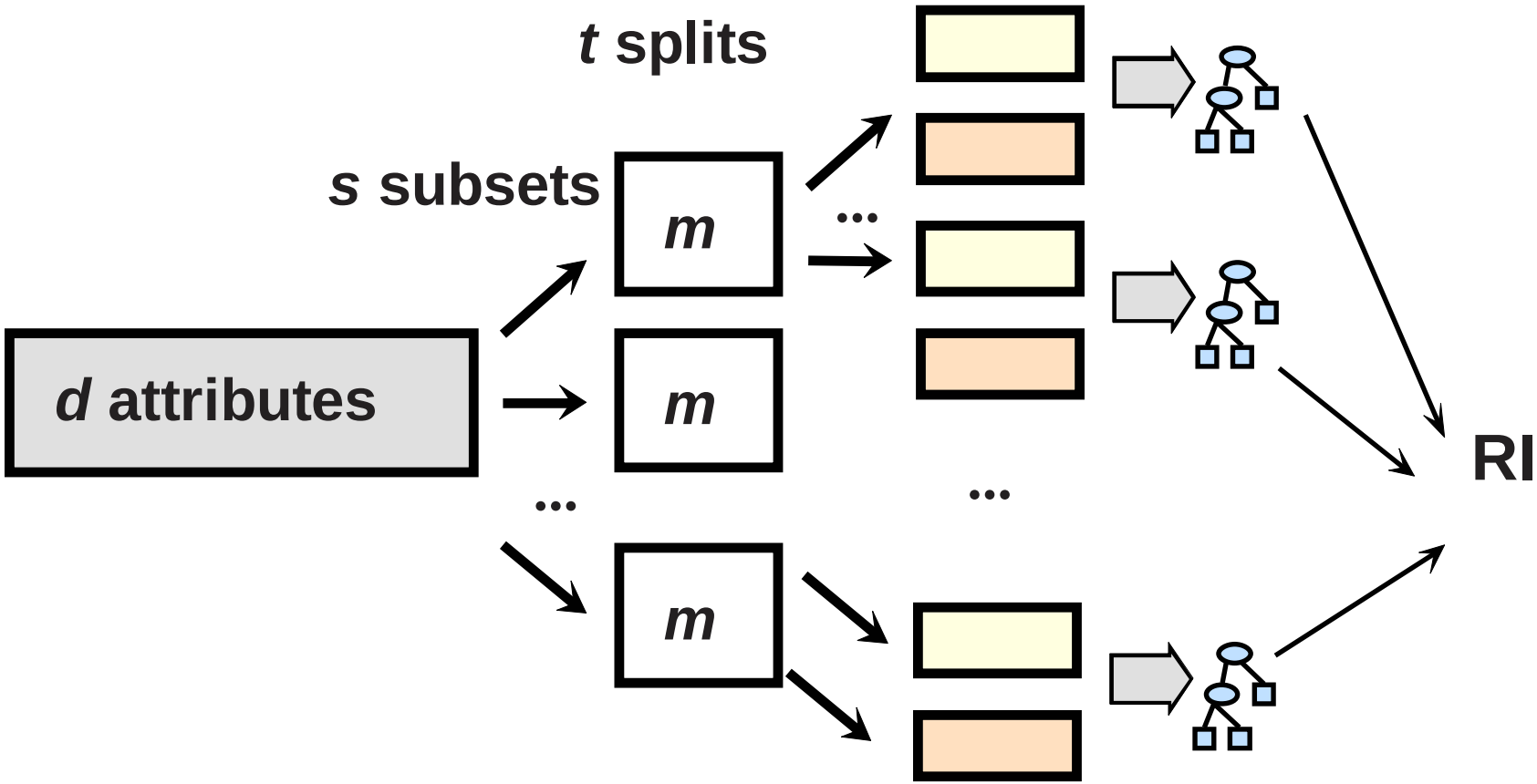
This "readiness" of a feature to take part in the classification process, termed relative importance of a feature, is measured via intensive use of classification trees.

In the main step of the procedure, we estimate relative importance of features by constructing thousands of trees for randomly selected subsets of features.

More precisely, out of all d features, s subsets of m features are selected, m being fixed and $m << d$, and for each subset of features, t trees are constructed and their performance is assessed.
Each of the t trees in the inner loop is trained and evaluated on a different, randomly selected training and test sets which come from a split of the full set of training data into two subsets: each time, out of all n samples, about 66% of samples are drawn at random for training (in such a way as to preserve proportions of classes from the full set of training data) and the remaining samples are used for testing.

MCFS-ID Algorithm: the MCFS part

The main step of the procedure



The relative importance of feature g_k, RI_{g_k}, can be defined as

$$\mathrm{RI}_{g_k} = \sum_{\tau=1}^{st} (wAcc)_\tau^u \sum_{n_{g_k}(\tau)} \mathrm{IG}(n_{g_k}(\tau)) \left(\frac{\text{no. in } n_{g_k}(\tau)}{\text{no. in } \tau} \right)^v ,$$

where summation is over all st trees and, within each τ-th tree, over all nodes $n_{g_k}(\tau)$ of that tree on which the split is made on feature g_k, $\mathrm{IG}(n_{g_k}(\tau))$ stands for information gain for node $n_{g_k}(\tau)$, $(\text{no. in } n_{g_k}(\tau))$ denotes the number of samples in node $n_{g_k}(\tau)$, $(\text{no. in } \tau)$ denotes the number of samples in the root of the τ-th tree, and u and v are fixed positive reals (now set to 1 by default).

Ranking as such does not enable one to discern between informative and non-informative features. A cut-off between these two types of features is needed.

We address this issue by comparing the ranking obtained for the original data with the one obtained for the data modified in such a way that the class attribute (label) becomes independent of the vector of all features. Such a data set is obtained via a random permutation of the values of the class attribute (i.e. of the class labels of the samples).

MCFS-ID Algorithm: the MCFS part
Validation and confirmatory steps

A thorough statistical validation of the results is of course a must.

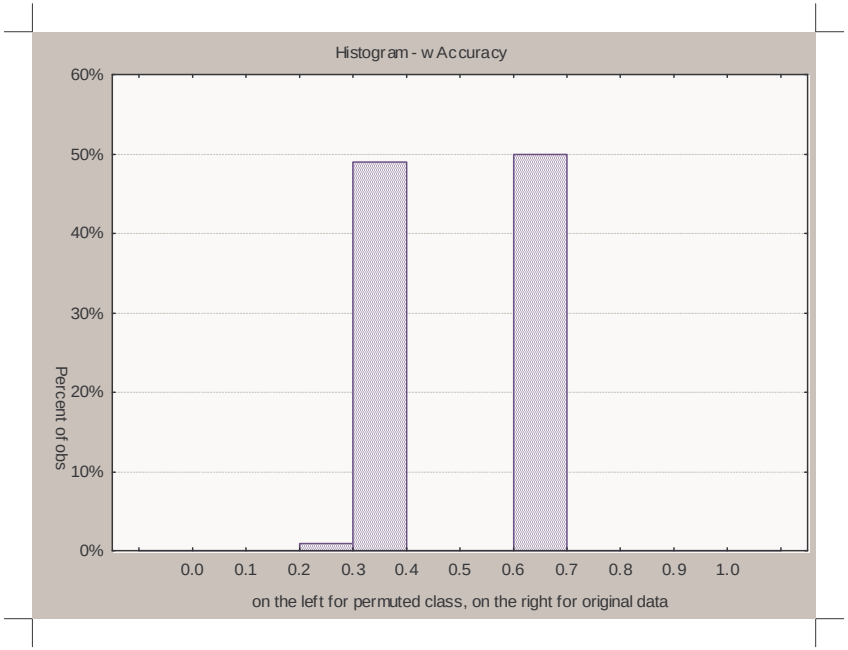
- First, we verify that the data are informative
- Second, we verify that the features found as most informative are such indeed
- Third, statistical significance of the results is confirmed.

- Moreover, independently of all the former considerations, we have also shown that, despite its simplicity and the use of tree classifiers, the algorithm is not biased towards features with many values (categories or levels).
- And finally, in our new implementations of the MCFS, other flexible rule-based classifiers have been used.

A digression on validation and confirmatory steps

For a given data set, the first validation step consists in repeating the main step of the procedure with, say, 50 different permutations of class labels of the samples.
The aim is to show that the classification results obtained earlier measured by the distribution of *wAcc* on all *st* trees are significantly different from what can be obtained under randomly permuting the labels (classes) of samples, hence making the class independent of feature values. It is thus a way to confirm that the data are informative.
This fact justifies the search for the most important features, based on the data provided.

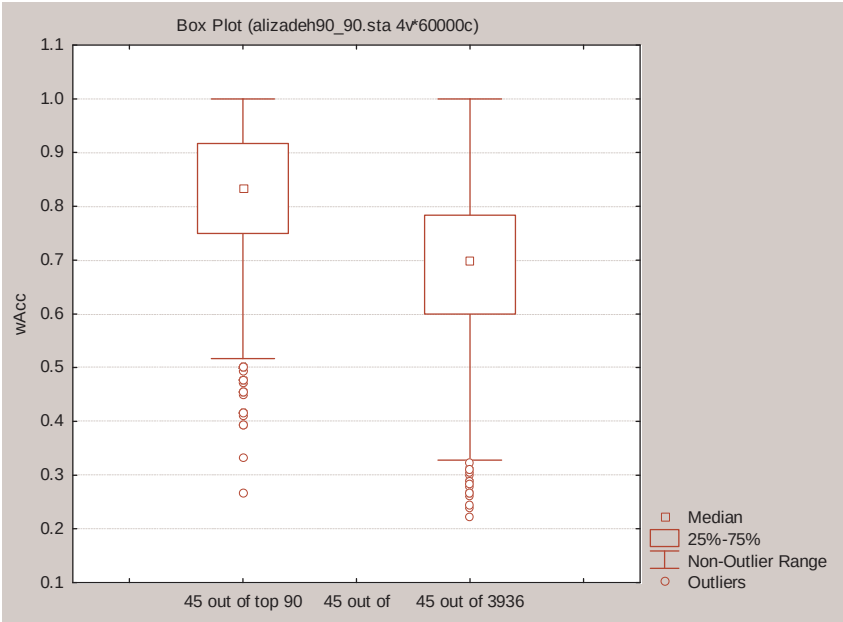
Validation and confirmatory steps
First validation step (lymphoma data of Alizadeh et al.)



Validation and confirmatory steps

The second validation step consists in showing that reliable class prediction can be performed using only a few, say b, randomly chosen features out of $2b$ features earlier found to be relatively most important.
This is done by again constructing thousands of trees on b features out of the $2b$ most important features, as well as on randomly selected sets of b features from the set of the remaining $d - 2b$ features.
For each set of b features many training and testing sets of samples are drawn at random from the original set of samples and trees are trained and tested on these sets.
As a result, two distributions of *wAcc* are obtained, one for b features from the $2b$ most important ones and another for b features chosen at random from all the remaining ones, with the goal to prove significance of classification results for the best features.

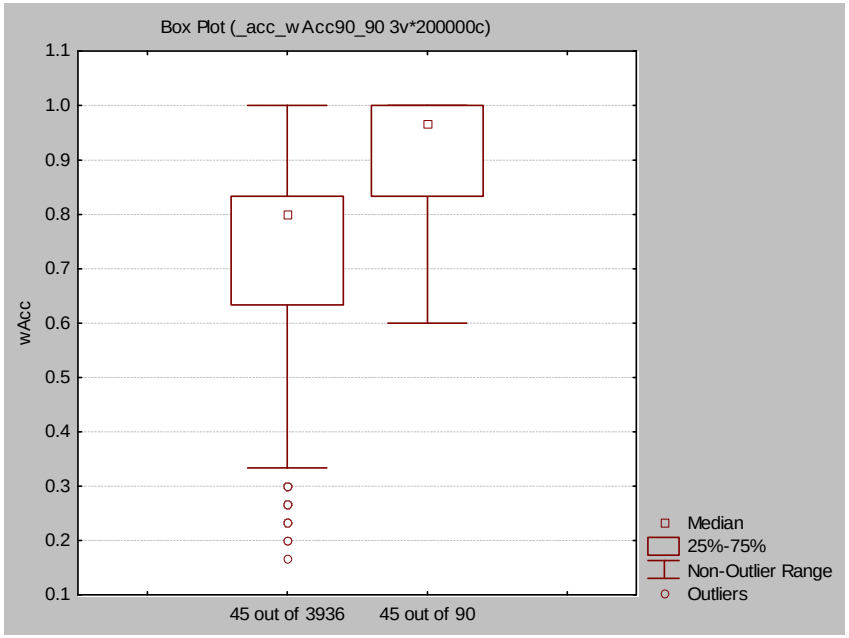
Validation and confirmatory steps
Second validation step (lymphoma data of Alizadeh et al.)



Validation and confirmatory steps

At the same time, however, our validation does not provide anything like an "exact level of significance" of the results obtained.
Extensive resampling introduces intrinsic interdependencies within the whole procedure, making results conditional on the data. Hence, we propose one additional confirmatory step in the procedure.
It consists in splitting first the data set into two subsets, comprised of about 75% and 25% of the whole data, respectively. The first subset is termed the final validation set and the latter – the final test set. Then, the main step of the procedure is run on the final validation set, and the earlier described second validation step is run with *wAcc*'s calculated on the basis of the final test set (not used in the main step in any way).

Validation and confirmatory steps
Confirmatory step (lymphoma data of Alizadeh et al.)



A word on experiments and results

Just one comment, stemming from a revealing comparison with the seminal work of Dudoit et al. (2002, 2003) and their analysis of the leukemia data of Golub et al. (1999):

No doubt, over-expression is **not** needed for a gene to contribute highly to classification. Most interestingly, our method has proved **capable of selecting features that are germane to the origins of the genesis of leukemia**.

Moreover, our method is **capable of exploiting interactions between features and hence finding groups of features which together contribute to classification**.

Interdependency Discovery, i.e., the ID part of the MCFS-ID Algorithm

Our approach to interdependency discovery is significantly different from known approaches which consist in finding correlations between features or finding groups of features that behave similarly in some sense across samples (e.g., as in finding co-regulated features).

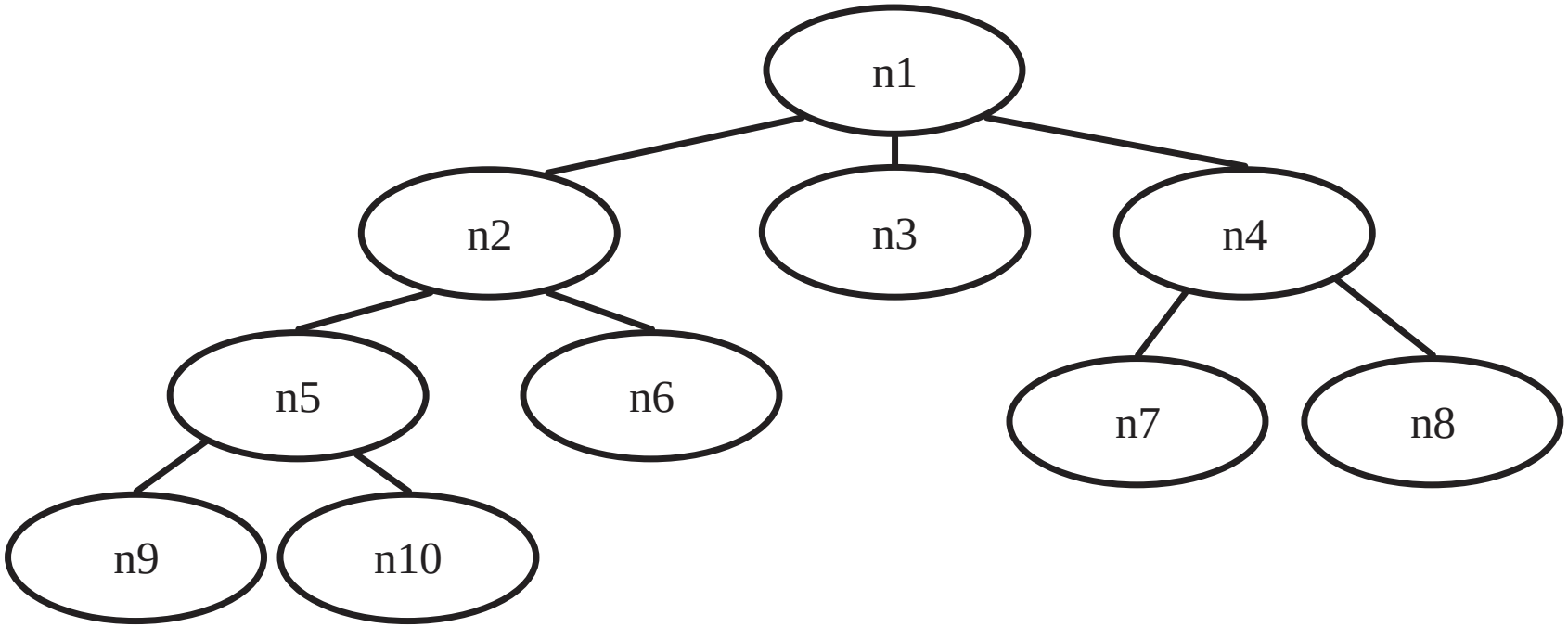
In our approach, we focus on identifying features that "cooperate" in determining that a sample belongs to a particular class.

Interdependency discovery: the ID part of MCFS-ID Algorithm

To be more specific, assume that for a given training set of samples, an ensemble of flexible rule-based classifiers has been constructed, where flexibility amounts to classifier's ability to produce rules as complex as is needed. Assume also that each of the decision rules provided by the classifiers has the form of a conjunction of conditions imposed on particular separate features.

Clearly then, each decision rule points to some interdependencies between the features appearing in the conditions. Indeed, the information included in such decision rules, when properly aggregated, reveals interdependencies (however complex they may prove) between features which are best "correlated" with or, as has been said, "cooperate" in determining, the samples' classes.

Interdependency discovery: strength of interdependence



$$\mathrm{Dep}(g_i, g_j) = \sum_{\tau=1}^{st} \sum_{\xi_\tau} \sum_{n_{g_i}(\xi_\tau), n_{g_j}(\xi_\tau)} \frac{1}{dist(n_{g_i}(\xi_\tau), n_{g_j}(\xi_\tau))},$$

A bit more on results

By way of example: HIV-1 resistance to seven RTI drugs

We applied the MCFS-ID to elucidate molecular interaction networks that underly resistance to seven reverse transcriptase inhibitors.

We were able to rediscover numerous mechanisms of drug-resistance and suggest several new mechanisms which should be further investigated. Importantly, our method offers deep insight into molecular mechanisms of drug resistance, since it shows interactions between physicochemical properties of mutating amino acids.

We were also able to show that the majority of the drug resistance mutations act simultaneously, in a co-operative and complex manner. See Draminski et al. (2010) and Kierczak et al. (2009) and (2010).

Selective bibliography on small n large p problems:

- Bogdan M., Ghosh J.K., Doerge R.W.: Modifying the Schwarz Bayesian Information Criterion to locate multiple interacting Quantitatve Trait Loci. Genetics. 2004; 167, 989-999.
- Bogdan, M., Chakrabarti, A., Frommlet, F., Ghosh, J.K.: Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. Annals of Statistics. 2011; 39(3), 1551-1579.
- Broman K.W, Speed T.P, A model selection approach for the identification of quantitative trait loci in experimental crosses. J. Royal Statist. Soc. 2002; B 64, 641-656.
- Bühlmann P., van de Geer S., *Statistics for High-Dimensional Data*, Springer, 2011.
- Chen J., Chen Z., Extended Bayesian information criterion for model selection with large model space. 2008; Biometrika, 94, 759-771.
- Chen J., Chen Z., Extended BIC for linear regression models with diverging number of relevant features and high or ultra-high feature spaces. 2011; arXiv:1107.2502
- Cui X., Churchill G.A., Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 2003; 4(4), 210; doi: 10.1186/gb-2003-4-4-210.
- Draminski M, Rada Iglesias A, Enroth S, Wadelius C, Koronacki J, Komorowski J.: Monte Carlo feature selection for supervised classification. Bioinformatics. 2008; 24, 110-117.
- Draminski M, Kierczak M, Koronacki J, Komorowski J.: Monte Carlo feature selection and interdependency discovery in supervised classification. In: Advances in Machine Learning, vol. 2; Springer, 2010.
- Dudoit S., van der Laan M. J., *Multiple Testing Procedures with Applications to Genomics*, Springer, 2011.
- Fan J., Lv J., Sure independence screening for ultra-high dimensional feature space. J. Royal Statist. Soc. 2008; B 70 (5), 849-911.
- Fan. J., Song R., Sure independence screening for generalized linear models with np-dimensionality. Ann. Statist. 2010; 38(6), 3567–3604.
- Fletcher T., Relevance vector machines explained. 2010; Tech. report, www.cs.ucl.ac.uk/staff/T.Fletcher
- Frommlet F., Ruhaltinger F., Twaróg P., Bogdan M.: Modified versions of the Bayesian Information Criterion for genome-wide association studies Computational Statist. and Data Anal. 2012; 56(5), 1038-1051.

Selective bibliography, contd.:

- Kerr M.K., Martin M., Churchill G.A., Analysis of variance for gene expression microarray data. J. Comput. Biology. 2000; 7:819-837.
- Kerr M.K., Churchill G.A., Experimental design for gene expression microarrays. Biostatistics. 2001; 2, 2, 183-201.
- Kerr M.K., Churchill G.A., Statistical design and the analysis of gene expression microarray data. Genet. Res., Camb. 2001; 77, 123-128.
- Kierczak M, Ginalski K, Draminski M, Koronacki J, Rudnicki W, Komorowski J. A rough set-based model of HIV-1 RT Resistome. Bioinformatics a. Biology Insights. 2009:3 109–127.
- Kierczak M, Draminski M, Koronacki J, Komorowski J. Computational analysis of local molecular interaction networks underlying change of HIV-1 resistance to selected reverse transcriptase inhibitors. Bioinformatics a. Biology Insights. 2010; 4: 137–146.
- Mielniczuk J., Teisseyre P.: Using Random Subset Method for prediction and variable importance assessment in linear regression. Computational Statist. and Data Anal. 2012; in press.
- Mielniczuk J., Teisseyre P.: Selection and Prediction for Linear Models using Random Subspace Methods.
- Saarela M., Elomaa T., Ruohonen K., An Analysis of Relevance Vector Machine Regression. In: Advances in Machine Learning, vol. 2; Springer, 2010.
- Tipping M.E., Sparse Bayesian learning and the relevance vector machine. 2001; Journal of Machine Learning Research 1, 211-244.

Remark on an instructive experiment of Simon *et al.* (2003):

Draw at random 20 vectors, each of size 6 000, from the standard normal distribution. Split the 20 vectors randomly into two classes, say, class 1 and 2, each of size 10, and consider them as your training set. Replicate this experiment 2 000 times.

For each training set build a classifier and evaluate its performance in the following three ways:

- Procedure 1: Select informative features, and use the 20 subvectors of the features selected to build a classifier; evaluate the classifier's accuracy by a substitution method.
- Procedure 2: Select informative features, and use the 20 subvectors of the features selected to build a classifier using the leave-one-out cross-validation (CV) to assess its accuracy.
- Procedure 3: Use leave-one-out CV in such a way that feature selection be made and classifier's accuracy be evaluated inside the CV loop.

An instructive experiment of Simon *et al.* (2003), contd.

The following, hardly surprising results were obtained:

- Procedure 1: 98.2% replications without a single misclassification (1.8% with just 1)!
- Procedure 2: 90.2% replications without a single misclassification (9.8% with just 1)!
- Procedure 3: From 0 to 20 misclassifications, with their median equal to 11.