Model assessment

Gilbert Saporta
Ndèye Niang
Chaire de Statistique Appliquée & CEDRIC, CNAM, 292 rue Saint Martin,
F-75003 Paris, e-mail : saporta@cnam.fr, niang@cnam.fr

In data mining and machine learning, models come from data and provide insights for understanding data (unsupervised classification) or making prediction (supervised learning) (Giudici, 2003, Hand, 2000).

Thus the scientific status of this kind of models is different from the classical view where a model is a simplified representation of reality provided by an expert of the field. In most data mining applications a good model is a model which not only fits the data but gives good predictions, even if it is not interpretable (Vapnik, 2006).

In this context, model validation and model choice need specific indices and approaches. Penalized likelihood measures (*AIC*, *BIC* etc.) may not be pertinent when there is no simple distributional assumption on the data and (or) for models like regularized regression, SVM and many others where parameters are constrained.
Complexity measures like the VC-dimension are more adapted, but very difficult to estimate.

In supervised classification, ROC curves and AUC are commonly used (Saporta & Niang, 2006). Comparing models should be done on validation (hold-out) sets but resampling is necessary in order to get confidence intervals.

1. Data and models

1.1 Traditional modelling

In « classical » statistical modelling, a model is a simplified representation of the real world of the form:

Data = Model + Error

where the "Model" part of this equation represents relationships between variables. There are many kinds of models, most of them being explanatory or supervised according to machine learning vocabulary.

The aim of a model is to give a good fit to observed data, in the sense where the error term may be considered as a white noise with a minimal variance.

Usually models come from a theory (biology, economics, physics etc.) and the role of statistics consists in:

a. estimating model parameters usually by Maximum Likelihood (ML) which has superseded other techniques like moments , minimum chi-square
b. checking if data are in agreement with the model (and vice-versa)

One can observe that model checking is frequently omitted in too many publications and that models are used to assess the influence of variables (risk factors) on a response rather than to predict individual behaviours. This may be in contradiction with the scientific exigence of having falsifiable models.

1.2 Model choice

When one has a nested family of parametric models with an unknown parameter θ, model choice based on penalized likelihood has given raise to a large literature. The two best known criteria being AIC and BIC:

$$AIC = -2\ln\left(L(\hat{\theta})\right) + 2k \text{ and } BIC = -2\ln\left(L(\hat{\theta})\right) + \ln(n)k$$

where k is the number of parameters and $\hat{\theta}$ the ML estimate of θ.
Despite their similarities *AIC* and *BIC* come from completely different theories.

AIC comes from Kullback-Leibler (KL) divergence. Let f and g be two probability density functions. If f is the true one and g an approximation the KL divergence or loss of information is:

$$I(f;g) = \int f(t)\ln\frac{f(t)}{g(t)}dt = \int \ln(f(t))f(t)dt - \int \ln(g(t))f(t)dt = E_f(\ln(f(t)) - E_f(\ln(g(t))$$

If we have to choose the best g (or θ) among a parametric family g(x:θ) one should maximize $E_f(\ln(g(t;\theta))$ but the expectation is calculated with respect to the true distribution which is unknown. The ML solution consists in using the parameter value maximising the density of the data according to g instead of f

Taking the expectation with respect to data and f we have to maximize $E_{\hat{\theta}}E_f(\ln(g(t;\hat{\theta}))$ which under regularity assumptions is asymptotically such that:

$$E_{\hat{\theta}}E_f(\ln(g(t;\hat{\theta})) \sim \ln(L(\hat{\theta}))-k$$

BIC comes from a completely different context : bayesian model choice. Let a finite family of parametric models M_i with priors *P(M* $P(M_i)$ and conditional priors for θ_i for each model $P(\theta_i / M_i)$. Then the posterior probability of M_i knowing the data **x** is proportional to $P(M_i)\, P(\mathbf{x}/M_i)$

With uniform priors $P(M_i)$, $P(\mathbf{x}/M_i) = \int P(\mathbf{x}/M_i;\theta_i)P(\theta_i/M_i)d\theta_i$. One has $\ln(P(\mathbf{x}/M_i) \sim \ln(P(\mathbf{x}/\hat{\theta}_i, M_i)-\frac{k}{2}\ln(n)$. The most probable model M_i *a posteriori* is the one with minimal *BIC*. Then the posteriors are $P(M_i/\mathbf{x})=\frac{e^{-0.5BIC_i}}{\sum_{j=1}^{m}e^{-0.5BIC_j}}$

It is well known that *BIC* favourises more parsimonious models than *AIC* due to its penalization and that *AIC* (not *BIC*) is biased in the following sense: if the true model belongs to the family M_i , the probability that *AIC* chooses the true model does not tend to one when the number of observations goes to infinity.

Other penalisations have been proposed such as $AIC3=-2\ln\left(L(\hat{\theta})\right)+3k$ but only from an empirical way, without any theoretical basis.

However the use of penalised likelihood is not a *panacea* since it only addresses a narrow set of models. Either the likelihood or the true number of parameters are not computable for many “modern” modelling techniques: decision trees, neural networks, ridge and PLS regression etc.

Furthermore, even for classical models one may have reasonable doubts about the convergence property towards the true model for BIC, since the “true” model generally does not exist, moreover if the number of observations tends to infinity. Let us remind that a model is a simplification of the real world helping the scientist to think and that as George Box said “All models are wrong. Some are useful”.

2. Models for prediction

In more and more applications (CRM, credit scoring etc.) models are used to make predictions. Thus the efficiency of a model should be measured by its capacity to make good predictions and not only to fit to the data (backforecasting is easier than forecasting).

2.1 The bias-variance trade-off (Hastie & al. 2001)

Let us consider a model like $y = f(\mathbf{x}) + \varepsilon$. f is estimated by $\hat{f}$ and we want to predict a new value y_0 of y for $\mathbf{x}_0$. The prediction error $y_0-\hat{y}_0 = f(x_0)+\varepsilon-\hat{f}(x_0)$ is

« twice » random : first, ε is not deterministic and second : the prediction $\hat{y}_0 = \hat{f}\,(\mathbf{x}_0)$ is random due to the use of a random sample of observations. The expected square error is:

$$E\left(y_0 - \hat{y}_0\right)^2 = \sigma^2 + E\left(f(x_0) - \hat{f}(x_0)\right)^2 = \sigma^2 + \left(E\left(\hat{f}(x_0)\right) - f(x_0)\right)^2 + V\left(\hat{f}(x_0)\right)$$

the first term is inherent to the phenomenon and cannot be reduced, the second term is the square bias of the model and the third is the prediction variance.

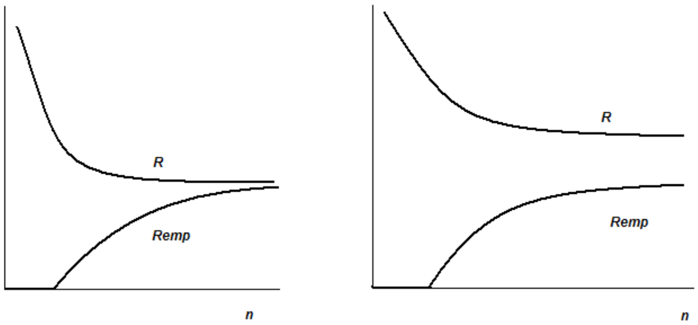
The more complex a model is, the lower is the bias but with a high variance. Thus there exist an optimal choice realizing a trade-off between bias (or goodness of fit to the observed data) and the prediction variance. But how can we measure the complexity of a model?

2.2 Statistical learning theory and the VC-dimension

V.Vapnik has shown that some models may not “generalize” in the following sense: for a prediction model let $L(y;\hat{y})$ be a loss function like $L(y;\hat{y}) = (y-\hat{y})^2$ in regression or in a binary classification problem where y and $\hat{y}$ take their values in {-1 ;+1}:

$$L(y;\hat{y}) = \frac{1}{2}\left|y-\hat{y}\right| = \frac{1}{2}\left(y-\hat{y}\right)^2$$

The risk is the expected loss $R = E(L) = \int L(z,\theta)dP(z)$ where $P(z)$ is the joint distribution of y and $\mathbf{x}$. The optimal parameter $\hat{\theta}$ should minimize R but it is an impossible task since $P(z)$ is unknown. The usual solution (least squares eg) consists in minimizing the empirical risk $R_{emp} = \frac{1}{n}\sum_{i=1}^{n} L(y_i;\, f(x_i;\theta))$ on a learning sample drawn from $P(z)$. With other definitions of L, one obtains the ML estimator or the Huber’s one etc. R_{emp} is a random variable. A model is consistent if R_{emp} converges towards R when n tends to infinity (figure 1).



consistent non consistent

Figure 1

A necessary and sufficient condition for consistency is that the Vapnik-Cervonenkis (VC) dimension should be finite. In binary supervised classification the VC-dimension h is a measure of complexity related to the separating capacity of a family of classifiers. h is the maximum number of points which can be separated by the family of functions whatever are their labels ± 1.

This does not mean that any configuration of h points might be « shattered » (one cannot for instance separate 3 points on the same line with a linear classifier in the plane), but for h+1 point there always exist a non-separable configuration.

The VC-dimension of unconstrained hyperplanes of $\mathbb{R}^p$ is p+1, but the VC dimension of a model is not identical to its number of parameters: it can be more or less.

In $\mathbb{R}$, the VC-dimension of f defined by $f(x)=1$ if $\sin(\theta x)>0$ and $f(x)=-1$ if $\sin(\theta x)<0$ is infinite since by increasing the unique parameter θ, one may separate an arbitrary number of points (see figure 2).

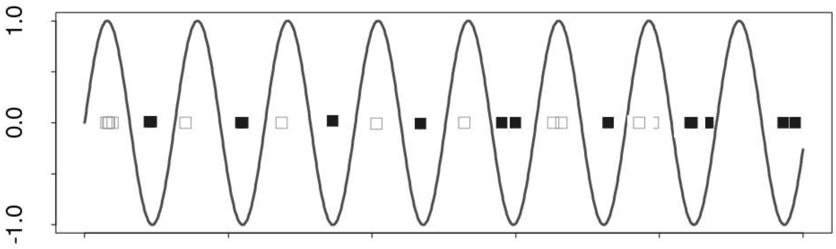


Figure 2

For constrained hyperplanes (ridge-regression) $f(X,w) = sign\left(\sum_{i=1}^{p}\left(w_i x_i\right)+1\right)$

where $\|\mathbf{w}\|^2 = \sum_{i=1}^{p} w_i^2 \le \frac{1}{C}$, the VC dimension may be far lower than p+1:

$$h \le \min\left[ent\left(\frac{R^2}{C^2}\right); p\right]+1$$

where R is the radius of the set of learning points.

2.3 Model choice by Structural Risk Minimization (SRM)

Vapnik's inequality relates the difference between R and R_{emp} to the VC-dimension h :

$$R < R_{\text{emp}} + \sqrt{\frac{h\left(\ln\left(2n/h\right)+1\right)-\ln\left(\alpha/4\right)}{n}}$$

where 1-α is the confidence level. This inequality proves that (provided h is finite) one may increase the complexity of a family of models (eg increase the degree of polynomials) when the number of learning cases increases, since it is the ratio h/n that is of interest.

Small values of h gives a low difference between R and R_{emp} . It explains why regularized (ridge) regression, as well as dimension reduction techniques, may provide better results in generalisation than ordinary least squares.

Based on the upper bound of R, SRM provides a model choice technique different from penalized likelihood, since no distributional assumptions are necessary .

Given a nested family of models, the principle is (for fixed n) to choose the model which minimizes the upper bound : this realizes a trade-off between the fit and the generalization capacity (see figure 3). Devroye & al (1996) and Vapnik (2006) have proved that for any distribution , the SRM provides the best solution with probability 1 (SRM is universally strongly consistent).

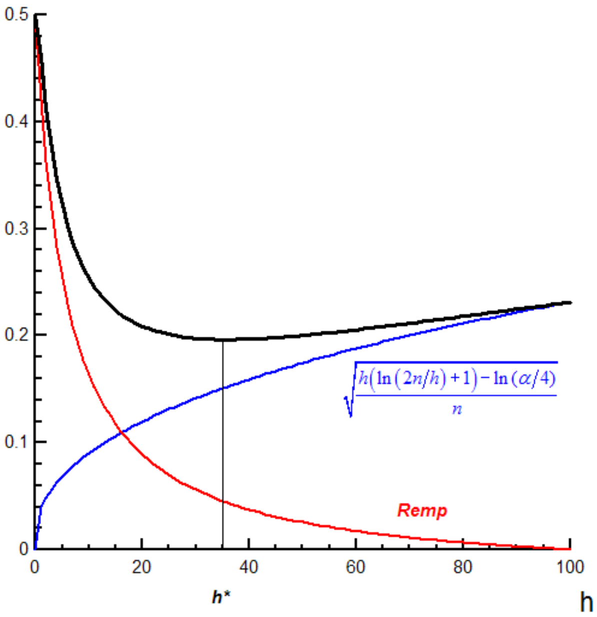


Figure 3

Since this is an universal inequality, the upper bound may be too large.

An other drawback is that the VC-dimension is very difficult to compute, and in most cases, one only knows upper bounds for *h*.

3. Empirical model choice

3.1. The 3 samples procedure

Even if the previous inequality is not directly applicable, SRM theory proved that the complexity is not equal to the number of parameters, and gives a way to handle methods where penalized likelihood is not applicable. One important idea is that one has to realize a trade-off between the fit and the robustness of a model.

An empirical way of choosing a model in the spirit of Statistical learning Theory is the following (Hastie & al., 2001):

Split the available data into 3 parts: the first set (training) is used to fit the various models of a family (parameter estimations), the second set (validation set) is used to estimate the prediction error of each previously estimated model and choose the best one, the last set (test set) is reserved to assess the generalization error rate of the best model. This last set is necessary, because the repeated use of the validation step is itself a “learning” step.

However split only once the data set into 3 parts is not enough, due to sampling variations. All this process should be repeated a number of times to get mean values and standard errors (see part 4).

Let us remark that after having selected the best model and assessing its error rate with the 3 sets methodology, the final model has to be reestimated using all available data in order to have the best parameter estimates.

3.2. Performance measures

We will focus here on supervised classification into 2 groups. Error rate estimation corresponds to the case where one applies a strict decision rule. But in many other applications one just uses a "score" S as a rating of the risk to be a member of one group, and any monotonic increasing transformation of S is also a score. Usual scores are obtained with linear classifiers (Fisher's discriminant analysis, logistic regression) but since the probability $P(G_1|\mathbf{x})$ is also a score ranging from 0 to 1, almost any technique gives a score.

The ROC curve synthesizes the performance of a score for any threshold s such that if $S(\mathbf{x}) > s$ then $\mathbf{x}$ is classified in group 1. Using s as a parameter, the ROC curve links the true positive rate to the false positive rate. The true positive rate (or specificity) is the probability of being classified in G_1 for a member of G_1 : $P(S>s|G_1)$). The false positive rate (or 1- sensitivity) is the probability of being wrongly classified to G_1 : $P(S>s|G_2)$.

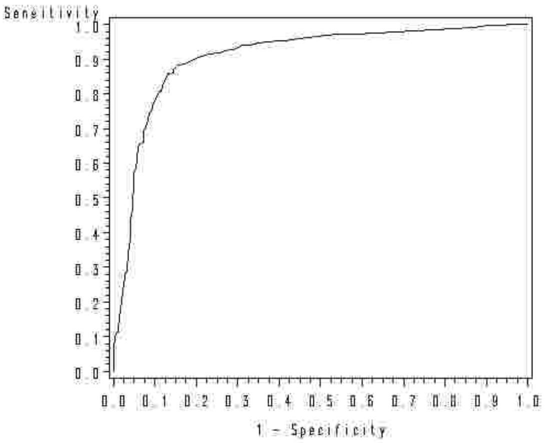


Figure 4

In other words, the ROC curve links the power of the procedure 1-β to α, the probability of error of first kind.

One of the main properties of the ROC curve is that it is invariant with respect to increasing (not only linear) transformations of S . Since the ideal curve is the one which sticks to the edges of the unit square, the favourite measure is given by the area under the ROC curve (AUC). Theoretical AUC is equal to the probability of "concordance" : AUC = $P(X_1>X_2)$ when one draws at random two observations independently from both groups. $AUC = \int_{s=+\infty}^{s=-\infty} (1-\beta(s))d\alpha(s)$. For two samples of n_1 and n_2 observations AUC is estimated by $c = \frac{n_c}{n_1 n_2}$ where n_c is the number of concordant pairs. AUC comes down to Mann-Whitney's U statistic : AUC = U/n_1n_2.

The diagonal corresponds to the worst case where score distributions are identical for both groups: some practitioners use then the so-called Gini index G instead of AUC. G is twice the area between the ROC curve and the diagonal G = 2AUC-1 . When there are no ties, G is equal to Somers'D.

ROC curves and AUC measures are commonly used to compare several scores or models, as long as there is no crossing. The best one has the largest AUC or G.

4. A case study

We exemplify the notions evocated in section 3 on a marketing data set (http://www.math.mcmaster.ca/peter/sora/case_studies_00/data_miningf). The sample consist of 2158 accounts. The response variable indicates whether or not a consumer answered to a direct mail campaign for a specific product. Among the 200 explanatory variables we selected 69 (including indicators for gender, recency, frequency, monetary type data for the specific accounts, census variables, "taxfiler" variables). We applied the two main classification techniques : Fisher's linear discriminant analysis (LDA) and logistic regression.

Both techniques lead to a score function $S(\mathbf{x})$ = $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ and a posterior probability for group 1 equal to

$$P(G_1|\mathbf{x}) = \frac{\exp(S(\mathbf{x}))}{1+\exp(S(\mathbf{x}))} = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

Modifying priors changes only the constant term in the score function.
The previous formula is obtained in LDA under normality and equal covariance matrices assumptions, while it is the model in logistic regression. Estimation techniques differs: least squares in LDA , conditional maximum likelihood in logistic regression.

Logistic regression is very popular since the β_j are related to odds-ratios. The probabilistic assumptions of logistic regression seem less restrictive than those of discriminant analysis, but discriminant analysis also has a strong non-probabilistic background being defined as the least-squares separating hyperplane between classes. Since the question is to find the best model in terms of prediction, the right thing to do is to compare their performance measured here by AUC.

Figure 5 shows very close results : logistic regression has a slightly greater AUC than discriminant analysis 0.830 instead of 0.829, but with a standard error of 0.009 the difference is not significant.[1] (Table 1).

[1] Analysis were performed with SAS 9.1. ROC curves and AUC were computed with SPSS 14

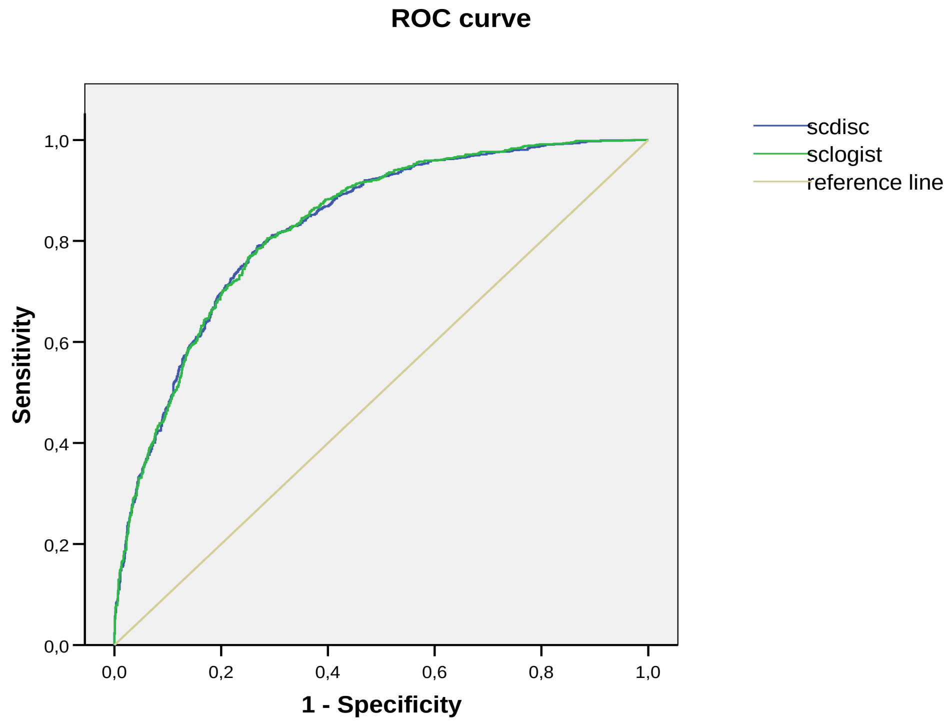

ROC curve

Figure 5

AUC

	Zone	Std. err	Asymptotic confidence interval 95%	
			Lower bound	Upper bound
scdisc	0,829	0,009	0,812	0,846
sclogist	0,830	0,009	0,813	0,847

Table 1

The above comparison was done with the total sample and may suffer from the resubstitution bias since the same data set is used twice : for estimating score and for prediction. As indicated in section 3.1 if we want to compare predicting capabilities of both methods, it is necessary to do so with an independent sample : one has to divide randomly the total sample into two parts : the training set and the validation set[2]. In order to avoid a too specific pattern, we did this random split 10 times using a stratified sampling (the strata are the two groups) without replacement of 70% for the training sample and 30 % for the validation sample. It is like a bootstrap technique but without replacement.

[2] We use here only two sets and not three, since the objective is to measure the accuracy of a method and not to do model selection.

The performance of both methods was measured by the AUC computed for each of the 10 validation samples.

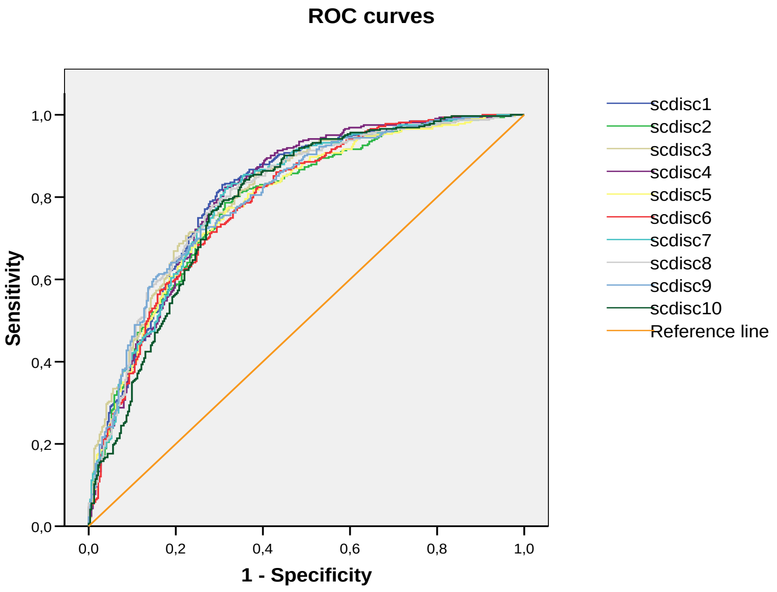


Figure 6 Discriminant analysis

Table 2

AUC

Variable(s)	Zone
scdisc1	,809
scdisc2	,787
scdisc3	,808
scdisc4	,808
scdisc5	,787
scdisc6	,789
scdisc7	,803
scdisc8	,806
scdisc9	,800
scdisc10	,789

$\overline{x} = 0.7986$

$s = 0.0095$

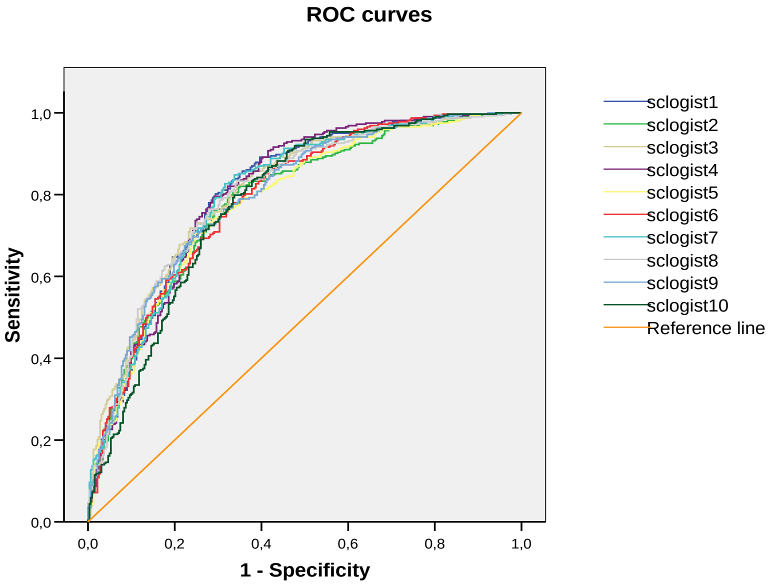


Figure 7 Logistic regression

Table 3

AUC

Variable(s)	Zone
sclogist1	,806
sclogist2	,786
sclogist3	,810
sclogist4	,806
sclogist5	,782
sclogist6	,791
sclogist7	,802
sclogist8	,803
sclogist9	,798
sclogist10	,782

$\overline{x} = 0.7966$

$s = 0.0105$

Figures 6 and 7 as well as tables 2 and 3 confirm that:

- Linear discriminant analysis performs as well as logistic regression
- AUC has a small (due to a large sample) but non neglectable variability
- Average AUC are lower than AUC computed on the total sample but are unbiased

5. Conclusion and perspectives

5.1 Predictive modelling

If the purpose of a model is prediction, one should use adequate and objective performance measures and not "ideology" to choose between models. Error rates are too specific of a threshold and depend strongly on prior probabilities and on group frequencies. AUC is a global measure which integrates all thresholds but may be too general. One certainly needs more specific measures focussing on the central part of the ROC curve.

Measurs based on penalized likelihood are intellectually appealing but of no help for complex models where parameters are constrained.

5.2 Predict or understand?

Some predictive methods used in data mining and machine learning (SVM, neural networks for instance) are so complex that they are « black-boxes ».

The concept of a model is different from the common meaning : it is no longer a (parsimonious) representation of real world coming from a scientific theory but merely a « blind » prediction technique.

If the problem is only to get good predictions, a model should be evaluated from the point of view of its efficiency and robustness. Is it possible to predict without understanding? This may be considered as shocking but the truth is yes, due mainly to advances in machine learning.

Many applications do not require a theory, which would be difficult to elaborate: for instance it is not necessary to have a theory of consumer to predict if someone will accept a commercial proposition. Statistics is in this case a (very efficient) decision support technique and not an auxiliary of science.

If the best model is the one which gives the best predictions, it has to be understood by users, especially when decisions implying citizens life are taken (reject a loan). In this respect, decision trees, linear models are commonly accepted but not non-parametric density estimation, non-linear SVM. But acceptability of methods change with time and a model which is considered as "complex" now, may be considered as standard 20 years later.

References

Devroye, L., Györfi, L., Lugosi, G. (1996) A Probabilistic Theory of Pattern Recognition, Springer

Giudici, P. (2003) *Applied Data Mining*, Wiley

Hand, D.J. (2000) Methodological issues in data mining, in J.G.Bethlehem and P.G.M. van der Heijden (editors)*, Compstat 2000 : Proceedings in Computational Statistics*, 77-85, Physica-Verlag

Hastie, T., Tibshirani, F., Friedman J. (2001) *Elements of Statistical Learning*, Springer

Saporta G., Niang N. (2006) Correspondence analysis and classification, in J.Blasius & M.Greenacre (editors) *Multiple Correspondence Analysis and Related Methods*, 371-392, Chapman & Hall

Vapnik, V. (2006) *Estimation of Dependences Based on Empirical Data*, 2nd edition, Springer