Model Selection and Predictive Inference

Gilbert Saporta

MODEL SELECTION AND PREDICTIVE INFERENCE

Gilbert Saporta
Chaire de Statistique Appliquée & CEDRIC
Conservatoire National des Arts et Métiers
292 rue Saint Martin, case 441
75141 Paris cedex 03, France
saporta@cnam.fr

Abstract: Methods based on penalized likelihood cannot be applied in many problems. Statistical learning theory provide the theoretical framework for predictive inference, but model choice based on VC dimension is often not feasible. In binary classification, ROC curve and AUC provide a reasonable criterium for model choice, combined with resampling.
Mathematics Subject Classification (2000): 62F35,62F40,62H30, 68T05
Key words: model selection, ridge regression, statistical learning, discriminant analysis

1.Introduction

A considerable amount of literature has been devoted to model selection by minimizing penalized likelihood criteria like AIC, BIC. This makes sense in the classical framework where a model is a simplified representation of the real world provided by an expert of the field. AIC and BIC have similar formulas but originates from different theories. Even in this context, penalized likelihood may not be applicable when there is no simple distributional assumption on the data (what is likelihood?) and (or) when one uses regularisation techniques like ridge or PLS regression where parameters are constrained (what is the actual number of parameters?).

In data mining and machine learning, models come from data and not from a theory behind it: models are used to make predictions (supervised learning) [5]. A good model not only fits the data but gives accurate predictions, even if it is not interpretable. A model is nothing else but an algorithm and the search for the true model is vain.

A more adapted measure of complexity is the VC-dimension which leads to the SRM principle for model selection which is universally strongly consistent, but the VC dimension is difficult to compute. Empirical measures of generalization are in general based on techniques like bootstrap or cross-validation [6].

We will focus on supervised binary classification: ROC curves and AUC are commonly used [7]. Comparing models should be done on validation (hold-out) sets and we will show on examples that resampling is necessary in order to get confidence intervals and how unexpected variability may occurr.

2.Model choice and penalized likelihood

A crude version of the likelihood principle which comes back to R.A.Fisher says that among several values of a parameter θ, one must choose the one which maximizes the probability density function $L(x_1,..,x_n;\theta)=\prod_{i=1}^{n} f(x_i;\theta)$ considered as a function of θ. This principle may be extended to the choice between different parametric families of distributions (variable selection being a particular case of model selection). However the likelihood increases with the number of parameters and in order to prevent from overfitting, one uses penalized versions of the likelihood which are modern version of Occam's razor or *lex parsimoniae*.

The two best known criteria are Akaike's AIC and Schwartz's BIC [3] :

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

BIC favourises more parsimonious models than *AIC* due to its penalization , *AIC* (not *BIC*) is biassed 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.
It must be stressed here that there is no rationale to use simultaneously *AIC* and *BIC*, since they come from different theories: *AIC* is an approximation of the Kullback-Leibler divergence between the true model and the estimated one, while *BIC* comes from a bayesian choice based on the maximisation of the posterior probability of the model, given the data.
There are severe practical limitations in the use of penalized likelihood which cannot be applied to many popular predictive models such as decision trees, ridge or PLS regression, since there may be no simple likelihood nor a simple number of parameters: what is eg the number of parameters for a ridge regression?
Both criteria assumes the existence of a "true" model belonging to the family of distributions of interest. This is of course questionable and we must remind of the famous dictum from G.E.P.Box [1]: "Essentially, all models are wrong, but some are useful". This is especially true for very large data sets where no simple parsimonious model can fit to the data.

3.Models for prediction

In data mining applications, a model is merely an algorithm, coming more often from the data than from a theory. The focus here is not on an accurate estimation of the parameters, or on the adequacy of a model on past observations but on the predictive ability, ie the capacity of making good predictions for new observations: forecasting differs from fitting.

3.1 The bias-variance trade-off [5]

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? The answer is given by statistical learning theory [8].

3.2 VC dimension and model choice through SRM

Let us consider a binary classification problem where y and $\hat{y}$ take their values in {-1 ;+1} with the following loss function

$$L(y;\hat{y}) = \frac{1}{2}\left|y - \hat{y}\right| = \frac{1}{4}\left(y - \hat{y}\right)^2$$

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. Ordinary least squares 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 $P(z)$. R_{emp} is a random variable. A

model is consistent if R_{emp} converges towards R when n tends to infinity. 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. h is not equal to the number of parameters: it may be smaller when there are regularization constraints.

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 that for h+1 points a non-separable configuration always exists. 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 which 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, 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. This is illustrated by figure1.

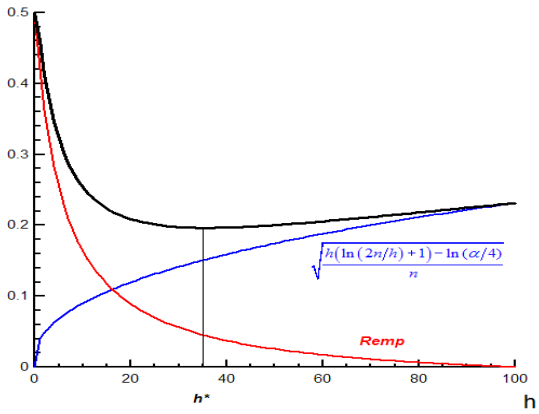

Figure 1: The SRM principle

Devroye [4] and Vapnik [8] have proved that for any distribution , the SRM provides the best solution with probability 1 (SRM is universally strongly consistent).

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.

4. Empirical model choice for binary supervised classification

4.1 Splitting the sample

Even if Vapnik's inequality is not directly applicable, SRM theory 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. [6]):

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. In order to avoid too specific patterns, all this process should be repeated a number of times to get mean values and standard errors.

In [2] extensive simulations showed that a resampled 10-fold cross-validation technique outperformed other estimators, such as bootstrap, for measuring the prediction error of a linear model.

4.2 ROC curve and AUC

Error rate estimation corresponds to the case where one applies a strict decision rule and depend strongly on prior probabilities and on group frequencies. But in many applications one just needs a "score" S ie 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 : $P(S>s|G_2)$. In other words, the ROC curve links the power of the procedure $1-\beta$ to α, the probability of error of first kind.

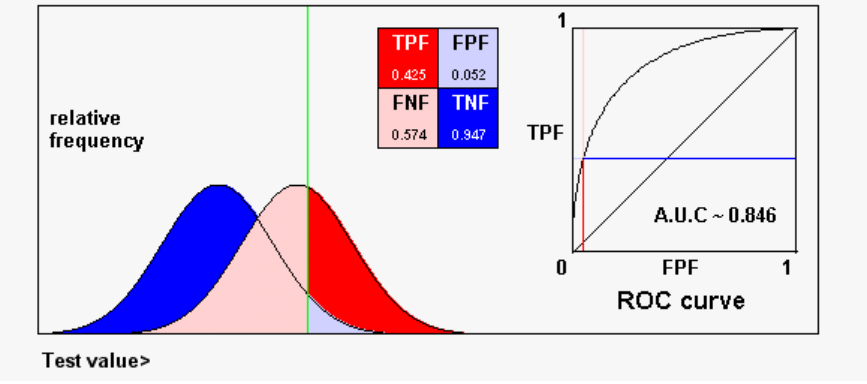

Figure 2 : Score distribution and the ROC curve (http://www.anaesthetist.com/mnm/stats/roc/)

ROC curve is invariant with respect to increasing 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.
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. Model choice should be done by using the split sample procedure with AUC instead of the error rate.

5. A case study

The data set (http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm) consists in 768 patients described by eight variables and a response variable which indicates whether or not the patient is diabetic. Two standard classification techniques: Fisher's linear discriminant analysis (LDA) and logistic regression are applied. Both techniques lead to a linear score function $S(x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p$. When using the total data set, figure 3 shows quite exactly the same curves for both methods.

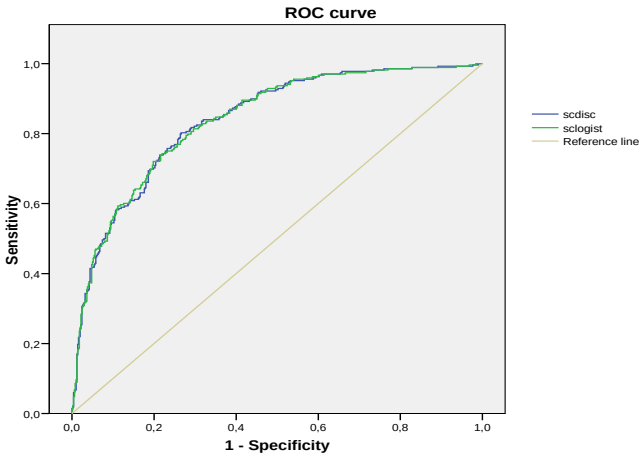

Figure 3

For the empirical comparison, we performed LDA and logistic regression on thirty training samples of 70% and evaluate the AUC on their corresponding validation sets (30%).
The results in figure 4 confirm that linear discriminant analysis performs as well as logistic regression, despite the (untrue) belief that "*logistic regression is a safer and more robust bet than the LDA model, relying on fewer assumptions*" [6]. AUC has a small but non neglectable variability and there is a large variability due to subset selection : ROC curves may sometimes show very specific and unexpected patterns.

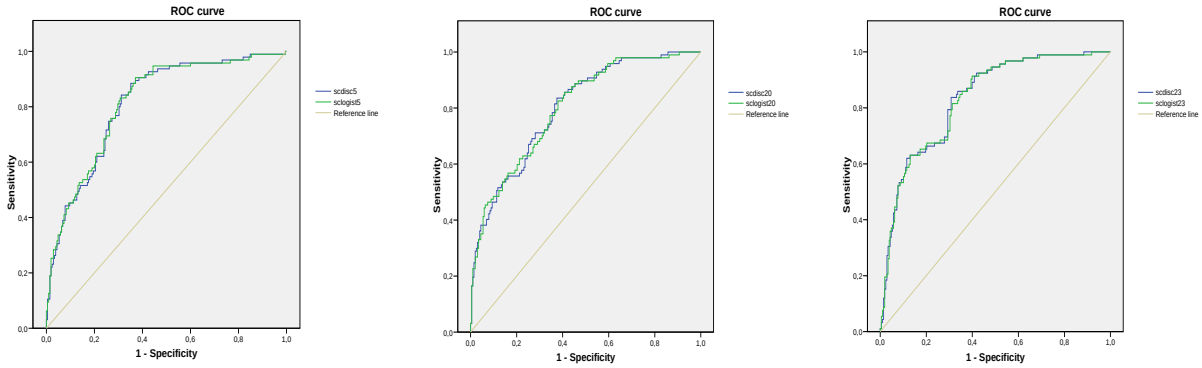

Figure4

6.Discussion

Statistical models have two meanings according to the problem : the first one is to help understanding a complex reality thanks to a simplified representation of the relations between variables, the second one in predictive inference is to give predictions.

In the first case, a good model must be parsimonious and have a good fit : measures based on penalized likelihood are very useful and intellectually appealing. In the second case models for prediction have to be efficient for new observations and penalized likelihood is of no help for complex models where parameters are constrained. In predictive inference models could be very complex, even a “blind” technique or a black box.

Statistical Learning Theory gives useful insights on the trade-off between fit and generalization. In predictive inference, one should use adequate and objective performance measures to choose between models. AUC is a very useful measure which integrates all thresholds but may be too general and one certainly needs more specific measures focussing on the central part of the ROC curve.

Resampling is necessary to estimate performance and may lead to unexpected patterns. A limitation for all these techniques is the assumption that future data will be drawn from the same distribution than the one observed in the past. Results are not valid when there are changes in the population.,

References

[1] Box, G.E.P. and Draper, N.R.: *Empirical Model-Building and Response Surfaces*, p. 424, Wiley, 1987

[2] Borra, S. and Di Ciaccio, A.: Measuring the prediction error. A comparison of cross-validation, bootstrap and hold-out methods, in Ferreira, C., Lauro, C., Saporta, G. and Souto de Miranda, M. (eds) *Proceedings IASC 07, Aveiro, Portugal,* 2007:ISBN 978-90-73592-26-1

[3] Burnham, K.P. and Anderson, D.R.:*Model Selection and Inference*, Springer, 2000

[4] Devroye, L., Györfi, L. and Lugosi, G.: *A Probabilistic Theory of Pattern Recognition*, Springer, 1996

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

[6] Hastie, T., Tibshirani, F., and Friedman J.: *The Elements of Statistical Learning*, Springer, 2001

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

[8] Vapnik, V.: *Estimation of Dependences Based on Empirical Data*, 2nd edition, Springer, 2006.