Models for Understanding versus Models for Prediction

Gilbert Saporta

Chaire de statistique appliquée & CEDRIC, CNAM
292 rue Saint Martin, Paris, France *saporta@cnam.fr*

Abstract. According to a standard point of view, statistical modelling consists in establishing a parsimonious representation of a random phenomenon, generally based upon the knowledge of an expert of the application field: the aim of a model is to provide a better understanding of data and of the underlying mechanism which have produced it. On the other hand, Data Mining and KDD deal with predictive modelling: models are merely algorithms and the quality of a model is assessed by its performance for predicting new observations. In this communication, we develop some general considerations about both aspects of modelling.

Keywords: model choice, data mining, complexity, predictive modelling

1 Models for understanding

A statistical model consists usually in the formulation of a parametric formula for the distribution of a (multidimensional) random variable. When the interest lies in a particular response, the usual form is $y = f(x;\theta) + \varepsilon$. When the model is completely specified by an expert (economist, biologist, etc.) the statistical work consists in estimating the unknown parameters, and (or) refute the model according to a goodness of fit test. If the model is rejected, the expert should think of an other one.
Generally a model should be simple, and parameters should be interpretable in terms of the application field : elasticity, odds-ratio, etc. The need for interpretation explains why for instance logistic regression is preferred to discriminant analysis by biostatisticians and econometricians, the coefficients being uniquely estimated and having the meaning of log-odds.
The purpose of a model is to give insights in the nature of a stochastic phenomenon, not necessarily to give accurate predictions. This may be viewed as a paradox, since *eg* in natural sciences, a good model must give good predictions, otherwise the model is replaced by an other one. It is due to the importance of the random term ε . In epidemiology for instance, it is more important to find risk indicators than having an accurate individual prediction of getting some disease.

1.1 Model estimation

Maximum likelihood estimation is by far the standard technique : 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 which is equal for iid observations to

$$L\left(x_1,..,x_n;\theta\right)=\prod_{i=1}^{n}f\left(x_i;\theta\right)$$

considered as a function of θ.
Advantages of ML estimation are in availability of asymptotic standard errors, as well as tests. The extensive use of ML estimation is recent: the first technique for estimating the logistic regression model proposed by Berkson in the 40's was the minimum chi-squared, see Berkson (1980).
Least squares estimation is often more robust and need less assumption (both are related) and is of common use, especially in exploratory analysis, including PLS structural equation modelling.

1.2 Model choice

Choosing between several models occurrs when the "expert" hesitates between several formulations. Statistics may help to choose among several models using some parsimony principle. This is conform to Occam's razor, which is often considered as a scientific principle against unnecessary hypotheses . The major use of model selection is for variable selection including interaction selection.
A considerable amount of literature has been devoted to model selection by minimizing penalized likelihood criteria like AIC, BIC, see Burnham and Anderson (2000).

$$AIC=-2\ln\left(L(\hat{\theta})\right)+2k\quad BIC=-2\ln\left(L(\hat{\theta})\right)+\ln(n)k \tag{1}$$

BIC favourizes more parsimonious models than AIC due to its penalization. AIC, but 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.
AIC and BIC have similar formulas but originates from different theories and there is no rationale to use simultaneously AIC and BIC: 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.

1.3 Some limitations

Even if we knew the *"true"* model, parameter estimation could be a difficult task when the number of cases is low. For example in a multiple regression model, this can lead to severe multicollinearity. If we want to estimate all parameters, without discarding variables (and we should not discard variables if we believe in our model), it is necessary to put some constraints or in other words to do some regularization. Ridge regression which is a direct application of Tikhonov regularisation is a well known remedy to multicollinearity. Projecting onto a lower dimension space is another kind of regularization and includes principal components regression as well as PLS regression.
Bayesian statistics provides an elegant solution: it balances the lack of observations by using prior information. For the normal regression model with normal priors on the parameters, bayesian estimation comes down to ridge regression and provides an enlightening interpretation of this technique.

Model choice by penalized likelihood suffers from practical limitations: penalized likelihood cannot be applied to ridge or PLS regression, since there may be no simple likelihood nor a simple number of parameters: what is eg the right number of parameters for a ridge regression? There are still p parameters but since they are constrained by a condition like $\|\beta\|^2 < k$, we need an equivalent number of parameters less than p depending on k but the exact formula is unknown. No need to say that penalized likelihood is hard to apply to choose a decision tree, or between a decision tree and a logistic regression. Let us also remark that the underlying hypothesis for BIC of having a uniform prior on models is not very realistic.

The ambition of finding the *"true"* model belonging to the family of distributions is questionable and we must remind of the famous dictum from George Box (Box and Draper,1987, p.424): *"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: it is well known that significance or goodness of fit tests always reject any precise null hypothesis when one has millions of observations: a correlation of 0.01 will be considered as significatively different from zero, but the point of interest is in the strength of the relationship, not in its existence (assuming that a correlation different from zero is a proof of existence...)

1.4 Models for Exploration

Most of what have been said before is about *regression* in the broad sense of relating a response to some inputs. There is a slightly different use of models, applied to some kind of exploratory problems.
Let us suppose that we analyze a sample drawn *iid* from a population defined by some model, then it is possible to do some inference for the outputs of

the analysis: confidence intervals for the eigenvalues of a PCA, confidence regions for the positions of points in multidimensional scaling, PCA, or Correspondence Analysis etc. The model is here an help to interpret and provide additional knowledge to what can be seen from graphical displays of exploratory analysis.
However the validity of the model is is often dubbious for numerical variables. The standard model for inference in PCA is the multivariate normal, as one can find in any textbook, and for most data sets this model is wrong : there is here a contradiction in using both an exploratory technique, which has for goal to reveal the underlying structure of frequently heterogenous data and a single simple model. The situation is more comfortable for categorical data where a multinomial scheme is realistic. For instance in contingency tables analysis it is true, provided that the sampling scheme is simply random, that the joint distribution of the frequencies n_{ij}, is the multinomial $M(n; p_{ij})$ where the p_{ij} are unknown parameters. This lead to exact results (however difficult) for the distribution of eigenvalues in correspondence analysis. This may be extended to the case where the margins of one variable are fixed (stratified sampling).
Many experiences have proved that resampling, eg bootstrap, provides more reliable results than using unrealistic models, see Hatabian and Saporta (1986) or Lebart (2006). Let us however mention the neglected fixed-effect model for PCA (Besse et al. , 1988) which is less demanding in its hypotheses than the multivariate normal and lead to inferences on dimensionality and on the displays.

2 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.
The "black-box model" (Vapnik, 2006, p.416) illustrates the differences with the previous conception of a model, while keeping the same mathematical formulation $y = f(x; \theta) + \varepsilon$. Statistical modelling (understanding data) look for a parsimonious function $f(x; \theta)$ belonging to a prespecified set. On the other hand, in predictive modelling, the aim is not to approximate the true function but to get a function which gives as accurate predictions as the model, since it is a stochastic one. The question is not to discover the hidden mechanism but to perform as well.
In many operational applications, like in Customer Relationship Management or pattern recognition, understanding the phenomenon would be a too complex and vain task: a banker does not need a theory for predicting if a loan will at risk or not, but only a good score function. In predictive inference,

models could be very complex like multilayer perceptrons or non-linear SVM, and we have the paradox that a good prediction does not need a deep understanding of what is observed. This may not be confused with the readability of a model: a decision tree is very simple to use for the end-user but is not a model of the hidden mechanism producing the data.

2.1 Risk minimization

Let L be a loss fonction and R=$E(L)$ its expectation, called here "risk".

$$R = E(L) = \int L(z, \theta) dP(z) \tag{2}$$

The risk is the average loss for new observations. The ideal would be to choose the model among some family of models in order to minimize R but it is an impossible task, since we do not know the true distribution P. Choosing the model which minimizes the empirical risk (ie the risk on observed data, or learning set)

$$R_{emp} = \frac{1}{n} \sum_{i=1}^{n} L(y_i; f(x_i; \theta)) \tag{3}$$

usually leads to overfitting.
For binary classification where one chooses as loss function the misclassification rate, Vapnik's inequality gives an upper bound relying on the VC-dimension h :

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

Based on the upper bound of R, the structured risk minization principle or SRM provides a model choice methodology different from penalized likelihood, since no distributional assumptions are necessary. Given a 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 inequality proves that (provided h is finite) one may increase the complexity of a family of models (eg in a simple case increase the degree of polynomials) when the number of learning cases increases, since it is the ratio h/n which is of interest. This shows a strong difference between SRM and model choice based on BIC, since the penalization in BIC increases with n and tends to choose simpler models for large n. Devroye et al. (1996) proved that SRM is strongly universally consistent.

2.2 AUC-like measures of efficiency

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 which is a rating of the risk to be a member of one group, and any monotonic increasing transformation of S is also a score like

$$S' = \frac{\exp(S)}{1+\exp(S)}$$

Usual scores are obtained with linear classifiers (Fisher's discriminant analysis, logistic regression) but since a probability is also a score (ranging from 0 to 1), almost any technique, even decision trees gives a score. SVM classifier also provide scores.

The ROC curve is a popular measure of efficiency which synthesizes the performance of a score for any threshold s such that if $S(x) > s$ then 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. 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 most popular measure is given by the area under the ROC curve (AUC); another measure is the so-called Gini index which is equal to twice the area between the ROC curve and the diagonal: $Gini = 2AUC-1$. Since 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 reduces to an old measure of nonparametric comparison: Mann-Whitney's U statistic.

ROC curve and AUC are extensively used in the banking industry to assess the quality of the credit risk rating system and are recommanded by the Basel Committee on Banking Supervision (see BCBS 2005).

Model choice using AUC should of course not be based on the learning sample. Inequalities similar to (4) , may be obtained for AUC but are not very useful in practice. Moreover ROC curve and AUC do not take into account some elements of interest in business applications like the error costs and the fact that very often the two subpopulations are not balanced at all.

2.3 Empirical model choice

Even if Vapnik's inequality is not directly applicable, for it is often difficult to evalute the VC dimension, 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 et al. 2001): Split the available data into 3 parts: the first set (training) is used to fit the various models of a family, 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 using repeatedly the validation step is itself a learning step.
However split only once the data set into 3 parts is not enough, and may lead to unexpected sampling variations, see Niang and Saporta (2007). In order to avoid too specific patterns, all this process should be repeated a number of times to get mean values and standard errors. For measuring the prediction error in regression, Borra and Di Ciaccio (2007) compared several resampling technique including bootstrap and .632 bootstrap; they showed by simulation that a resampled 10-fold cross-validation technique outperformed other estimators. Since Fisher's supervised classification in 2 classes is a special case of the linear model, the latter results may be also valid for discrimination.

3 Conclusions

Two very different conceptions correspond to the same name of "model", and this may be a cause of misunderstanding. As Cherkassky and Mulier (1998) wrote: *"Too many people use different terminology to solve the same problem; even more people use the same terminology to address completely different issues"*. Models for understanding data correspond to the part of statistics considered as an auxiliary of science. Models for prediction belong to the other face of statistics as an help for decision. There are more job opportunities for graduate students in predictive modelling but also more competitors coming from other disciplines.
But one may question this opposition between science and action : when a technique gives really good predictions, it is also an improvement of the knowledge we have on data. Predictive modelling belongs to empiricism which is itself a theory of knowledge.

References

BCBS (2005): Studies on the Validation of Internal Rating Systems, *Basel Committee on Banking Supervision, Bank of International Settlements*, *http : //www.bis.org/publ/bcbs_wp*14.*htm*

BERKSON, J. (1980): Minimum chi-square, not maximum likelihood! *Annals of Mathematical Statistics* 8, 457-487.

BESSE, P., CAUSSINUS, H., FERRE, L., FINE, J. (1988): Principal Components Analysis and Optimization of Graphical Displays, *Statistics*, 19, 301-312.

BORRA, S. and Di CIACCIO, A.(2007): 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*

BOX, G.E.P. and DRAPER, N.R. (1987): *Empirical Model-Building and Response Surfaces*, Wiley

BURNHAM, K.P. and ANDERSON, D.R. (2000): *Model Selection and Inference*, Springer

CHERKASSKY, V., MULIER, F., (1998): *Learning from data* , Wiley

DEVROYE, L., GYÖRFI L. and LUGOSI, G. (1996): *A Probabilistic Theory of Pattern Recognition*, Springer

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

HASTIE, T., TIBSHIRANI, F., and FRIEDMAN J. (2001): *Elements of Statistical Learning*, Springer

HATABIAN G., SAPORTA G. (1986): Régions de confiance en analyse factorielle, in Diday E. (ed) *Data Analysis and Informatics IV*, North-Holland, 499-508

LEBART L. (2006): Validation Techniques in Multiple Correspondence Analysis, in Greenacre M. and Blasius J. (eds) *Multiple Correspondence Analysis and related techniques*, Chapman and Hall/CRC, 179-196

NIANG N. and SAPORTA G. (2007): Resampling ROC curves, in Ferreira, C., Lauro, C., Saporta, G. and Souto de Miranda, M. (eds), *Proceedings IASC 07, Aveiro, Portugal*

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