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

A considerable amount of literature has been devoted to model selection by minimizing penalized likelihood criteria like *AIC*, *BIC* (Burnham et al. 2000). 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) (Hand, 2000). 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 (Devroye et al., 1996, Vapnik, 2006), but the *VC* dimension is difficult to compute. Empirical measures of generalization are in general based on techniques like bootstrap or cross-validation (Hastie et al., 2001).

We will focus on supervised binary classification: *ROC* curves and *AUC* are commonly used (Saporta & Niang, 2006). 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.

References

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

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

[3] 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

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

[5] 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

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