Supervised Classification and AUC

Ndèye Niang[1], Gilbert Saporta[1]

Chaire de Statistique Appliquée & CEDRIC
CNAM, 292 rue Saint Martin, 75141 Paris Cedex 03, France

Abstract In supervised classification, ROC curves and AUC are commonly used to evaluate and to compare models performances. Evaluations of AUC are usually done on one validation (hold-out) set. Resampling procedures allow a better use of ROC curves and AUC for predictive purposes.

Keywords: ROC curve, AUC, resampling

1 Measures of performance: Roc curve, Lift curve and Gini index

We focus on supervised classification into two 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 (eg Fisher's discriminant analysis, logistic regression) but since the probability $P(G_1|\mathrm{x})$ of classifying an observation $\mathbf{x}$ in the group G1 is also a score ranging from 0 to 1, almost any technique gives a score.

The Receiver Operating Characteristic (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 G1. Using s as a parameter, the ROC curve links the true positive rate (or specificity) to the false positive rate (or 1- sensitivity). One of the main properties of the ROC curve is that it is invariant with respect to any 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 of performance 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 $AUC = \int_{s=+\infty}^{s=-\infty} (1 - \beta(s))d\alpha(s)$ where $1 - \beta$ is the power of the procedure, and α is the probability of the first kind error. The diagonal corresponds to the worst case where score distributions are identical for both groups. Some practitioners use the lift curve and the area under it (AUL) instead of AUC. The lift curve links the true positive rate (or specificity) to $P(S > s)$. AUC and AUL are linked through the Gini index G : G is the area between the lift curve and the diagonal divided by the area between the ideal lift curve and the diagonal and also twice the area between the ROC curve and the diagonal: $G = 2AUC - 1$.

2 Evaluation of AUC

Let us consider two samples of n_1 and n_2 observations drawn from both groups and some score function S related to the probability of belonging to group 1. A pair of observations x_1 and x_2, one from each group is said to be concordant if $S(x_1) > S(x_2)$. A non parametric estimate of AUC is thus given by the proportion of concordant pairs. The number of concordant pairs is equal to the well known Mann-Whitney's U statistic. Using the relationship between the U statistic and the Wilcoxon W statistic for group1: $W = U + n_1(n_1 + 1)/2$. Hanley et al. [1] obtained the standard error of the empirical AUC as :

$$SE = \sqrt{(A(1-A) + (n_1 - 1)(Q_1 - A^2) + (n_2 - 1)(Q_2 - A^2))/n_1 n_2}$$

where A is the true or theoretical AUC, an unbiased estimates of which being the empirical AUC, $Q_1 = A/(2-A)$ and $Q_2 = 2A^2/(1+A)$. The question is to know how the model will perform for future data (the generalization capacity), provided that future data will be drawn from the same distribution. Evaluating models on the basis of the learning sample may be misleading. If we want to predict capabilities of a method, it is necessary to do so with independent data : it is generally advised to divide randomly the total sample into two parts : the training set and the validation set according to a stratified sampling scheme (the strata are the two groups) without replacement of eg 70% for the training sample and 30% for the validation sample. However in order to avoid a too specific pattern, this random split should be repeated. The performance of the method can then be measured by the AUC computed for all the validation samples : the empirical mean and standard error give an unbiased estimation of future AUC and its standard error and therefore asymptotic confidence interval can be derived.

3 A case study

We exemplify the notions evocated in the previous section on a diabetis data set (http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm). We applied two standard classification techniques : Fisher's linear discriminant analysis (LDA) and logistic regression leading both to a score function. Evaluation of their performances is done by computing AUC for thirty validation sets. The results show the variability of ROC curves which may have very specific and unexpected patterns [3]. It is also shown that AUC has a small but non neglectable variability, average AUC for both methods are lower than AUC computed on the total sample but are unbiased, and LDA performs as well as logistic regression.

References

[1] Hanley, J.A. and McNeil, B.J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, **142** (1982) 29-36.

[2] Hanley. J.A, and McNeil, B.J. : A method of comparing the areas under receiver operating characteristic (ROC) curves derived from the same cases. Radiology, **148**(1983) 839-843.

[3] Saporta, G. and Niang, N.: Resampling ROC curves. In IASC meeting on Statistics for Data Mining, Learning and Knowledge Extraction, IASC07 August 30-September 1, Aveiro, Portugal, (2007).