Resampling ROC curves

Ndèye Niang and Gilbert Saporta

Chaire de Statistique Appliquée & CEDRIC, CNAM, 292 rue Saint Martin,
F-75003 Paris

Key words: ROC curve, AUC, resampling

Abstract

Resampling procedures allows a better use of ROC curves and AUC for predictive purposes. We also address a drawback of AUC for the comparison of ROC curves which are crossing, by recommending the use of partial AUC.

1 ROC curve and AUC as a measure of performance

Receiver operating characteristic (ROC) curves are very popular for evaluating a diagnostic test or score performances in various decision making applications: medicine, marketing, credit scoring etc. The ROC curve provides a concise graphical representation of the trade off between sensitivity and specificity. We will focus here 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 (Fisher's discriminant analysis, logistic regression) but since the probability $P(G_1 \mid \mathrm{x})$ is also a score ranging from 0 to 1, almost any technique gives a score.

Several indices have been used to summarize the information contained in a ROC curve. The most common index is the area under it or AUC, Hanley et al. (1982). ROC curves and AUC are used for comparing models and techniques for two groups supervised classification, Saporta et al.(2006).

The major drawback of these indices is that the evaluation of the performances is global, over the entire range of sensitivity and specificity. But there are cases where this may not be appropriate eg when two ROC curves intersect.

Methods for analysing only a portion of the ROC curve have been proposed with formulas to evaluate standard errors useful to build statistical tests for comparing curves. They are in general based on the assumption of normality for both groups and numerical integration is used to calculate the partial area and the delta method to obtain confidence intervals. The estimates obtained are biased and the formulas are somewhat complicated and need specific software to be computed.

The widespread availability of ever-faster computers has made resampling methods increasingly accessible and convenient tools for data analysis. Bootstrap may be specially helpful in determining confidence intervals, but also other resampling procedures. Resampling methods provide nonparametric (distribution independent) and more appropriate ways for deriving confidence interval and hypothesis testing when analysing portions of ROC curves as well as for the global AUC.

After presenting the basic AUC index for evaluating model quality in the first section, in the second we present how to compare ROC curves and their AUC for two classification methods, taking into account that the two ROC curves are derived from the same sample. We propose resampling techniques to derive an unbiased estimation of AUC and empirical confidence intervals. We exemplifying these notions on a medical data set. Then in the third section we address the major drawback of the AUC and the resampling solution we propose to analyse a portion of a ROC curve will be illustrated on a simulated example.

1.1 The theoretical ROC curve

The ROC curve synthesizes the performance of a score for any threshold s such that if $S(\mathbf{x}) > s$ 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 frequency (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 frequency (or 1- sensitivity) is the probability of being wrongly classified in G_1: $P(S>s|G_2)$.

Figure 1 (from http://www.anaesthetist.com/mnm/stats/roc/) shows the relationship between score distribution and the ROC curve.

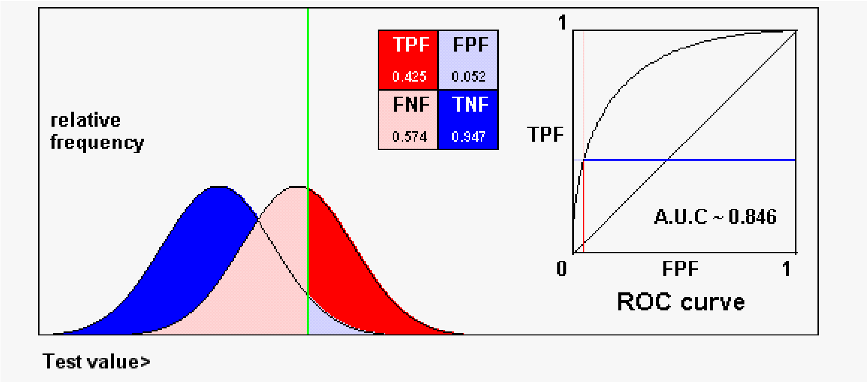


Figure 1

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

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. $AUC = \int_{s=+\infty}^{s=-\infty} (1-\beta(s))d\alpha(s)$.

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.

1.2 Estimation

Let us consider two samples of n_1 and n_2 observations drawn from two 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 the $S(x_1) > S(x_2)$.

A non parametric estimate of AUC is thus given by the proportion of concordant pairs $c = \frac{n_c}{n_1 n_2}$.

The number of concordant pairs n_c is nothing else than the well known Mann-Whitney's U statistic. Note that when there are no ties, the Gini index G is equal to Somers'D.

Using the relationship between the U statistic and the Wilcoxon W statistic for group1: $W = U + \frac{n_1(n_1+1)}{2}$, Hanley et al. (1982) obtained the standard error of the empirical AUC as:

$$SE = \sqrt{(A(1-A)+(n_1-1)(Q_1-A^2)+(n_2-1)(Q_2-A^2))/n_1n_2}$$

where A is the true or theoretical AUC and $Q_1 = A / (2-A)$ and $Q_2 = 2A^2 / (1+A)$.

A is estimated by the empirical AUC which is unbiased, but *SE* is biased and the bias depend on the shapes of the distributions. Confidence intervals are conservative and the estimates of SE will tend to be too large rather than too small Hanley et al. (1982), even if the distributions are not normal.

2 Comparing models

2.1 Testing the difference between two AUC

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. But more formal statistical criteria are needed to judge whether observed differences are significant.

Since the same data set is used to compare two scores or two ROC curves, the estimates of the AUC are not independent and the standard error of the difference between two areas cannot be computed as $SE(A1 - A2) = \sqrt{(SE^2(A_1) + SE^2(A_2)}$.

The procedure of comparing both confidence intervals is also wrong for the same reason.

The true formula is:

$$SE(A1 - A2) \ = \sqrt{(SE^2(A_1) + SE^2(A_2) - 2rSE(A_1)SE(A_2)}$$

where *r* is the correlation coefficient between the two area estimates.

Thus the general approach to evaluate whether observed difference in the two areas are significant, that is if the difference is random or real, consists on a standard statistical test of the null hypothesis that the observed difference is merely a result of random sampling, with the following decision variable:

$$z = \ (A1 - A2) / \ \sqrt{(SE^2(A_1) + SE^2(A_2) - 2rSE(A_1)SE(A_2)}$$

For deriving *r*, Hanley and McNeil (1983) introduced the two within class correlation coefficients between scores on each group which are then converted into *r* via a table provided by the authors. The former approach is somewhat complicated, needs lots of calculations which are difficult to be automatised. Moreover the *SE* statistic z is also biased.

2.2 Empirical resampling technique

Comparing models on the basis of the learning sample may be misleading since the comparison will be generally in favour of the more complex model. The question is to know which model will perform the best for future data (the generalization capacity), provided that future data will be drawn from the same distribution; when there are changes in the populations, see Hand (2006). If we want to compare predicting capabilities of several methods, 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 each 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.

2.3 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). The sample consists of 768 females with eight continuous variables measured for each unit. The response variable indicates whether or not a patient is diabetic. We applied two standard classification techniques: Fisher's linear discriminant analysis (LDA) and logistic regression.

Both techniques lead to a score function $S(x)=\beta_0+\beta_1 x_1+...+\beta_p x_p$ and a posterior probability for group 1 equal to $P(G_1|\mathbf{x})=\dfrac{\exp(S(\mathbf{x}))}{1+\exp(S(\mathbf{x}))}=\dfrac{e^{\beta_0+\beta_1x_1+...+\beta_px_p}}{1+e^{\beta_0+\beta_1x_1+...+\beta_px_p}}$

It is known that 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. 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 2 shows quite exactly the same curves for the two methods and exactly the same results for the AUC with a small but not neglectable standard error of 0.015. The difference is not significant.[1] (Table 1).

Area under ROC curve

	AUC	Std Err.	Asymptotic confidence Interval 95%	
			Lower bound	Upper bound
Scdisc	0.839	0.015	0.810	0.868
Sclogist	0.839	0.015	0.811	0.868

Table 1

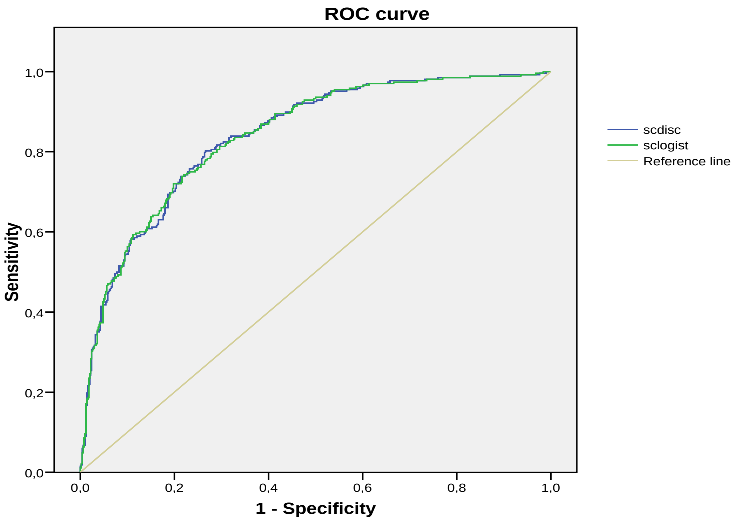


Figure 2

For the empirical comparison, we performed LDA and logistic regression on thirty training samples and evaluate the AUC on their corresponding validation sets.
The results in figure 3 as well as in table 2 confirm that:

- Linear discriminant analysis performs as well as logistic regression, their respective empirical AUC are 0.8267 and 0.8263, their standard error are 0.0169 and 0.0166.
- AUC has a small but non neglectable variability

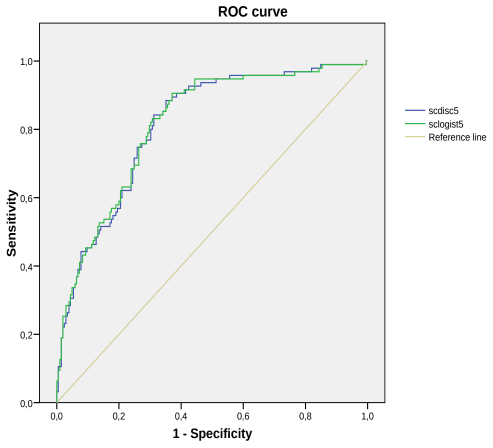
[1] Analysis were performed with SAS 9.1. ; ROC curves and AUC were computed with SPSS 14

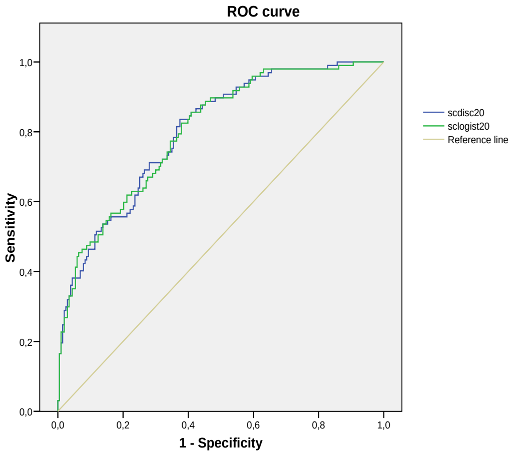
- Average AUC are lower than AUC computed on the total sample (table 1) but are unbiased.

One interesting fact is the variability of ROC curves which may have very specific and unexpected patterns like in figure 3 top left and bottom right.

sample	LDA AUC	Logistic AUC
1	0.819	0.819
2	0.83	0.831
3	0.85	0.849
4	0.817	0.816
5	0.813	0.815
6	0.827	0.825
7	0.835	0.835
8	0.822	0.821
9	0.838	0.837
10	0.821	0.821
11	0.805	0.81
12	0.82	0.816
13	0.81	0.812
14	0.821	0.822
15	0.838	0.835
16	0.843	0.844
17	0.855	0.856
18	0.834	0.835
19	0.864	0.863
20	0.801	0.801
21	0.825	0.821
22	0.829	0.83
23	0.833	0.831
24	0.816	0.813
25	0.809	0.81
26	0.859	0.856
27	0.847	0.847
28	0.804	0.801
29	0.808	0.808
30	0.81	0.81
Mean	0.8267	0.8263
Sdt err	0.0169	0.0166

Table 2





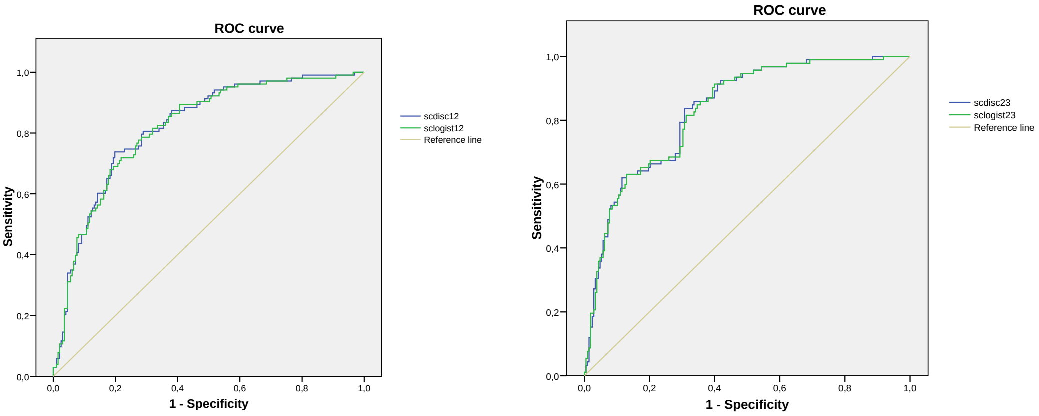


Figure 3

3 Partial ROC and AUC

In the former section we have presented methods for calculating and comparing the area under ROC curves derived from the same sample of observations. The AUC is used to quantify in a single numerical value the overall “position” of a ROC curve compared to the diagonal. A major drawback of AUC as an index of performance is its global nature. The AUC index can be in some situations not appropriate: ROC curves that cross may have similar AUC but could be different in their diagnostic efficacy and specific diagnostic (figure 4).

Important differences between ROC curves in specific regions may be muted when the total AUC is used. Moreover curves with different global AUC may be similar in particular regions. In such cases, the adequate method is to analyze portions of the curves, on predetemined regions. Methods for analyzing only a portion of the ROC curve have been proposed with formulas to evaluate standard errors useful to build statistical tests for comparing curves. As for the total AUC, these formulas are in general based on normal distribution. They are still very complicated and need specific software to be evaluated. After presenting a brief review of the literature on this subject, we propose a simple way based on trapezoidal estimate of the partial AUC and resampling techniques to evaluate its standard errors for analyzing a portion of a ROC curve and comparing ROC curves in a prespecified interval of specificity.

Area under ROC curve

	AUC	Std Err.	Asymptotic confidence Interval 95%	
			Lower bound	Upper bound
X	0.629	0.039	0.552	0.706
Y	0.712	0.037	0.640	0.783

Table 3

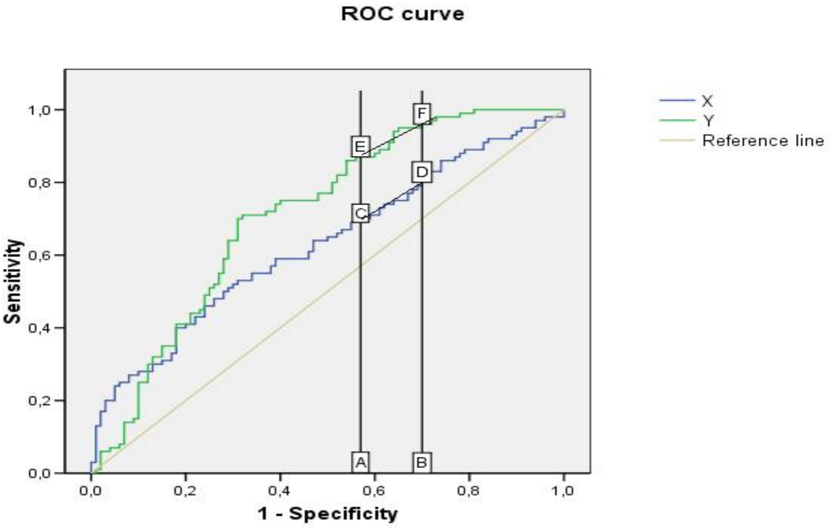


Figure 4

3.1. Definition

The partial area index proposed by Jiang et al. (1996) is a regional assessment based on partial area under the ROC curve for a range of true positive fraction (TPF) value greater than or equal to an arbitrary value TPF*.

The index is defined as the partial area divided by its maximum possible value, 1- TPF*:

$$A(TPF^*) = \frac{\int_{TPF^*}^{1} [1 - FPF(TPF)] dTPF}{1 - TPF^*}$$

The partial area under the ROC curve has also been studied by McClish (1989) in an approach that slightly differs from that of Jiang et al. McClish proposes to determine the ROC area for a range of false positive fraction (FPF) values rather than for TPF values and the upper end of the range of specificity values is not restricted to 1, so the McClish method is more flexible.

Andrew J. Dwyer (1997) compares both approaches. The works of McClish and Jiang et al. are based on the same assumption of normal distributions, use numerical integration to calculate the partial area and the delta method (Oehlert, G. W. (1992)) to obtain confidence intervals and to test the statistical significance of the difference between ROC curves.

Even if the assumption of normality produces reasonably good fits to a wide variety of ROC curves, some data sets may not fit this model. In such situations resampling techniques provide nonparametric and more appropriate methods for deriving confidence interval and hypothesis testing.

3.2 Empirical estimation

We propose to approximate the partial AUC denoted PA in the following $PA(FPF_1, FPF_2) = \int_{FPF_1}^{FPF_2} [1 - FPF(TPF)] dFPF$ by $(FPF_2 - FPF_1) * (TPF_1 + TPF_2)/2$ which is the surface of the trapeze ABDC defined by the four points (FPF_1, 0) (FPF_2, 0) (FPF_2 , TPF_2) and (FPF_1 , TPF_1) (figure 5).

On the two simulated ROC curves of figure 5 we use this approximation for the partial AUC with FPF_1= 0.05 and FPF_2 = 0.15. We obtain the corresponding partial AUC, PA = 0.0255 for the ROC curve associated to X and PA = 0.0195 for the second ROC curve. Then whereas the total AUC of the ROC curve of Y is greater than the X one (table 3), for the partial AUC it is reversed; the PA for the X ROC curve is larger.

As for the global AUC, resampling techniques may be applied to partial ROC data analysis in order to derive confidence interval and hypothesis testing for more formal comparison.

4 Conclusion and perspectives

We advocated a general approach to the problem of evaluating and comparing ROC curves based on resampling techniques. Looking at the entire ROC curves gives a useful global view but in some situations, for example when the curves cross, a partial index is more appropriate. With the approach we have proposed, once FPF values are fixed, it is quite simple to derive the corresponding areas and the standard errors to compare partially the curves.

References

Dwyer, A.J. (1997). In pursuit of a piece of the ROC. *Radiology*, **202** 621-625.

Hand, D.J. (2006). Classifier technology and the illusion of progress (with discussion). *Statistical Science*, **21** 1-34.

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

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

Jiang, Y. , Metz, C.E. and Nishikawa, R.M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*, **201** 745-750.

McClish, D.K. (1989). Analyzing a portion of the ROC curve. *Med. Decision Making*, **9** 190-195.

Oehlert, G. W. (1992). A Note on the Delta Method. *The American Statistician*, Vol. 46, No.1, 27-29.

Saporta, G. and Niang, N. (2006) Model assessment. In *KNEMO: Knowledge Extraction and Modeling, Capri,4-6 september*, IASC-INTERFACE-IFCS Workshop.