Supervised classification and AUC

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

ndeye.niang_keita@cnam.fr, gilbert.saporta@cnam.fr

Outline

1. Introduction
2. ROC curve and AUC
3. Estimation of AUC
4. Application: LDA and Logistic regression
5. Conclusion & future work

1. Introduction

- Framework of supervised classification
  - The aim : to construct a decision rule to assign new observations to one of pre-specified set of classes using information about the observations
  - The data set : descriptive variables for a sample of observations for which the true class is known (the response)

Several methods

- Many approaches : LDA, Logistic regression, ridge and PLS regression, decision trees, neural networks… (Hand 1997, Hastie & al. 2001)
- The natural question : which method to apply? Is it appropriate? Is it the best one?
- No simple answer : many factors to take into account for comparison, many criteria for evaluating performances, (Hand 2006, Jamain & Hand 2008)

Several criteria

- One of the most important issues in model comparison is the choise of a criterion among many such as :
  - likelihood ratios,
  - error rate,
  - AUC or other measures based on specificity and sensitivity (defined next)

Data Mining context
Predictive modelling

- Models are used to make predictions
  - Our interest: model efficiency: capacity to make good predictions and not only to fit to the data
  - Focus on two class-problems
  - Performance evaluation and comparison: ROC curve and associated index AUC

2. ROC curve and AUC

- A classifier corresponds to a **score function** S(x)
- the decision rule associated to a threshold *s* is :
  - classify **x** in group 1 if $S(\mathbf{x}) > s$
- Usual scores : linear classifiers (Fisher's LDA, logistic regression), non linear (Neural networks)
- Probability P(G1/x) is also a score ranging from 0 to 1 (The optimal threshold is not always 0.5)
- Almost any technique gives a score.

2. ROC curve and AUC

- **Summary of a given decision rule**

True positives	False negatives
False positives	True negatives

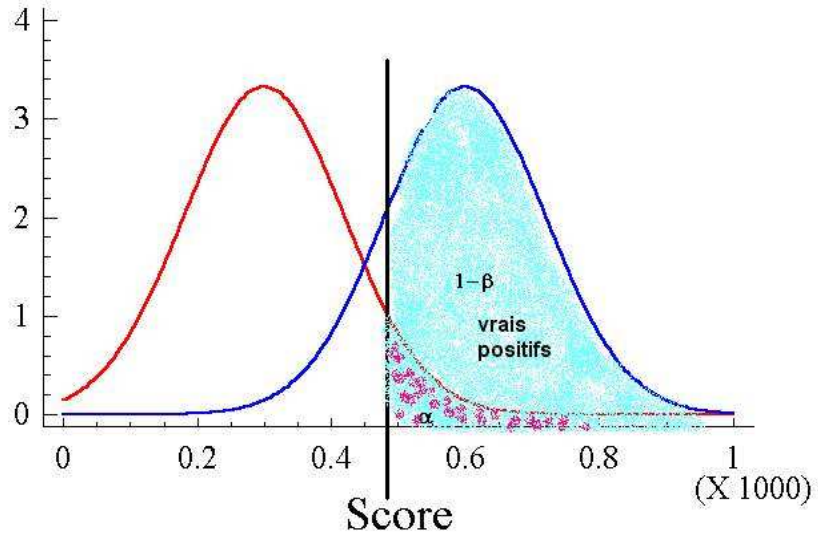
- Sensitivity = % of true positives
- Specificity = % of true negatives

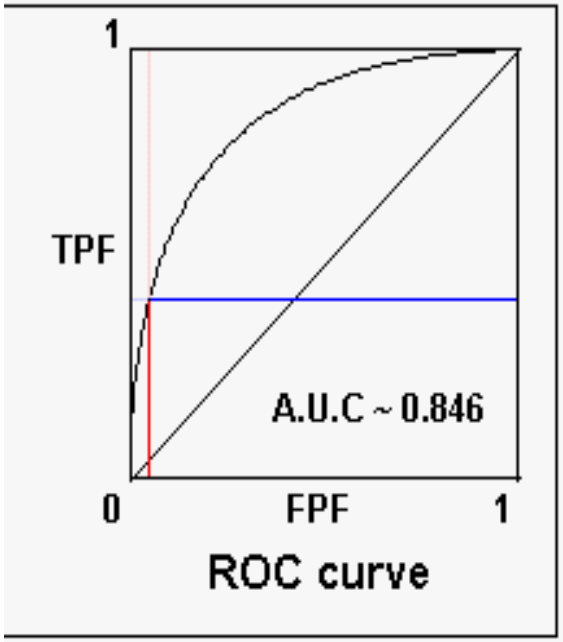
2. ROC curve and AUC

- Misclassification rate and score performance
  - Error rate implies a strict decision rule, and a fixed threshold
  - Better alternative: Area under ROC curves, overall performance of the score function, no need to specify a threshold, no misclassification cost to determine (Hand 2009)

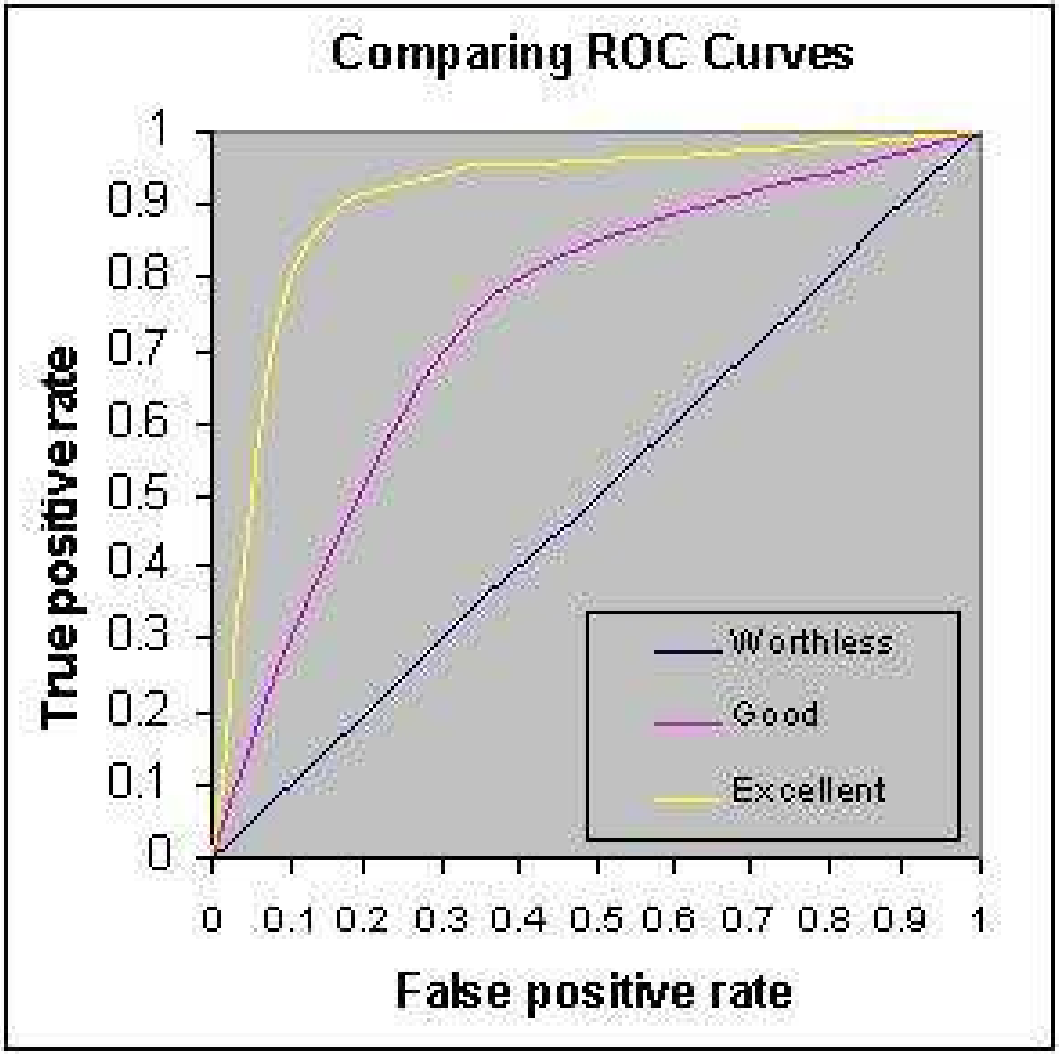
ROC curve

- A synthesis of score performance for any threshold *s*
- Using s as a parameter, the ROC curve links the true positive fraction TPF to the false positive fraction FPF





Comparing Roc Curves



Area Under Roc Curve: AUC

- Widely used measure of score performance
- Single number derived from classification rule
- Objective : no parameters to choose
- Straigthforward way for comparison
- **Theoretical AUC** = probability of concordance $P(X_1>X_2)$

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

AUC properties

- Related to **Mann-Whitney's U** statistic : $AUC = U/n_1n_2$

- Gini index: twice the area between the ROC curve and the diagonal G = 2AUC-1

3. Estimation of AUC

- 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
- Concordant pair x_1 and x_2 : $S(x_1) > S(x_2)$

- **Non parametric estimation of AUC** = $n_c / n_1 n_2$ = proportion of concordant pairs among n_1 n_2

- n_c is the Mann-Whitney's U

Standard error

- Standard error of AUC (Hanley et al., 1982) :

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

$Q_1 = AUC / (2\text{-}AUC)$ and $Q_2 = 2\,AUC^2 / (1 + AUC)$

Comparison through AUC

- As long as there is no crossing: the best method is the one with the largest AUC

- Comparison should be done taking into account the non independance of AUC estimates due to the use of the same cases

Comparison through AUC

- test statistic

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

- Somewhat complicated and too many calculations for automating the comparison

- Our proposal : empirical resampling technique

Empirical resampling

- Biased if estimated on the learning data set
- overfitting
  - **Split sample:**
    - Learning data set: estimates model parameters
    - Test data set: estimates the performance for future data
  - Performance comparisons should be done on test data
- **Resample :** Repeat the random split to avoid too specific patterns

Empirical resampling

- Calculation done on hold-out data prevent overfitting
- Test samples give unbiased estimations of AUC and its SE
- Paired comparaison test: Student, Wilcoxon,…

4. Application

- A medical data set of 768 females with eight continuous variables measured for each unit (http://ida.first.fraunhofer.de/projects/bench/benchmarks.htm).
- The response variable indicates whether or not a patient is diabetic.

Comparing LDA and logistic

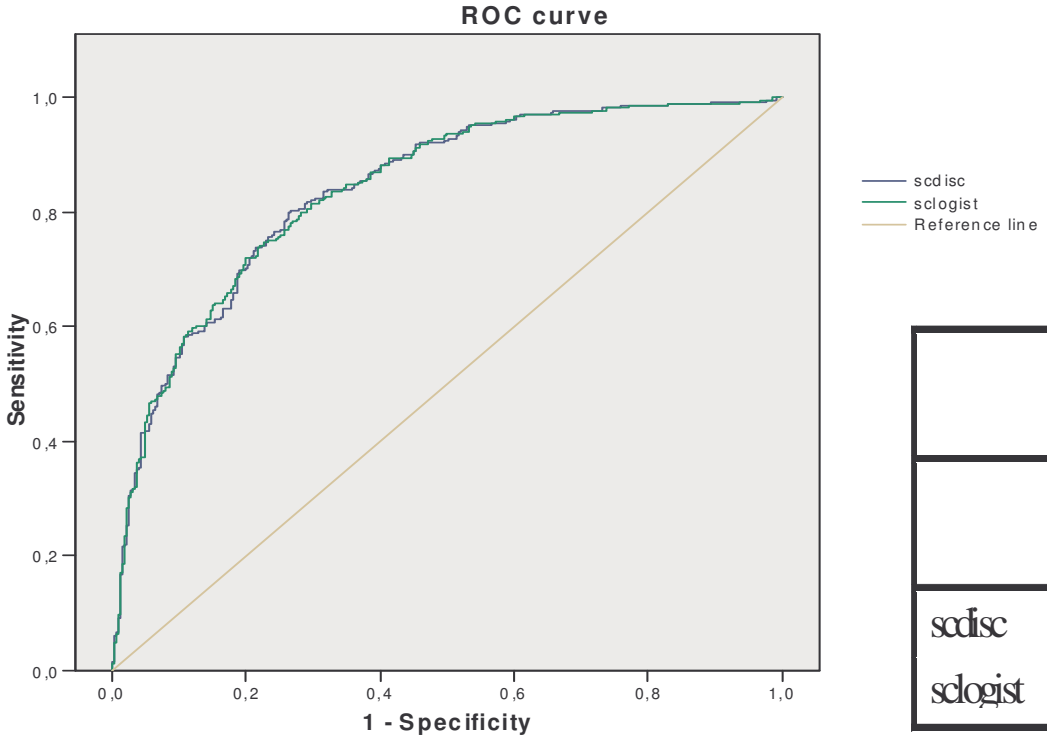
- Both techniques lead to a linear score function

$$S(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

- Estimation techniques differ:
  - Least squares in LDA
  - Conditional maximum likelihood in logistic regression.

$$P(G_1|\mathbf{x}) = \frac{\exp(S(\mathbf{x}))}{1+\exp(S(\mathbf{x}))} = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_p x_p}}$$

Performance comparisons



AUC

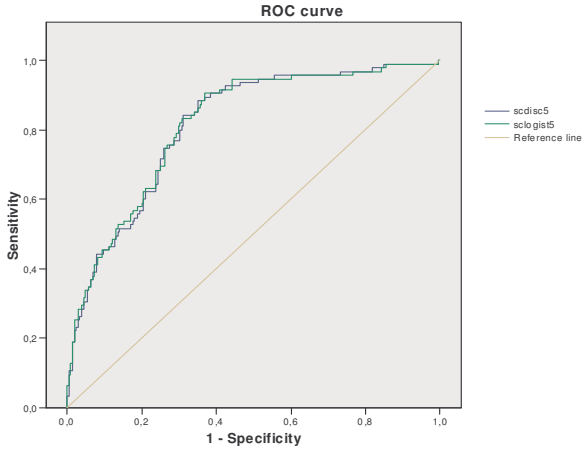
	Zone	Std. err	Asymptotic confidence interval 95%	
			Lower bound	Upper bound
scdisc	0,839	0,015	0,810	0,868
sclogist	0,839	0,015	0,811	0,868

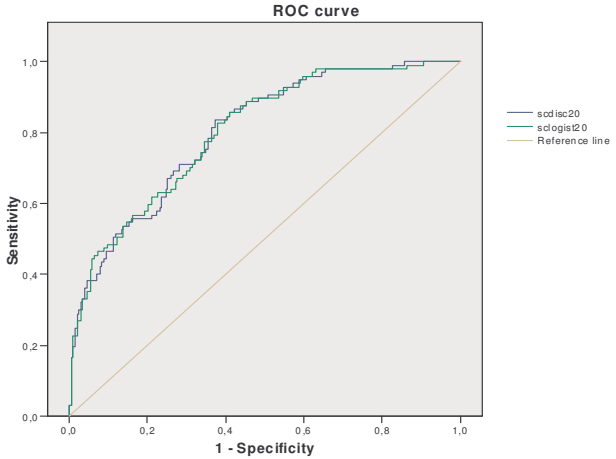
Variability

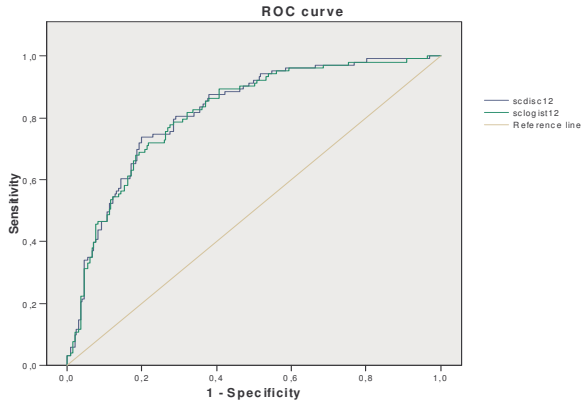
- Training data set 70%, test data set 30%, 30 times

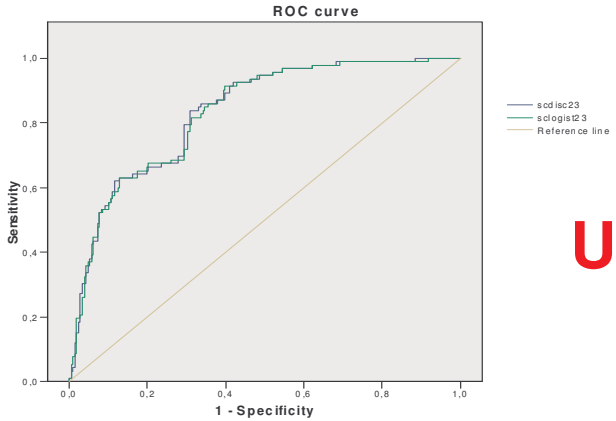
sample	LDA AUC	Logistic AUC
1	0.819	0.819
2	0.83	0.831
3	0.85	0.849
4	0.817	0.816
5	0.813	0.815
6	0.827	0.825
7	0.835	0.835
8	0.822	0.821
9	0.838	0.837
10	0.821	0.821
11	0.805	0.81
12	0.82	0.816
13	0.81	0.812
14	0.821	0.822
15	0.838	0.835
16	0.843	0.844
17	0.855	0.856
18	0.834	0.835
19	0.864	0.863
20	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

Variability









UNEXPECTED!

Comments

- Linear discriminant analysis performed as well as logistic regression
- AUC has a small but non neglectable variability
- Average AUC are lower than AUC computed on the total sample but are unbiased
- Also variability in subset selection (Saporta, Niang, 2006)

Conclusion & future work

- General approach to the problem of evaluating and comparing two class supervised methods
- Adequate and objective performance measures ROC curves and AUC
- Estimation based on resampling techniques:
  - confidence interval and hypothesis testing for comparing no crossing ROC curves
  - Unexpected variability in ROC curves.
- More comparisons needed with Cross-validation, leave-one-out, bootstrap

Conclusion & future work

- AUC is global, inadequate when curves cross
- Use partial roc curves, focus on a portion of the ROC curve
- The adequate portion depends on one's interest
- Resampling techniques can be used to estimate partial AUC (Niang & Saporta 2007)
- Last but not least important disavantage (Hand 2009):

  AUC = weighted average over thresholds, over misclassification cost ratios, but the weights depends on the classifier. Comparison with different metrics!

References

- Hand, D.J. (2006). Classifier technology and the illusion of progress (with discussion). *Statistical Science*, **21** 1-34.
- Hand, D.J. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. (To appear in Machine learning)
- 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.
- Saporta, G. and Niang, N. (2006) Model assessment. In *KNEMO: Knowledge Extraction and Modeling, Capri,4-6 septembre*,. IASC-INTERFACE-IFCS Workshop.