Estimating the ROC Curve of Linearly Combined Dichotomizers

Claudio Marrocco, Mario Molinara, and Francesco Tortorella

Dipartimento di Automazione, Elettromagnetismo,
Ingegneria dell'Informazione e Matematica Industriale,
Università degli Studi di Cassino,
03043 Cassino (FR), Italy
{c.marrocco, m.molinara, tortorella}@unicas.it

Abstract. A well established technique to improve the classification performances is to combine more classifiers. In the binary case, an effective instrument to analyze the dichotomizers under different class and cost distributions providing a description of their performances at different operating points is the Receiver Operating Characteristic (ROC) curve. To generate a ROC curve, the outputs of the dichotomizers have to be processed. An alternative way that makes this analysis more tractable with mathematical tools is to use a parametric model and, in particular, the binormal model that gives a good approximation to many empirical ROC curves. Starting from this model, we propose a method to estimate the ROC curve of the linear combination of two dichotomizers given the ROC curves of the single classifiers. A possible application of this approach has been successfully tested on real data set.

1 Introduction

Dichotomizers (i.e. two-class classifiers) are used in many critical applications (e.g., automated diagnosis, fraud detection, currency verification) which require highly discriminating classifiers. In order to improve the classification performance a well established technique is to combine more classifiers so as to take advantage of the strengths of the single classifiers and avoid their weaknesses. To this aim, a huge number of possible combination rules have been proposed up to now which generally try to decrease the classification error. However, the applications considered frequently involve cost matrices and class distributions both strongly asymmetric and dynamic and in such cases the overall error rate, usually employed as a reference performance measure in classification problems, is not a suitable metric for evaluating the quality of the classifier [1].

A more effective tool for correctly quantifying the accuracy and analyzing the dichotomizer under different class and cost distributions is given by the Receiver Operating Characteristic (ROC) curve. It provides a description of the performance of the dichotomizer at different operating points, which is independent of the prior probabilities of the two classes. ROC analysis is based in statistical decision theory and was first employed in signal detection problems [2]; it is now common in medical diagnosis and particularly in medical imaging [3]. In the Pattern Recognition field, ROC analysis is increasingly adopted for many central issues such as the evaluation

F. Roli and S. Vitulano (Eds.): ICIAP 2005, LNCS 3617, pp. 778 – 785, 2005.

of machine learning algorithms [4], the robust comparison of classifier performance under imprecise class distribution and misclassification costs [5] and the definition of a reject option for dichotomizers [6]. In this framework, the analysis is commonly performed on an empirical ROC curve which is generated by processing the outputs provided by the dichotomizer on a labeled data set (an efficient algorithm is described in [7]). An alternative way is to use a parametric model for the ROC curve so as to make the analysis more tractable with mathematical tools. A parametric model which gives a good approximation to many empirical ROC curves is the *binormal model*, which assumes that a pair of latent normal decision-variable distributions underlies ROC data [2, 8]. Such model plays a central role in ROC analysis, similar to the normal distribution in statistics [9].

On the basis of this model, we propose a method for estimating the ROC curve of the linear combination of two dichotomizers given the ROC curves of the single classifiers. This is a useful result since it makes possible to have an immediate preview of the performance of the system obtained by applying the combination without evaluating the outputs on the samples of the data set.

In the next section we present a short description of the binormal model and the method to obtain a parametric model of the ROC curve of the combination. The conclusive section describes a possible application of the method and shows some results obtained from experiments performed on real data sets.

2 Linear Combination of Two Classifier Based on ROC Curve

The ROC Curve correlates the *True Positive Rate* (TPR) with the *False Positive Rate* (FPR), i.e. the fraction of positive cases correctly classified with the fraction of negative cases incorrectly classified as positive. Let us define P and N as, respectively, the positive and the negative class in a two-class problem, for each fixed threshold value t, it is possible to define the rates as:

$$FPR(t) = \#\{x \in N \mid y(x) \geq t\} / \#N \qquad TPR(t) = \#\{x \in P \mid y(x) \geq t\} / \#P \tag{1}$$

where $\boldsymbol{x}$ is the considered sample and $y(\boldsymbol{x})$ is the output of the classifier. Let us, now, consider these quantities as probabilities; in this way it is possible to write:

$$\begin{aligned} FPR(t) &\cong \Pr\{y(x) \geq t \mid x \in N\} = 1 - \Pr\{y(x) < t \mid x \in N\} = 1 - F_{y|N}(t) \\ TPR(t) &\cong \Pr\{y(x) \geq t \mid x \in P\} = 1 - \Pr\{y(x) < t \mid x \in P\} = 1 - F_{y|P}(t) \end{aligned} \tag{2}$$

where $F_{y|N}(t)$ e $F_{y|P}(t)$ are two Cumulative Distribution Functions (CDF) characterizing the output of the classifier conditionally to the class of the sample x.

The binormal model is a parametric model for the ROC curve which assumes that the ROC data arise from a pair of latent normal distributions with:

$$\boldsymbol{F}_{y|N}(\boldsymbol{t}) = \Phi(\boldsymbol{t}) \quad , \quad \boldsymbol{F}_{y|P}(\boldsymbol{t}) = \Phi(\boldsymbol{bt - a}) \tag{3}$$

where a and b are two constants and $\Phi(t)$ is the CDF of a Gaussian distribution with zero mean and unit variance. Therefore:

$$FPR(t) = 1 - \Phi(t) \quad , \quad TPR(t) = 1 - \Phi(bt - a) \tag{4}$$

A method to obtain a maximum-likelihood estimate of the parametric model from continuously distributed data has been proposed by Metz et al. in [8].

Let us now examine the classifier obtained by the linear combination of two dichotomizers: if y_1 and y_2 are the outputs of the two classifiers, the output of the combined classifier is given by $z = \alpha_1 y_1 + \alpha_2 y_2$ where α_1 and α_2 are suitably chosen weights. The TP and FP rates of this classifier are:

$$\begin{aligned} FPR_Z(t) &= \Pr\{\alpha_1 y_1 + \alpha_2 y_2 \geq t \mid x \in \mathrm{N}\} = 1 - \Pr\{\alpha_1 y_1 + \alpha_2 y_2 < t \mid x \in \mathrm{N}\} = \\ &= 1 - F_{\alpha_1 y_1 + \alpha_2 y_2 \mid \mathrm{N}}(t) \\ TPR_Z(t) &= \Pr\{\alpha_1 y_1 + \alpha_2 y_2 \geq t \mid x \in \mathrm{P}\} = 1 - \Pr\{\alpha_1 y_1 + \alpha_2 y_2 < t \mid x \in \mathrm{P}\} = \\ &= 1 - F_{\alpha_1 y_1 + \alpha_2 y_2 \mid P}(t) \end{aligned} \tag{5}$$

The two CDFs $\boldsymbol{F}_{\alpha_1 y_1 + \alpha_2 y_2 \mid \mathrm{N}}(t)$ and $\boldsymbol{F}_{\alpha_1 y_1 + \alpha_2 y_2 \mid P}(t)$ can be obtained from the CDFs of the single classifiers by means of the Stieltjes integral [10]:

$$\boldsymbol{F}_{\alpha_1 y_1 + \alpha_2 y_2 \mid \mathrm{N}}(t) = \int_{-\infty}^{+\infty} \boldsymbol{F}_{y_2 \mid \mathrm{N}}\left(\frac{t - \alpha_1 \tau}{\alpha_2}\right) d\boldsymbol{F}_{y_1 \mid \mathrm{N}}(\tau) \quad , \quad \boldsymbol{F}_{\alpha_1 y_1 + \alpha_2 y_2 \mid \mathrm{P}}(t) = \int_{-\infty}^{+\infty} \boldsymbol{F}_{y_2 \mid \mathrm{P}}\left(\frac{t - \alpha_1 \tau}{\alpha_2}\right) d\boldsymbol{F}_{y_1 \mid \mathrm{P}}(\tau) \tag{6}$$

Let us now assume that the binormal models of the ROC curves of the base classifiers are available and let their parameters be (a_1,b_1) and (a_2,b_2). This allows us to obtain a closed form of the integrals in eq. (6). To this aim, let us consider eq. (6) in terms of the corresponding density functions:

$$\begin{aligned} f_{\alpha_1 y_1 + \alpha_2 y_2 \mid \mathrm{N}}(t) &= \int_{-\infty}^{+\infty} \frac{1}{\alpha_2} f_{y_2 \mid \mathrm{N}}\left(\frac{t - \alpha_1 \tau}{\alpha_2}\right) f_{y_1 \mid \mathrm{N}}(\tau) d\tau \\ f_{\alpha_1 y_1 + \alpha_2 y_2 \mid \mathrm{P}}(t) &= \int_{-\infty}^{+\infty} \frac{1}{\alpha_2} f_{y_2 \mid \mathrm{P}}\left(\frac{t - \alpha_1 \tau}{\alpha_2}\right) f_{y_1 \mid \mathrm{P}}(\tau) d\tau \end{aligned} \tag{7}$$

Taking into account that

$$\begin{aligned} f_{y_1|N}(t) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \quad f_{y_1|P}(t) = \frac{b_1}{\sqrt{2\pi}} \exp\left(-\frac{(b_1 t - a_1)^2}{2}\right) \\ f_{y_2|N}(t) &= \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{t^2}{2}\right) \quad f_{y2|P}(t) = \frac{b_2}{\sqrt{2\pi}} \exp\left(-\frac{(b_2 t - a_2)^2}{2}\right) \end{aligned} \tag{8}$$

after some algebraic manipulations, we finally obtain:

$$f_{\alpha_1 y_1 + \alpha_2 y_2 | N}(t) = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{\alpha_1^2 + \alpha_2^2}} \exp\left(-\frac{t^2}{2\alpha_2^2}\left(1 - \frac{\alpha_1^2}{\alpha_1^2 + \alpha_2^2}\right)\right)$$

$$\begin{aligned} f_{\alpha_1 y_1 + \alpha_2 y_2 | P}(t) &= \frac{b_2 b_1}{\sqrt{2\pi}\sqrt{b_2^2 \alpha_1^2 + b_1^2 \alpha_2^2}} \exp\left(-\frac{b_2^2 t^2 - 2a_2 b_2 \alpha_2 t + \alpha_2^2 \left(a_1^2 + a_2^2\right)}{2\alpha_2^2}\right) \cdot \\ &\cdot \exp\left(\frac{\left(b_2^2 \alpha_1 t - a_2 b_2 \alpha_1 \alpha_2 + a_1 b_1 \alpha_2^2\right)}{2\alpha_2^2 \left(b_2^2 \alpha_1^2 + b_1^2 \alpha_2^2\right)}\right) \end{aligned} \tag{9}$$

The corresponding CDFs can now be determined:

$$F_{\alpha_1 y_1+\alpha_2 y_2|N}(t)=\frac{1}{2}\left[erf\left(\frac{t}{\sqrt{2\left(\alpha_1^{\,2}+\alpha_2^{\,2}\right)}}\right)+1\right]$$
$$F_{\alpha_1 y_1+\alpha_2 y_2|P}(t)=\frac{1}{2b}\left[erf\left(\frac{\sqrt{2}}{2}\frac{1}{\sqrt{b_2^{\,2}\alpha_1^{\,2}+b_1^{\,2}\alpha_2^{\,2}}}\left(b_1b_2t-\left(\alpha_2a_2b_1+\alpha_1a_1b_2\right)\right)\right)+1\right] \tag{10}$$

Finally, from eq. (5) we obtain

$$FPR_z(t)=1-\frac{1}{2}\left[erf\left(\frac{t}{\sqrt{2\left(\alpha_1^{\,2}+\alpha_2^{\,2}\right)}}\right)+1\right]$$
$$TPR_z(t)=1-\frac{1}{2b}\cdot\left[erf\left(\frac{\sqrt{2}}{2}\frac{1}{\sqrt{b_2^{\,2}\alpha_1^{\,2}+b_1^{\,2}\alpha_2^{\,2}}}\left(b_1b_2t-\left(\alpha_2a_2b_1+\alpha_1a_1b_2\right)\right)\right)+1\right] \tag{11}$$

3 Experimental Results

The proposed method can be applied in many situations to be faced when linearly combining two dichotomizers. For example, if the application at hand is cost-sensitive it could be useful to know the performance of the built classification system in some regions of the ROC curve [5]. Another important operation which can take advantage of our method is the evaluation of the area under the ROC curve [11] which provides an efficient way to measure the ranking quality of the classifier obtained by combining the two dichotomizers.

In this paper we have evaluated the proposed method on the latter problem. We have employed two dichotomizers: a Support Vector Machine (SVM) with linear kernel and a Multi Layer Perceptron (MLP) with five units in the hidden layer. The former has been implemented by means of SVM^{light} tool (available at http://svmlight.joachims.org) while for the latter we used the NODElib library [12]. The training of the MLP has been performed on 10000 epochs using the back propagation algorithm with a learning rate of 0.01.

The datasets used are publicly available at the UCI Machine Learning Repository [13]; all of them have two classes and a variable number of numerical input features. The features were previously rescaled so as to have zero mean and unit standard deviation. More details are given in table 1.

Table 1. Datasets used in the experiments

Data Set	# Feature	# Sample	% Positive	% Negative	Train Set	Test Set	Valid Set
Pima Indian Diabetes	8	768	34.9	65.1	384	192	192
German Credit	24	1000	30	70	500	250	250
Contraceptive Method Choice	9	1473	57.3	42.7	737	368	368

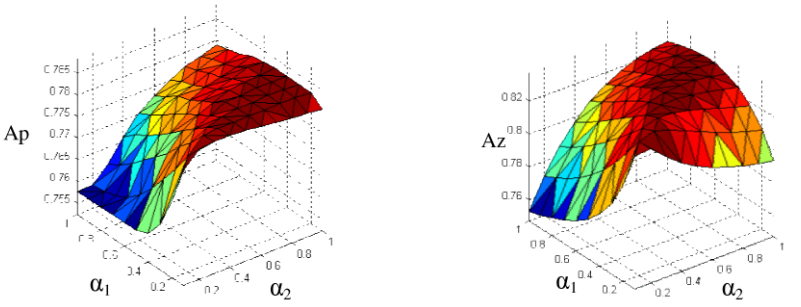
The classifier resulting from the linear combination of the two dichotomizers has output $z=\alpha_{SVM}y_{SVM}+\alpha_{MLP}y_{MLP}$, where y_{SVM} and y_{MLP} are the outputs of the two classifiers and α_{SVM} and α_{MLP} are the relative weights. We have considered a hundred of different combinations, obtained varying each of the weights from zero to one with a 0.1 step. For each dataset, once the parameters (a,b) of the binormal model of the ROC curve of each dichotomizer have been obtained by means of the method described in [8], we have generated the parametric ROC curve of the built classifier for each combination of the weights by applying eq. (11). At the same time, an empirical estimate of the same curves have been generated by means of an algorithm described in [7]. For each combination, the area under the empirical curve (Az) and the area under the parametric curve (Ap) have been compared. To avoid any bias in the comparison, 12 runs of a multiple hold-out procedure were performed on all datasets. In each run, the dataset was split in three subsets: a training set (containing 50% of the samples of each class) used in the learning phase of the dichotomizers, a validation set and a test set (each containing 25% of the samples of each class); the final size of each of these sets is given in table 1. For each run, the parameters of the binormal model were estimated on the validation set, while the empirical ROC curve was obtained from the test set. The results of the comparison for the three datasets are presented in figs 1-3 and in tables 2-4. Each cell of the tables contains a value corresponding to the mean plus or minus the standard deviation of the area under the parametric and the empirical curves relative to a particular weights combination; for the sake of readability only half of the studied combinations are shown. The figures plot the mean values of Az and Ap for all the considered combinations.

In all cases the obtained results show that the points of maximum and minimum of Az correspond to the points of maximum and minimum of Ap, thus our method is able to individuate the better and the worst couple of weights of the linear combination. In this way, it is possible to make a faster computation of all the combination without evaluating the empirical ROC and so to choice the better combination without generating the ROC curve of the combined classifiers.

However, some problems can arise when the binormal model does not fit satisfactorily the empirical ROC curve. This happens when the dichotomizer does not perform very well and the empirical ROC curve presents some concavities which the binormal model is not able to fit. Some examples are shown in fig. 4.

Table 2. Results on German Credit dataset for a linear combination of an SVM with linear kernel and an MLP

α_{MLP} \ α_{SVM}		0.2	0.4	0.6	0.8	1.0
0.2	Az	**0.7789**±0.0004	**0.7827**±0.0004	**0.7829**±0.0004	**0.7829**±0.0004	**0.7826**±0.0004
	Ap	**0.8280**±0.0011	**0.8287**±0.0005	**0.8204**±0.0003	**0.8132**±0.0003	**0.8074**±0.0002
0.4	Az	**0.7701**±0.0005	**0.7789**±0.0004	**0.7825**±0.0004	**0.7827**±0.0003	**0.7827**±0.0003
	Ap	**0.8064**±0.0018	**0.8282**±0.0011	**0.8313**±0.0006	**0.8286**±0.0005	**0.8242**±0.0004
0.6	Az	**0.7628**±0.0006	**0.7743**±0.0004	**0.7789**±0.0004	**0.7811**±0.0004	**0.7822**±0.0004
	Ap	**0.7898**±0.0020	**0.8173**±0.0016	**0.8282**±0.0011	**0.8310**±0.0007	**0.8298**±0.0006
0.8	Az	**0.7589**±0.0007	**0.7701**±0.0005	**0.7763**±0.0005	**0.7789**±0.0004	**0.7805**±0.0004
	Ap	**0.7787**±0.0020	**0.8063**±0.0018	**0.8208**±0.0014	**0.8274**±0.0011	**0.8290**±0.0008
1.0	Az	**0.7578**±0.0008	**0.7661**±0.0006	**0.7728**±0.0005	**0.7767**±0.0004	**0.7789**±0.0004
	Ap	**0.7708**±0.0019	**0.7968**±0.0020	**0.8125**±0.0017	**0.8210**±0.0013	**0.8245**±0.0010

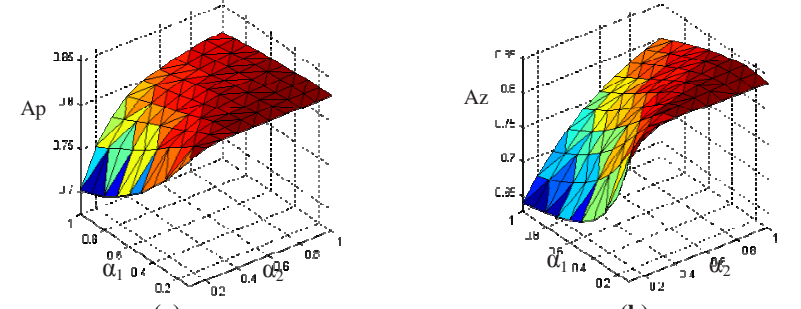


(a) (b)

Fig. 1. Plot of the mean values of the area under the empirical **(a)** and the parametric **(b)** ROC curve for all the considered combination on the German Credit dataset for a linear combination of an SVM with linear kernel and an MLP

Table 3. Results on Pima Indian Diabetes dataset for a linear combination of an SVM with linear kernel and an MLP

α_{MLP} \ α_{SVM}		0.2	0.4	0.6	0.8	1
0.2	Az	**0.8188**±0.0028	**0.8268**±0.0027	**0.8289**±0.0025	**0.8296**±0.0025	**0.8297**±0.0025
	Ap	**0.8016**±0.0022	**0.8348**±0.0023	**0.8406**±0.0023	**0.8414**±0.0023	**0.8408**±0.0023
0.4	Az	**0.7993**±0.0026	**0.8188**±0.0028	**0.8251**±0.0028	**0.8268**±0.0027	**0.8281**±0.0027
	Ap	**0.7410**±0.0017	**0.8016**±0.0022	**0.8252**±0.0023	**0.8347**±0.0023	**0.8385**±0.0023
0.6	Az	**0.7785**±0.0024	**0.8086**±0.0027	**0.8188**±0.0028	**0.8242**±0.0029	**0.8261**±0.0027
	Ap	**0.7057**±0.0014	**0.7677**±0.0019	**0.8016**±0.0022	**0.8195**±0.0022	**0.8288**±0.0022
0.8	Az	**0.7616**±0.0022	**0.7993**±0.0026	**0.8120**±0.0027	**0.8188**±0.0028	**0.8233**±0.0029
	Ap	**0.6845**±0.0012	**0.7410**±0.0017	**0.7781**±0.0020	**0.8012**±0.0021	**0.8150**±0.0022
1	Az	**0.7460**±0.0018	**0.7887**±0.0025	**0.8050**±0.0026	**0.8137**±0.0027	**0.8188**±0.0028
	Ap	**0.6706**±0.0011	**0.7208**±0.0015	**0.7575**±0.0018	**0.7829**±0.0020	**0.7996**±0.0021



(a) (b)

Fig. 2. Plot of the mean values of the area under the empirical **(a)** and the parametric **(b)** ROC curve for all the considered combination on the Pima Indian Diabetes dataset for a linear combination of an SVM with linear kernel and an MLP

Table 4. Results on Contraceptive Method Choice dataset for a linear combination of an SVM with linear kernel and an MLP

α_{MLP} \ α_{SVM}		0.2	0.4	0.6	0.8	1.0
0.2	Az	**0.7280**±0.0011	**0.7257**±0.0010	**0.7225**±0.0009	**0.7200**±0.0008	**0.7188**±0.0008
	Ap	**0.7768**±0.0014	**0.7669**±0.0011	**0.7553**±0.0010	**0.7471**±0.0010	**0.7413**±0.0009
0.4	Az	**0.7250**±0.0013	**0.7280**±0.0011	**0.7274**±0.0010	**0.7257**±0.0010	**0.7240**±0.0009
	Ap	**0.7635**±0.0015	**0.7768**±0.0014	**0.7734**±0.0012	**0.7668**±0.0011	**0.7603**±0.0011
0.6	Az	**0.7220**±0.0013	**0.7270**±0.0013	**0.7280**±0.0011	**0.7277**±0.0011	**0.7268**±0.0010
	Ap	**0.7497**±0.0015	**0.7716**±0.0015	**0.7768**±0.0014	**0.7751**±0.0013	**0.7707**±0.0012
0.8	Az	**0.7192**±0.0013	**0.7250**±0.0013	**0.7275**±0.0012	**0.7280**±0.0011	**0.7279**±0.0011
	Ap	**0.7400**±0.0015	**0.7635**±0.0015	**0.7738**±0.0014	**0.7764**±0.0013	**0.7748**±0.0013
1.0	Az	**0.7172**±0.0013	**0.7234**±0.0013	**0.7263**±0.0013	**0.7278**±0.0012	**0.7280**±0.0011
	Ap	**0.7331**±0.0014	**0.7558**±0.0015	**0.7684**±0.0015	**0.7738**±0.0014	**0.7746**±0.0013

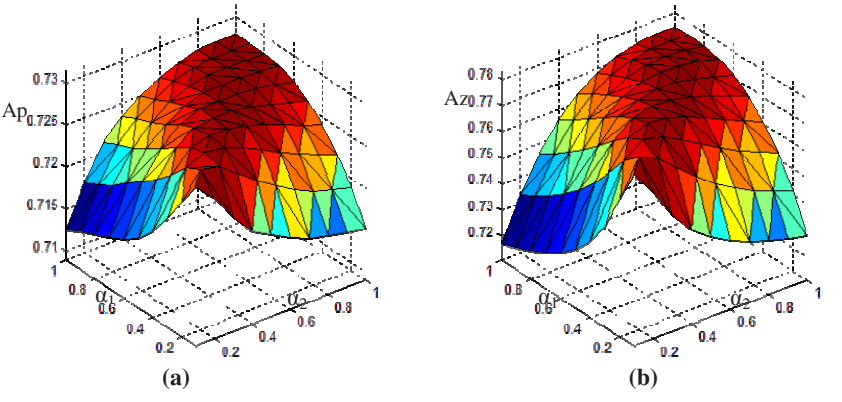


Fig. 3. Plot of the mean values of the area under the empirical **(a)** and the parametric **(b)** ROC curve for all the considered combination on the Contraceptive Method Choice dataset for a linear combination of an SVM with linear kernel and an MLP

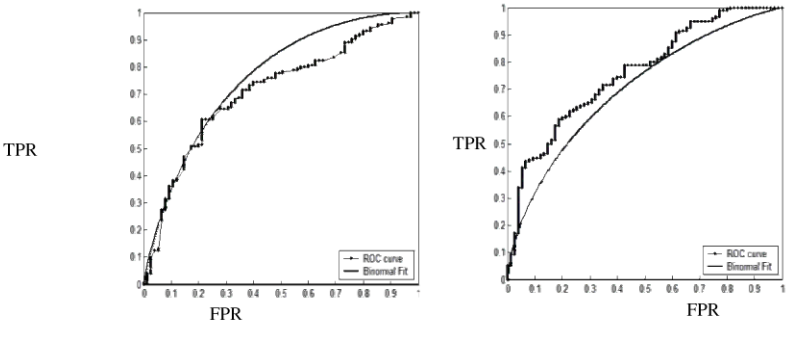


Fig. 4. Two examples of ROC curve with concavities: the fitting model does not perform well

Starting from these results, future developments of our work will examine both other algorithms for the fitting of the binormal model and other parametric models for the ROC curve.

References

[1] Provost, F., Fawcett, T., Kohavi, R.: The case against accuracy estimation for comparing induction algorithms. Proc. 15th Intl. Conf. on Machine Learning, Morgan Kaufmann (1998), 445-453.

[2] Egan, J.P.: Signal detection theory and ROC analysis. Series in Cognition and Perception, Academic Press, New York (1975).

[3] Metz, C.E., ROC methodology in radiologic imaging. Invest. Radiol. 21 (1986), 720-733.

[4] Bradley, A.P.: The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition 30 (1997), 1145-1159.

[5] Provost, F., Fawcett, T., Robust classification for imprecise environments. Machine Learning 42 (2001), 203-231.

[6] Tortorella, F.: A ROC-based Reject Rule for Dichotomizers. Pattern Recognition Letters 26 (2005), 167-180.

[7] Fawcett, T.: ROC graphs: notes and practical considerations for data mining researchers. HP Labs Tech Report HPL-2003-4 (2003).

[8] Metz, C.E., Herman, B.A., Shen, J.H.: Maximum-likelihood estimation of ROC curves from continuously-distributed data. Statistics in Medicine 17 (1998), 1033-1053.

[9] Pepe, M.S. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford Statistical Science Series, Oxford University Press (2003)

[10] Papoulis, A., Probability, Random Variables, and Stochastic Processes. McGraw-Hill, New York (2001)

[11] Cortes, C., Mohri, M.: AUC Optimization vs. Error Rate Minimization. Advances in Neural Information Processing Systems, NIPS 2003, (2004).

[12] Flake, G.W., Pearlmuter, B.A.: Differentiating Functions of the Jacobian with Respect to the Weights. In S. A. Solla, T. K. Leen, and K. Müller, eds., Advances in Neural Information Processing Systems, vol. 12, The MIT Press (2000).

[13] Blake, C., Keogh, E., Merz, C.J.: UCI Repository of Machine Learning Databases. (1998) [www.ics.uci.edu/~mlearn/MLRepository.html]

[14] Metz, C.E., Pan, X.: Proper binormal ROC curves: theory and maximum-likelihood estimation. J Math Psych 43 (1999), 1-33.