Tree-structured ranking rules and approximation of the optimal ROC curve

Stéphan Clémençon, Nicolas Vayatis

Tree-structured ranking rules and approximation of the optimal ROC curve

Stéphan Clémençon
Telecom Paristech (TSI) - LTCI UMR Institut Telecom/CNRS 5141
stephan.clemencon@telecom-paristech.fr

Nicolas Vayatis
ENS Cachan & UniverSud - CMLA UMR CNRS 8536
vayatis@cmla.ens-cachan.fr

Abstract

We consider the extension of standard decision tree methods to the *bipartite ranking* problem. In ranking, the goal pursued is global: define an order on the whole input space in order to have positive instances on top with maximum probability. The most natural way of ordering all instances consists in projecting the input data x onto the real line using a real-valued *scoring function* s and the accuracy of the ordering induced by a candidate s is classically measured in terms of the AUC. In the paper, we discuss the design of tree-structured scoring functions obtained by maximizing the AUC criterion. In particular, the connection with recursive piecewise linear approximation of the optimal ROC curve both in the L_1-sense and in the L_∞-sense is discussed.

1 Introduction

The statistical ranking problem can broadly be considered as the problem of ordering instances from an abstract space or a high-dimensional Euclidean space. A natural approach consists of "projecting" these instances onto the real line through some real-valued scoring function. Such a function would allow to rank any list of instances in the initial space. Depending on the available information, various approaches can be developed. For instance, both preference learning ([HGBSO98], [CSS98], [dEW06]) and ordinal regression ([HGO00], [CS01]) deal with statistical ranking but under different label information. We focus here on the setup where a binary label characterizing each instance is given. This problem is known as the *bipartite ranking problem* ([FISS03], [AGH$^+$05], [CLV05]). The calibration of ranking rules can be performed in various ways. In scoring applications, the standard approach is mostly in the spirit of logistic regression and relies on the statistical modeling of the regression function using additive models ([HT90]). The statistical learning approach is different insofar as it avoids the difficult problem of estimating the distribution in high dimensions and focuses on prediction. Statistical learning strategies can be thought as the optimization of performance measures based on data. In the case of bipartite ranking, the development of the statistical learning approach is involved with AUC maximization. Indeed, a standard performance measure for a scoring function in the presence of classification data is the Receiver Operating Characteristic (ROC) curve, together with the Area Under the ROC Curve, known as the AUC (see [DS66], [vT68], [Ega75], [HM82]). But, since their introduction, ROC curves and the AUC used to serve mostly for validation and not as the basis for optimization principles. More recently, several aspects of AUC maximization have been discussed in the machine learning literature ([CM04], [Rak04], [YDMW03]) and also from a statistical learning perspective ([AGH$^+$05], [CLV05], [CLVar]). A particular class of learning algorithms will be at the center of the present paper, namely decision trees in the spirit of CART [BFOS84]. The investigation of decision trees in the context of ranking was initiated only recently in the field of machine learning ([FFHO02], [PD03], [XZW06]).

In the present work, we consider the problem of recovering the optimal ROC curve from the perspective of *approximation theory*, in conjunction with the one of adaptively building a scoring function from training data with a ROC curve close to the optimal approximate version. Our primary goal is to relate linear-by-parts approximations with finite-dimensional (piecewise constant) approximations of optimal scoring functions. As the ROC curve provides a performance measure of functional nature, the approximation can be conceived in a variety of ways depending on the topology equipping the space of ROC curves. For instance, the AUC is related to the L_1-distance but we will also consider convergence to the optimal ROC curve in a stronger sense carried by the L_∞-distance. A recursive implementation of the approximation procedure naturally leads to a tree-like structure for underlying scoring functions. We suggest that such a tree-based ranker could serve as a weak learner and feed a boosting-type algorithm such as RankBoost ([FISS03]). We also provide mathematical results in terms of the approximation error and statistical consistency.

The paper is organized as follows. In Section 2, we present a general approach for assessing optimality in the bipartite ranking problem. We also recall the main concepts and discuss the issue of AUC maximization. In Section 3, we discuss the approximation of the optimal ROC curve with piecewise constant scoring functions and provide an adaptive tree-structured recursive procedure for which an approximation error result is established. This approximation scheme can be carried out over empirical data by the means of the TREERANK algorithm described in Section 4. The statisti-

cal consistency of the method is also studied. All proofs are postponed to the Appendix section.

2 The nature of the ranking problem

We start off by describing the optimal elements for the bipartite ranking problem ([FISS03]). The use of the ROC curve as a performance measure for bipartite ranking is then strongly advocated by this enlightening approach, under which the problem boils down to recovering the collection of *level sets* of the regression function.

2.1 Setup and goal of ranking.

We study the ranking problem for classification data with binary labels. This is also known as the bipartite ranking problem. The data are assumed to be generated as copies of a random pair $(X, Y) \in \mathcal{X} \times \{-1, +1\}$ where X is a random descriptor living in the measurable space $\mathcal{X}$ and Y represents its binary label (relevant vs. irrelevant, healthy vs. sick, ...). We denote by $P = (\mu, \eta)$ the distribution of (X, Y), where μ is the marginal distribution of X and η is the *regression function* (up to an affine transformation): $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$, $x \in \mathcal{X}$. We will also denote by $p = \mathbb{P}\{Y = 1\}$ the proportion of positive labels. In the sequel, we assume that the distribution μ is absolutely continuous with respect to Lebesgue measure.

The goal of a ranking procedure is to provide an ordering of the elements of $\mathcal{X}$ based on their labels. We expect to end up with a list with positive labels at the top and negative labels at the bottom. However, label information does not permit to derive a total order on $\mathcal{X}$ and among relevant (positively labelled) objects in $\mathcal{X}$, some might be more relevant than others. In short, a good ranking should preserve the ordering induced by the likelihood of having a positive label, namely the regression function η. We consider the approach where the ordering can be derived by the means of a *scoring function* $s : \mathcal{X} \to \mathbb{R}$. The following definition sets the goal of learning methods in the setup of bipartite ranking.

Definition 1 (Optimal scoring functions) *A scoring function $s^* : \mathcal{X} \to \mathbb{R}$ is said to be optimal if it induces the same ordering over $\mathcal{X}$ as the function $\eta(x) = \mathbb{P}\{Y = 1 \mid X = x\}$, $\forall x \in \mathcal{X}$. In other words:*

$$\forall x, x' \in \mathcal{X}\ , \quad s^*(x) - s^*(x') > 0 \Rightarrow \eta(x) - \eta(x') > 0\ .$$

According to the previous definition, the next proposition is a trivial characterization of the class of optimal scoring functions.

Proposition 2 *The class of optimal scoring functions is given by the set*

$$\mathcal{S}^* = \{\ s^* = T \circ \eta\ \mid\ T : [0,1] \to \mathbb{R} \textit{ strictly increasing}\ \}.$$

Interestingly, it is possible to make the connection between an arbitrary (bounded) optimal scoring function $s^* \in \mathcal{S}^*$ and the distribution P (through the regression function η) completely explicit.

Proposition 3 (Optimal scoring functions representation) *A bounded scoring function s^* is optimal if and only if there exist a nonnegative integrable function w and a continuous random variable V in $(0,1)$ such that:*

$$\forall x \in \mathcal{X}\ , \quad s^*(x) = \inf_{\mathcal{X}} s^* + \mathbb{E}\left(w(V) \cdot \mathbb{I}\{\eta(x) > V\}\right)$$

Remark 1 In the case of the regression function η, we have the following identity :

$$\forall x \in \mathcal{X}\ , \quad \eta(x) = \mathbb{E}\left(w(U)\mathbb{I}\{\eta(x) > U\}\right)$$

where U is a uniform random variable on $[0,1]$ and the function w is the indicator of the support of the random variable $\eta(X)$.

A crucial consequence of the last proposition is that solving the bipartite ranking problem amounts to recovering the collection $\{x \in \mathcal{X} \mid \eta(x) > u\}_{u \in (0,1)}$ of level sets of the regression function η. Hence, the bipartite ranking problem can be seen as a collection of *overlayed classification problems*. This view was first introduced in [CV07]. Moreover, the representation of optimal scoring functions provides the intuition for the approximation procedure of Section 3 and the subsequent TREERANK algorithm of Section 4. By checking the proof of the Proposition, it looks like the weight function w only plays the role of a scaling function. However, the general representation may suggest various estimations schemes of the Monte-Carlo type in order to recover optimal scoring functions.

2.2 (True) ROC curves

We now recall the concept of ROC curve and explain why it is a natural choice of performance measure for the ranking problem with classification data. In this section, we only consider *true* ROC curves which correspond to the situation where the underlying distribution is known.

Before recalling the definition, we need to introduce some notations. For a given scoring rule s, the conditional cdfs of the random variable $s(X)$ are denoted by G_s and H_s. We also set, for all $z \in \mathbb{R}$:

$$\bar{G}_s(z) = 1 - G_s(z) = \mathbb{P}\left\{s(X) > z \mid Y = +1\right\}\ ,$$
$$\bar{H}_s(z) = 1 - H_s(z) = \mathbb{P}\left\{s(X) > z \mid Y = -1\right\}\ .$$

to be the residual conditional cdfs of the random variable $s(X)$. When $s = \eta$, we shall denote the previous functions by G^*, H^*, $\bar{G}^*$, $\bar{H}^*$ respectively. We will also use the notation, for all t:

$$\alpha(t) = \bar{H}^*(t) = \mathbb{P}\{\eta(X) > t \mid Y = -1\}\ ,$$
$$\beta(t) = \bar{G}^*(t) = \mathbb{P}\{\eta(X) > t \mid Y = 1\}\ .$$

We introduce the notation $Q(Z, \alpha)$ to denote the quantile of order $1 - \alpha$ for the distribution of a random variable Z conditioned on the event $Y = -1$. In particular, the following quantile will be of interest:

$$Q^*(\alpha) = Q(\eta(X), \alpha) = \bar{H}^{*-1}(\alpha)\ ,$$

where we have used here the notion of generalized inverse F^{-1} of a càdlàg function F:

$$F^{-1}(z) = \inf\{t \in \mathbb{R} \mid F(t) \geq z\}\ .$$

A classical way to assess the performance of a scoring function s in separating the two populations (positive vs. negative labels) is the *Receiver Operating Characteristic* known as the ROC curve ([vT68], [Ega75]).

Definition 4 (True ROC curve) *The ROC curve of a scoring function s is the parametric curve:*

$$z \mapsto \left(\bar{H}_s(z), \bar{G}_s(z)\right)$$

for thresholds $z \in \mathbb{R}$. It can also be defined as the plot of the function:

$$\alpha \in [0,1] \mapsto \bar{G}_s \circ \bar{H}_s^{-1}(\alpha) = \bar{G}_s\left(Q(s(X), \alpha)\right) = \mathrm{ROC}(\mathrm{s}, \alpha).$$

By convention, points of the curve corresponding to possible jumps (due to possible degenerate points for H_s or G_s) are connected by line segments, in order that the ROC curve is always continuous.
For $s = \eta$, we take the notation $\mathrm{ROC}^*(\alpha) = \mathrm{ROC}(\eta, \alpha)$.

The residual cdf $\bar{G}_s$ is also called the true positive rate and $\bar{H}_s$ is the false positive rate, so that the ROC curve is the plot of the true positive rate against the false positive rate. Basic properties of ROC curves can be found in the Appendix A.

The ROC curve provides a visual tool for comparing the ranking performance of two scoring rules.

Definition 5 *Consider two scoring functions s_1 and s_2. We say that s_1 provides a better ranking than s_2 when:*

$$\forall \alpha \in (0,1)\ , \quad \mathrm{ROC}(\mathrm{s}_1, \alpha) \geq \mathrm{ROC}(\mathrm{s}_2, \alpha)\ .$$

Remark 2 (GLOBAL VS. LOCAL PERFORMANCE.) Note that, as a functional criterion, the ROC curve induces a partial order over the space of all scoring functions. Some scoring function might provide a better ranking on some part of the observation space and a worst one on some other. A natural step to take is to consider local properties of the ROC curve in order to focus on best instances but this is not straightforward as explained in [CV07].

Therefore, we expect optimal scoring functions to be those for which the ROC curve dominates all the others for all $\alpha \in (0,1)$. The next proposition highlights the fact that the ROC curve is relevant when evaluating performance in the bipartite ranking problem.

Proposition 6 *The class $\mathcal{S}^*$ of optimal scoring functions provides the best possible ranking with respect to the ROC curve. Indeed, for any scoring function s, we have:*

$$\forall \alpha \in (0,1)\ , \quad \mathrm{ROC}^*(\alpha) \geq \mathrm{ROC}(\mathrm{s}, \alpha)\ ,$$

and

$$\forall s^* \in \mathcal{S}^*\ ,\ \forall \alpha \in (0,1)\ , \quad \mathrm{ROC}(\mathrm{s}^*, \alpha) = \mathrm{ROC}^*(\alpha)\ .$$

Moreover, if we set the notations:

$$R^*_\alpha = \{x \in \mathcal{X} \mid \eta(x) > Q^*(\alpha)\}$$
$$R_{s,\alpha} = \{x \in \mathcal{X} \mid s(x) > Q(s(X), \alpha)\}$$

then we have, for any s and any α such that $Q^*(\alpha) < 1$:

$$\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(\mathrm{s}, \alpha) = \frac{\mathbb{E}(|\eta(X) - Q^*(\alpha)|\ \mathbb{I}\{X \in R^*_\alpha \Delta R_{s,\alpha}\})}{p(1 - Q^*(\alpha))}$$

where Δ denotes the symmetric difference between sets.

The last statement reveals that the pointwise difference between the dominating ROC curve and the one related to a candidate scoring function s may be interpreted as the error made in recovering the specific level set R^*_α through $R_{s,\alpha}$.

A simple consequence of the previous result (and its proof) is that the one-dimensional statistic $\eta(X)$ (instead of the supposedly high-dimensional observation X) suffices to recover the optimal ROC curve. In other words, projecting the original data onto $(0,1)$ using the regression function leaves the ROC curve untouched.

Corollary 7 *Consider the statistical model corresponding to the observation of the random pair $(\eta(X), Y)$. Then the optimal ROC curve under this statistical model is exactly the same as the optimal ROC curve for the random pair (X, Y).*

The following result will be needed later.

Proposition 8 (Derivative of the ROC) *We assume that the optimal ROC curve is differentiable. Then, we have, for any α such that $Q^*(\alpha) < 1$:*

$$\frac{d}{d\alpha}\mathrm{ROC}^*(\alpha) = \frac{1-p}{p} \cdot \frac{Q^*(\alpha)}{1 - Q^*(\alpha)}\ .$$

2.3 AUC maximization

Although the ROC curve is a useful graphical tool for evaluating the performance of a scoring function, its use as the target of an optimization strategy to estimate ROC-optimal scoring functions turns out to be quite challenging. Indeed, selecting a scoring function by empirical maximization of the ROC curve over a class $\mathcal{S}$ of scoring functions is a highly complex task because of the functional nature of the ROC curve criterion.

Of course, the closer to ROC^* the ROC curve of a candidate scoring function $s \in \mathcal{S}$, the more pertinent the ranking induced by s. However, various metrics can be considered for measuring the distance between curves. We focus on two essential cases:

- the L_1 metric

$$d_1(s^*, s) = \int_0^1 \left(\mathrm{ROC}(\mathrm{s}^*, \alpha) - \mathrm{ROC}(\mathrm{s}, \alpha)\right)\ d\alpha.$$

- the L_∞ metric

$$d_\infty(s^*, s) = \sup_{\alpha \in (0,1)} \left(\mathrm{ROC}(\mathrm{s}^*, \alpha) - \mathrm{ROC}(\mathrm{s}, \alpha)\right)\ .$$

Remark 3 In order to avoid a possible confusion due to the notation, we bring to the reader's attention the fact that d_1 and d_∞ do not denote metrics on the space of scoring functions $\mathcal{S}$, but on the set of ROC curves.

As far as we know, the L_∞ metric has not been considered in the literature yet, although it is a natural choice given the view on the goal of ranking previously developed, *i.e.* recovering the collection of level sets $\{R^*_\alpha\}_{\alpha \in (0,1)}$ (see

subsection 2.2). Of course, L_∞-convergence implies convergence in the L_1-sense, while the reverse is generally false. However, the L_1-metric actually corresponds to a very popular criterion which is at the heart of most practical ranking methods. It is known as the Area Under an ROC Curve (or AUC, see [HM82]).

Definition 9 (AUC) *For any scoring function s, define the* AUC *as:*

$$\mathrm{AUC(s)} = \int_0^1 \mathrm{ROC(s},\alpha)\,\mathrm{d}\alpha\ ,$$

and set $\mathrm{AUC}^* = \mathrm{AUC}(\eta)$. *We then have:*

$$d_1(s^*, s) = \mathrm{AUC}^* - \mathrm{AUC(s)}\ .$$

When it comes to finding a scoring function, based on empirical data, which will perform well with respect to the AUC criterion, various strategies can be considered.

A possible angle is the *plug-in* approach ([DGL96]). The idea of plug-in consists in using an estimate $\hat{\eta}$ of the regression function as a scoring function. It is expected that, whenever $\hat{\eta}$ is close to η in a certain sense, then $\mathrm{ROC}(\hat{\eta},\cdot)$ and ROC^* are also close.

Proposition 10 *Consider $\hat{\eta}$ an estimator of η. We have:*

$$\mathrm{AUC}^* - \mathrm{AUC}(\hat{\eta}) \leq \frac{1}{\mathrm{p(1-p)}}\mathbb{E}\left(|\hat{\eta}(\mathrm{X}) - \eta(\mathrm{X})|\right)\ .$$

Assume furthermore that H^* has a density which is bounded by below on $[0,1]$: $\exists c > 0$ such that $\forall \alpha \in [0,1]$, $\frac{dH^*}{d\alpha}(\alpha) \geq c^{-1}$. Then, we have: $\forall \alpha \in [0,1]$ such that $Q^*(\alpha) < 1$,

$$\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(\hat{\eta},\alpha) \leq \frac{\mathrm{c}\mathbb{E}\left(|\mathrm{H}^*(\eta(\mathrm{X})) - \mathrm{H}_{\hat{\eta}}(\hat{\eta}(\mathrm{X}))|\right)}{\mathrm{p}(1-\mathrm{Q}^*(\alpha))}.$$

However, plug-in rules face difficulties when dealing with high-dimensional data ([GKKW02]). Another drawback of plug-in rules is that they are not consistent with respect to the supremum norm. This observation provides an additional motivation for exploring algorithms based on empirical AUC maximization.

A nice feature of the AUC performance measure is that it may be interpreted in a probabilistic fashion.

Proposition 11 ([CLVar]) *For any scoring function s such that G_s and H_s are continuous cdfs, we have:*

$$\begin{aligned}\mathrm{AUC(s)} &= \mathbb{P}(s(X) > s(X') \mid Y = 1, Y' = -1)\\ &= \frac{1}{2p(1-p)}\mathbb{P}\{(s(X) - s(X'))(Y - Y') > 0\}\ .\end{aligned}$$

where (X,Y) and (X',Y') are i.i.d. copies.

From this observation, ranking can be interpreted as classification of pairs of observations. We refer to [CLVar] for a systematic study of related empirical and convex risk minimization strategies which involve U-statistics. From a machine learning perspective, there is a growing literature in which existing algorithms are adapted in order to perform AUC optimization (such as, for instance: [CM04], [Rak04], [YDMW03]). The tree-based method we propose in the sequel consists in an adaptive recursive strategy for building a piecewise constant scoring function with nearly maximum AUC.

3 Piecewise linear approximation of the optimal ROC curve

In this section, we assume that the distribution, and hence the optimal ROC curve, are known. We also assume that the optimal ROC curve is differentiable and concave (check Proposition 22). We consider the problem of building, in a stepwise manner, a scoring function whose ROC curve is a piecewise linear approximation/interpolation of the optimal curve ROC^*.

3.1 Piecewise constant scoring functions

The motivation for considering piecewise constant scoring functions comes from the representation result on optimal scoring functions given in Proposition 3. When it comes to approximations of the optimal s^*, a natural idea is to introduce discrete versions and to replace the expectation by a finite sum.

We recall that a partition of $\mathcal{X}$ is a finite class $\mathcal{C}_N = (C_j)_{1\leq j\leq N}$ of sets such that $\bigcup_j C_j = \mathcal{X}$ and $C_i \cap C_j = \emptyset$ for $i \neq j$.

We now introduce D-representation of a piecewise constant scoring function where the 'D' stands for 'disjoint'.

Definition 12 (D-representation) *The D-representation of a piecewise constant scoring function s_N taking values in $\{a_1,\ldots,a_N\}$ is given by:*

$$\forall x \in \mathcal{X}\ ,\quad s_N(x) = \sum_{j=1}^N a_j\,\mathbb{I}\{x \in C_j\}\ ,$$

for some decreasing sequence $(a_j)_{j\geq 1}$ and some partition *$\mathcal{C}_N = (C_j)_{1\leq j\leq N}$ of $\mathcal{X}$.*

We now list some obvious properties of piecewise constant scoring function.

Proposition 13 *Consider some piecewise constant scoring function s_N taking N different values.*

(i) The ROC *curve of s_N is piecewise linear with N linear parts.*

(ii) The ROC *curve of s_N does not depend on the particular values of the sequence $(a_j)_{j\geq 1}$ appearing in its D-representation but only on their ordering.*

We introduce the class $\mathcal{S}_N$ of piecewise constant scoring functions which take N distinct values.

Definition 14 (Class $\mathcal{S}_N$) *We define $\mathcal{S}_N$ to be the class of scoring functions with D-representations of order N:*

$$\begin{aligned}\mathcal{S}_N = \{s_N = \sum_{j=1}^N a_j\,\mathbb{I}_{C_j}\ :\ &(C_j)_{j\geq 1} \textit{ is a disjoint partition}\ ,\\ &(a_j)_{j\geq 1} \textit{ is a decreasing sequence}\ \}\ .\end{aligned}$$

Our purpose in this section is to design an iterative procedure which outputs a piecewise constant scoring function $s_N \in \mathcal{S}_N$ whose ROC curve is as close as possible to the optimal ROC^*. Closeness between ROC curves will be measured both in terms of AUC and in the L_∞-sense. The iterative procedure described in the sequel satisfies the following approximation error result, see Proposition 15's proof.

Proposition 15 *Assume that the optimal* ROC *curve is twice differentiable and concave and that its second derivative takes its values in a bounded interval which does not contain zero. There exists a sequence of piecewise constant scoring functions $(s_N)_{N\geq 1}$ such that, for any $N \geq 1$, $s_N \in \mathcal{S}_N$ and:*

$$\mathrm{AUC}^* - \mathrm{AUC}(\mathrm{s_N}) = \mathrm{d_1}(\mathrm{s}^*, \mathrm{s_N}) \leq C \cdot N^{-2} \ ,$$
$$d_\infty(s^*, s_N) \leq C \cdot N^{-2} \ ,$$

where the constant C depends only on the distribution.

The proof can be found in the Appendix. The approximation rate $O(N^{-2})$ is actually reached by any piecewise linear approximant provided that the mesh length is of order $O(N^{-1})$. This result is well-known folklore in approximation theory, see [DL93]. We underline that the piecewise linear approximation method we describe next is adaptive in the sense that breakpoints are not fixed in advance and strongly depend on the target curve (which suggests that this scheme possibly yields a sharper constant C). It highlights the explicit relationship between the ROC^* approximant and the corresponding piecewise constant scoring function. The ranking algorithm proposed in the sequel (Section 4) will appear as a statistical version of this variable knot approximation, where the unknown quantities driving the recursive partitioning will be replaced by their empirical counterparts.

3.2 An alternative representation of scoring functions

It will be useful to consider another possible representation of piecewise constant scoring functions which is based on increasing sequences of sets.

Definition 16 (Increasing sequence of sets) *We call an* increasing sequence of sets *of $\mathcal{X}$ a finite class of sets $\mathcal{R}_N = (R_j)_{1\leq j\leq N}$ such that $\bigcup_j R_j = \mathcal{X}$ and $R_i \subset R_j$ for $i < j$. In particular, we have $R_N = \mathcal{X}$.*

Definition 17 (I-representation) *Consider a piecewise constant scoring function s_N taking values in $\{1, \ldots, N\}$. Its I-representation is given by:*

$$\forall x \in \mathcal{X} \ , \quad s_N(x) = \sum_{j=1}^{N} \mathbb{I}\{x \in R_j\} \ ,$$

for some increasing sequence $\mathcal{R}_N = (R_j)_{1\leq j\leq N}$ *of subsets of $\mathcal{X}$.*

The relationship between D- and I-representations is straightforward. Assume that s_N takes values in $\{1, \ldots, N\}$ and consider the sequence $\mathcal{R}_N$ arising from the I-representation. We can then obtain the D-representation by taking $C_1 = R_1$ and:

$$\forall i > 2 \ , \ C_i = R_i \setminus R_{i-1} \quad \text{and} \quad \forall j \ , \ a_j = N - j + 1 \ .$$

3.3 One-step approximation to the optimal ROC curve

We now provide some insights on the general construction by describing the one-step modification of a given piecewise constant scoring function s_N. As advocated by Proposition 3, modifications are picked up in the class $\mathcal{G}$ of level sets of the regression function η:

$$\mathcal{G} = \{\{x \in \mathcal{X} \ : \ \eta(x) > t\} : t \in (0, 1)\}$$

Definition 18 (One-step approximation) *Given $s_N \in \mathcal{S}_N$, we define:*

$$\sigma_N = \arg\max_{\sigma \in \mathcal{G}} d_1(s_N, s_N + \sigma).$$

Then, the one-step approximation sequence to some optimal scoring function s^* is defined as the sequence $(s_N)_{N\geq 1}$ of scoring functions such that:

$$s_1 = \mathbb{I}_{\mathcal{X}}$$
$$s_{N+1} = s_N + \sigma_N \ , \quad N \geq 1 \ .$$

At this point, we shall consider the I-representation of piecewise constant scoring functions. A constructive procedure will rely on a particular choice of subsets $(R_j)_{j\geq 1}$. Following the result from Proposition 3, we focus on partitions with sets of the form:

$$R_j = \{x \in \mathcal{X} \ : \ \eta(x) > u_j\}$$

for some positive decreasing sequence $(u_j)_{j\geq 1}$ with $u_1 > 0$.

First iteration. We initialize the procedure for $N = 1$ with the scoring function:

$$\forall x \in \mathcal{X} \ , \quad s_1(x) = \mathbb{I}\{x \in \mathcal{X}\} \equiv 1 \ ,$$

which ranks all instances equally. It is clear that adding up the indicator of any region of the form $\{\eta(x) > t\}$ for some $t \in (0, 1)$ would provide a piecewise linear approximation of the optimal ROC curve. We choose the one which maximizes the AUC criterion.

Proposition 19 (First iteration) *Assume that the optimal* ROC *curve is differentiable and concave. Then the one-step approximation at the first iteration is given by the piecewise constant scoring function:*

$$\forall x \in \mathcal{X} \ , \quad s_2(x) = \mathbb{I}\{x \in \mathcal{X}\} + \mathbb{I}\{\eta(x) > t^*\} \ .$$

with $t^* = p$, where $p = \mathbb{P}\{Y = 1\}$. We also have:

$$(d\beta/d\alpha)(t^*) = 1.$$

Remark 4 (RANKING VS. CLASSIFICATION.) We point out that the optimal binary-valued scoring function in the AUC sense does not correspond to the Bayes classifier $g^*(x) = 2\mathbb{I}\{\eta(x) > 1/2\} - 1$, except when $p = 1/2$. Indeed, if we consider classifiers $g_t(x) = 2\mathbb{I}\{\eta(x) > t\} - 1$ of the form and look for the minimizer of the classification error:

$$\mathbb{P}\{Y \neq g_t(X)\} = p(1 - \alpha(t)) + (1 - p)\beta(t),$$

which is minimum for t such that $\frac{d\beta}{d\alpha}(t) = (1 - p)/p$ (if such a value can be reached), and hence $t = 1/2$ by Proposition 8. Denote by $r_{\max} = (d/d\alpha)(\mathrm{ROC}^*)(0)$ and $r_{\min} = (d/d\alpha)(\mathrm{ROC}^*)(1)$. When p falls out of the interval $((1 + r_{\max})^{-1}, (1 + r_{\min})^{-1})$ then one of the two extremal values will give the solution.

It is noteworthy that the one-step approximation obtained by optimization of the AUC criterion is the same as the one obtained through optimization of the sup-norm. The proof of the following proposition is simple and left to the reader.

Proposition 20 *Consider the increments at the first step:*

$$\sigma_1 = \arg\max_{\sigma \in \mathcal{G}} d_1(s_1, s_1 + \sigma)$$
$$\tilde{\sigma}_1 = \arg\max_{\sigma \in \mathcal{G}} d_\infty(s_1, s_1 + \sigma) \, .$$

We have: $\tilde{\sigma}_1 = \sigma_1$.

N-th iteration. Now consider a piecewise constant scoring function $s_N \in \mathcal{S}_N$. The ROC curve of s_N is a broken line with N linear pieces defined by the sequence of points $((\alpha_j, \beta_j))_{0 \leq j \leq N}$ where $(\alpha_0, \beta_0) = (0, 0)$ and $(\alpha_N, \beta_N) = (1, 1)$.

We look for the optimal splitting which would increase the AUC by adding a knot $(\alpha(t), \beta(t))$ such that $\alpha(t)$ is between α_j and α_{j+1}. We take the notation

$$s^{(j)}_{N+1,t}(x) = s_N(x) + \mathbb{I}\{\eta(x) > t\} \, ,$$

with $t \in (Q^*(\alpha_{j+1}), Q^*(\alpha_j))$. The AUC can then be written, for some constant c_j, as:

$$\begin{aligned} A_{N+1}(t) &= \mathrm{AUC}(\mathrm{s}^{(\mathrm{j})}_{\mathrm{N+1,t}}) \\ &= c_j + \frac{1}{2}(\alpha_{j+1} - \alpha_j)\beta(t) - \frac{1}{2}\alpha(t)(\beta_{j+1} - \beta_j) \, , \end{aligned}$$

which is maximized at t^* such that:

$$d\beta(t^*) = \left(\frac{\beta_{j+1} - \beta_j}{\alpha_{j+1} - \alpha_j} \right) d\alpha(t^*) \, .$$

We can set $\alpha^*_j = \alpha(t^*)$ and we get, thanks to Proposition 8, the following relationship:

$$\frac{1-p}{p} \cdot \frac{Q^*(\alpha^*_j)}{1 - Q^*(\alpha^*_j)} = \frac{\beta_{j+1} - \beta_j}{\alpha_{j+1} - \alpha_j} \, .$$

This leads to a one-step optimal splitting point (α^*_j, β^*_j) on the ROC curve such that:

$$\alpha^*_j = \bar{H}^* \left(\Delta_j \right) \quad \text{and} \quad \beta^*_j = \bar{G}^* \left(\Delta_j \right)$$

where

$$\Delta_j = \frac{p(\beta_{j+1} - \beta_j)}{(1-p)(\alpha_{j+1} - \alpha_j) + p(\beta_{j+1} - \beta_j)} = t^* \, .$$

Remark 5 (INTERPRETATION IN TERMS OF PARTITIONS.) The insertion of the new knot (α^*_j, β^*_j) is materialized by the splitting of subset R_{j+1} with a subset R^*_j containing R_j and we have:

$$R^*_j = \{x \in \mathcal{X} \, : \, \eta(x) > Q^*(\alpha^*_j)\} \, ,$$

while $R_j = \{x \in \mathcal{X} \, : \, \eta(x) > Q^*(\alpha_j)\}$. In terms of D-representations, we can write:

$$s_N = \sum_{j=1}^{N} (N - j + 1) \, \mathbb{I}_{C_j}$$

where

$$C_j = \{x \in \mathcal{X} \, : \, Q^*(\alpha_{j+1}) < \eta(x) \leq Q^*(\alpha_j)\} \, .$$

After the splitting, in the new partition, the set C_{j+1} is replaced by C^*_j and $C_{j+1} \setminus C^*_j$ where

$$C_{j+1} = \{x \in \mathcal{X} \, : \, Q^*(\alpha_{j+1}) < \eta(x) \leq Q^*(\alpha^*_j)\} \, .$$

The previous computations quantify the improvement in terms of AUC after adding one knot for each linear part of the ROC curve at step N. Instead of sticking to one-step approximations, we can introduce an approximation scheme which will add 2^N knots after the N-th iteration.

3.4 A tree-structured recursive approximation scheme

We now turn to the full recursive procedure. At each step, an adaptively chosen knot is added between all consecutive points of the current meshgrid. We take $N = 2^D$ with $D \geq 0$ and we describe iterations over D for constructing a sequence of piecewise constant scoring functions. It will be easier to work with D-representations of the form:

$$\forall x \in \mathcal{X} \, , \quad s_D(x) = \sum_{k=0}^{2^D - 1} (2^D - k) \, \mathbb{I}\{x \in C_{D,k}\} \, ,$$

where, for fixed D, the class of sets $(C_{D,k})_{0 \leq k \leq 2^D - 1}$ is a disjoint partition of $\mathcal{X}$.

We will use the following notations:

$$\alpha(C) = \mathbb{P}\{X \in C \mid Y = -1\}$$
$$\beta(C) = \mathbb{P}\{X \in C \mid Y = 1\} \, .$$

The iterative procedure goes as follows.

Initialization ($d = 0$ and $d = 1$). For the extremal points, we set:

$$\forall d \in \mathbb{N} \, , \; \alpha^*_{d,0} = \beta^*_{d,0} = 0 \quad \text{and} \quad \alpha^*_{d,2^d} = \beta^*_{d,2^d} = 1 \, ,$$

and for the first iteration points ($d = 1$):

$$\alpha^*_{1,1} = \bar{H}^*(p) \quad \text{and} \quad \beta^*_{1,1} = \bar{G}^*(p) \, ,$$

From d to $d+1$, for $d \geq 1$. We are given the collection of points $\{(\alpha^*_{d,k}, \beta^*_{d,k})\}_{k=0,\ldots,2^d-1}$. On each interval $(\alpha^*_{d,k}, \alpha^*_{d,k+1})$, we apply the one-step approximation. Hence, the new point is given by:

$$\begin{aligned} \alpha^*_{d+1,2k+1} &= \bar{H}^* \left(\Delta^*_{d+1,2k+1} \right) , \\ \beta^*_{d+1,2k+1} &= \bar{G}^* \left(\Delta^*_{d+1,2k+1} \right) , \end{aligned}$$

where

$$\Delta^*_{d+1,2k+1} = \frac{p(\beta^*_{d,k+1} - \beta^*_{d,k})}{(1-p)(\alpha^*_{d,k+1} - \alpha^*_{d,k}) + p(\beta^*_{d,k+1} - \beta^*_{d,k})} \, .$$

Moreover, the previous cut-off point is renamed:

$$\alpha^*_{d+1,2k} = \alpha^*_{d,k} \quad \text{and} \quad \beta^*_{d+1,2k} = \beta^*_{d,k} \, ,$$

and also $\Delta^*_{d+1,2k} = \Delta^*_{d,k}$.

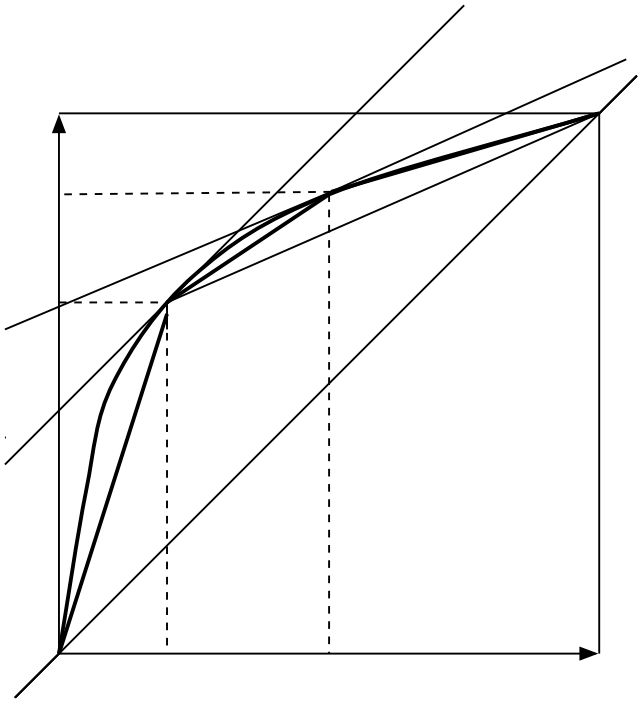

Figure 1: Piecewise linear approximation of the ROC curve.

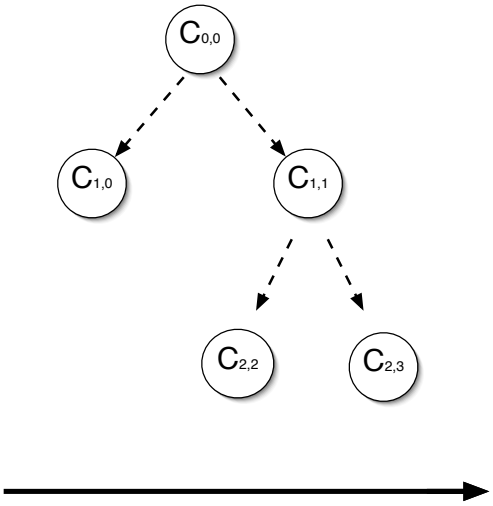

Figure 2: Numbering of the nodes and order for reading the ranks.

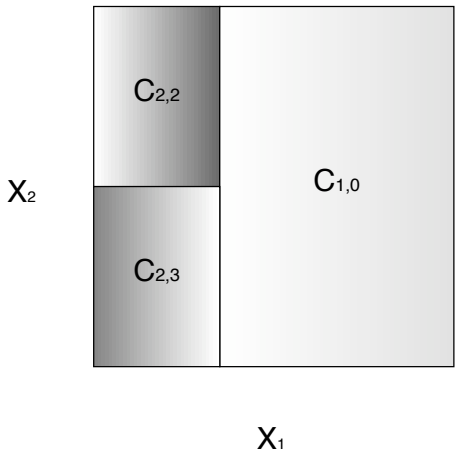

Figure 3: Partitioning induced by the tree structure with perpendicular splits.

Note that, for each level d, the resulting partition is given by the class of sets:

$$C^*_{d,k} = \{x \in \mathcal{X} \,:\, \Delta^*_{d,k} < \eta(x) \leq \Delta^*_{d,k+1}\},$$

for all $k = 0, \ldots, 2^d - 1$ with the convention that $\Delta^*_{d,0} = 0$ and $\Delta^*_{d,2^d} = 1$ for all $d \geq 0$.

We also define the sets $R^*_{d,k}$ by: $R^*_{d,k} = C^*_{d,k} \cup R^*_{d,k-1}$ with $R^*_{d,0} = C^*_{d,0}$.

Remark 6 (A TREE-STRUCTURED RECURSIVE INTERPOLATION SCHEME.) A nice feature of the recursive approximation procedure is its binary-tree structure. Owing to their crucial practical advantages regarding implementation and interpretation, tree-structured decision rules have been proved useful for a wide range of statistical tasks and are in particular among the most popular methods for regression and classification (we refer to Chapter 20 in [DGL96] for an excellent account of tree decision rules in the context of classification).

Remark 7 (A PIECEWISE CONSTANT APPROXIMANT OF THE REGRESSION FUNCTION.) Although the angle embraced in this paper consists of directly building a partitioning of the input space corresponding to a nearly optimal ranking in the spirit of popular machine-learning algorithms, we point out that, as a byproduct, the resulting partition provides a stepwise approximation of the regression function:

$$\tilde{\eta}(x) \quad = \quad \sum_{k=0}^{2^D-1} \Delta^*_{D+1,2k+1}\, \mathbb{I}\{x \in C^*_{D,k}\}$$

Provided that H^* is strictly increasing, the scoring function $s(x) = H^*(\eta(x))$ is also optimal and is approximated by:

$$\tilde{s}(x) = \sum_{j=0}^{2^D-1} (\alpha^*_{D,j+1} - \alpha^*_{D,j})\, \mathbb{I}\{x \in R^*_{D,j}\},$$

which should be seen as a Riemann's discretization of the integral $\int_0^1 \mathbb{I}\{\eta(x) > Q^*(\alpha)\}\, d\alpha$ (see Remark 1).

4 A tree-structured weak ranker

It is time to exploit the theory developed in the previous sections to deal with empirical data. We formulate a practical algorithm which implements a top-down strategy to build a binary tree-structured scoring function. This algorithm mimics the ideal recursive approximation procedure of the optimal ROC curve from Section 3, where probabilities are replaced by their empirical counterparts.

4.1 The TREERANK algorithm

We assume now that a training data set

$$\mathcal{D}_n = \{(X_1, Y_1),\ \ldots,\ (X_n, Y_n)\}$$

of n independent copies of the pair (X, Y) is available. We set

$$n_+ = \sum_{i=1}^n \mathbb{I}\{Y_i = 1\} \quad \text{and} \quad n_- = \sum_{i=1}^n \mathbb{I}\{Y_i = -1\}\,.$$

We introduce the following data-based quantities, for any subset C:

$$\hat{\alpha}(C) = \frac{1}{n_-} \sum_{i=1}^n \mathbb{I}\{X_i \in C,\ Y_i = -1\}$$

$$\hat{\beta}(C) = \frac{1}{n_+} \sum_{i=1}^n \mathbb{I}\{X_i \in C,\ Y_i = +1\}$$

which correspond respectively to the empirical false positive rate and the empirical true positive rate of a classifier predicting +1 on the set C.

For notational convenience, we set $\alpha_{d,0} = \beta_{d,0} = 0$ and $\alpha_{d,2^d} = \beta_{d,2^d} = 1$ for all $d \geq 0$. We assume that we are given a class $\mathcal{C}$ of subsets of $\mathcal{X}$.

TREERANK ALGORITHM

1. **Initialization.** Set $C_{0,0} = \mathcal{X}$ and $\alpha_{0,1} = \beta_{0,1} = 1$.

2. **Iterations.** For $d = 0,\ \ldots,\ D-1$ and for $k = 0,\ \ldots,\ 2^d - 1$:

   (a) (OPTIMIZATION STEP.) Set the entropy measure:

   $$\begin{aligned}\text{Ent}_{d,k+1}(C) = (\alpha_{d,k+1} - \alpha_{d,k})\hat{\beta}(C) \\ - (\beta_{d,k+1} - \beta_{d,k})\hat{\alpha}(C)\,.\end{aligned}$$

   Find the best subset $C_{d+1,2k}$ of rectangle $C_{d,k}$ in the AUC sense:

   $$C_{d+1,2k} = \operatorname*{arg\,max}_{C \in \mathcal{C},\ C \subset C_{d,k}} \text{Ent}_{d,k+1}(C)\,.$$

   Then, set $C_{d+1,2k+1} = C_{d,k} \setminus C_{d+1,2k}$.

   (b) (UPDATE.) Set

   $$\begin{aligned}\alpha_{d+1,2k+1} &= \alpha_{d,k} + \hat{\alpha}(C_{d+1,2k}) \\ \beta_{d+1,2k+1} &= \beta_{d,k} + \hat{\beta}(C_{d+1,2k})\end{aligned}$$

   and

   $$\begin{aligned}\alpha_{d+1,2k+2} &= \alpha_{d,k+1} \\ \beta_{d+1,2k+2} &= \beta_{d,k+1}\,.\end{aligned}$$

3. **Output.** After D iterations, we get the piecewise constant scoring function s_D:

$$s_D(x) = \sum_{k=0}^{2^D-1} (2^D - k)\, \mathbb{I}\{x \in C_{D,k}\}$$

The main features of the TREERANK algorithm are listed in the following remarks.

Remark 8 (READING THE RANKS.) The resulting ranking induced by the scoring function s_D may be read from the left to the right looking at the terminal nodes (see Figure 2).

Remark 9 (A SIMPLISTIC STOPPING CRITERION.) If there is more than one subrectangle solution in the OPTIMIZATION STEP, take the larger. Hence, if there is no improvement in terms of AUC maximization when splitting the current rectangle $C_{d,k}$, set $C_{d+1,2k} = C_{d,k}$, so that $C_{d+1,2k+1} = \emptyset$.

Remark 10 (ON THE SPLITTING CRITERION.) In the context of classification, this splitting rule has been considered previously in [FFHO02]. We point out that, in contrast to tree-based classification methods, such as CART, the splitting criterion depends on the node through the parent's false and true positive rates $\hat{\alpha}(C)$ and $\hat{\beta}(C)$. This can be explained by the fact that the goal pursued in the ranking problem is global: one attempts to order all input data with respect to each other.

Remark 11 (ORTHOGONAL SPLITS.) As a practical strategy, one could use a class $\mathcal{R}$ of decision stumps which are obtained by cutting a certain coordinate of the input vector X at a certain level (the split variable and the level being chosen so as to maximize the AUC). The subclass to be enumerated is then the intersection of decision stumps with the set represented in the parent node.

4.2 Consistency of TREERANK

We now provide a consistency result for the class of partitions induced by the TREERANK algorithm. The formulation (and the proof) mimics Theorem 21.2 from [DGL96].

Theorem 21 *We consider scoring functions s_n corresponding to partitions $\mathcal{F}_n$ of $\mathcal{X}$. We assume that the $\mathcal{F}_n$'s are random partitions of $\mathcal{X}$ resulting from runs of* TREERANK *with training sets of size n. We also assume that $\mathcal{X}$ is bounded and that the partitions $\mathcal{F}_n$ belong to a VC class of sets with VC dimension V, for any n and any training set. If the diameter of any cell of $\mathcal{F}_n$ goes to 0 when n tends to infinity, then we have that:*

$$\mathrm{AUC}(\mathrm{s}^*) - \mathrm{AUC}(\mathrm{s_n}) = \mathrm{d_1}(\mathrm{s}^*, \mathrm{s_n}) \to 0$$

almost surely, as n goes to ∞.

If we have, in addition that H^* has a density which is bounded by below on $[0,1]$ and that, for any α, $Q^*(\alpha) < 1 - \epsilon$, for some $\epsilon > 0$, then:

$$d_\infty(s^*, s_n) \to 0$$

almost surely, as n goes to ∞.

Remark 12 (BOUNDEDNESS OF $\mathcal{X}$.) This assumption is a simplification which can be removed at the cost of a longer proof (the core of the argument can be found in [DGL96]).

Remark 13 (COMPLEXITY ASSUMPTION.) Instead of assuming a finite VC dimension, a weaker assumption on the combinatorial entropy of the class of partitions may be provided (again check [DGL96] for this refinement).

Sketch of proof. The proof of the consistency result in the case of decision trees for classification is based on the control of the excess risk in terms of the L_1-distance between the regression function and its plug-in estimator obtained as a local estimation on one cell. In the case of ranking, we can use a similar argument both for the AUC criterion and the supremum norm over the ROC curves thanks to Proposition 10. For a given sample $\mathcal{D}_n$, consider the sequences of sets $(R_{d,k})_{d,k}$, $(C_{d,k})_{d,k}$ and the sequences $\{(\alpha_{d,k}, \beta_{d,k})\}_{d,k}$ arising from a run of TREERANK with depth $N = 2^D$. We can then deal with the two metrics in a similar way:

- L_1 metric (AUC) - we can consider the following plug-in estimator of the regression function (see Remark 7):

$$\hat{\eta}(x) = \sum_{k=0}^{2^D-1} \Delta_{D+1,2k+1}\, \mathbb{I}\{x \in C_{D,k}\},$$

where

$$\Delta_{D+1,2k+1} = \frac{n_+(\beta_{d,k+1} - \beta_{d,k})}{n_-(\alpha_{d,k+1} - \alpha_{d,k}) + n_+(\beta_{d,k+1} - \beta_{d,k})}.$$

Now denote by j_0 the index of the set such that $x \in C_{D,j_0}$ and we have:

$$\hat{\eta}(x) = \Delta_{D+1,2j_0+1}\ .$$

Then use the inequality from Proposition 10:

$$\mathrm{AUC}^* - \mathrm{AUC}(\hat{\eta}) \leq \frac{1}{\mathrm{p}(1-\mathrm{p})} \mathbb{E}\left(|\hat{\eta}(\mathrm{X}) - \eta(\mathrm{X})|\right)\ .$$

- L_∞ metric - here we introduce the estimator:

$$\hat{s}(x) = \sum_{j=0}^{2^D-1} (\alpha_{D,j+1} - \alpha_{D,j})\, \mathbb{I}\{x \in R_{D,j}\}$$

for $H^* \circ \eta$. But we have, by construction:

$$\mathbb{I}\{x \in R_{D,j}\} = \sum_{k=0}^{j} \mathbb{I}\{x \in C_{D,k}\}\ .$$

As before, take j_0 to be the index of the set such that $x \in C_{D,j_0}$ and we have:

$$\hat{s}(x) = 1 - \alpha_{D,j_0}\ .$$

Then we have, also by Proposition 10, for any α:

$$\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(\hat{\mathrm{s}}, \alpha) \leq \frac{\mathrm{c}\mathbb{E}\left(|\mathrm{H}^*(\eta(\mathrm{X})) - \hat{\mathrm{s}}(\mathrm{X}))|\right)}{\mathrm{p}(1 - \mathrm{Q}^*(\alpha))}.$$

Note that

$$\alpha_{D,j_0} = \frac{1}{n_-} \sum_{i=1}^{n} \mathbb{I}\{X_i \in C_{D,j_0}, Y_i = -1\}$$

$$\beta_{D,j_0} = \frac{1}{n_+} \sum_{i=1}^{n} \mathbb{I}\{X_i \in C_{D,j_0}, Y_i = 1\}\ .$$

This observation indicates that the same argument will work for the two metrics. From there, the rest of the proof is exactly as in Theorem 21.2 from [DGL96], except that n_+, n_- are random. We can write, for instance:

$$\frac{1}{n_-} = \frac{1}{n_-} - \frac{1}{n(1-p)} + \frac{1}{n(1-p)}\ ,$$

and we can see that there will be a corrective term of the order of $n^{-1/2}$ which will not affect the convergence.

Appendix A - Properties of ROC curves

We now recall some simple properties of ROC curves (see [vT68], [HT96]).

Proposition 22 (Properties of the ROC curve) *For any distribution P and any scoring function $s : \mathcal{X} \to \mathbb{R}$, the following properties hold:*

1. **Limit values.** *We have:* $\mathrm{ROC}(\mathrm{s}, 0) = 0$ *and* $\mathrm{ROC}(\mathrm{s}, 1) = 1$
2. **Invariance.** *For any strictly increasing function $T : \mathbb{R} \to \mathbb{R}$, we have, for all $\alpha \in (0, 1)$:* $\mathrm{ROC}(\mathrm{T} \circ \mathrm{s}, \alpha) = \mathrm{ROC}(\mathrm{s}, \alpha)$.
3. **Concavity.** *If the likelihood ratio dG_s/dH_s is a monotone function then the* ROC *curve is concave.*
4. **Linear parts.** *If the likelihood ratio dG_s/dH_s is constant on some interval in the range of the scoring function s then the* ROC *curve will present a linear part on the corresponding domain.*
5. **Differentiability.** *Assume that the distribution μ of X is continuous. Then, the* ROC *curve of a scoring function s is differentiable if and only if the conditional distribution of $s(X)$ given Y is continuous.*

Appendix B - Proofs

Proof of Proposition 3

First note that, for any scoring function s with range equal to (m, M), if U has a uniform distribution in (m, M), then:

$$\forall x \in \mathcal{X} \ , \quad \mathbb{E}(\mathbb{I}\{s(x) > U\}) = \frac{s(x) - m}{M - m} \ .$$

Assume that the range of η has no holes. Then for $s^* \in \mathcal{S}^*$ with range equal to $[m, M]$, there exists a strictly increasing function $T : (0, 1) \to [m, M]$ such that $s^* = T \circ \eta$. We have:

$$\forall x \in \mathcal{X} \ , \quad s^*(x) = m + (M - m)\mathbb{E}(\mathbb{I}\{\eta(x) > T^{-1}(U)\}) \ .$$

We can set $V = T^{-1}(U)$ and $w(V) = M - m$, and the 'only if' part is proved in the case where $\eta(X)$ has a support equal to $[0, 1]$. For the general case, we only have to take w to be the indicator of the support of $\eta(X)$.

Now assume that s^* has the given form. In order to show that s^* is an optimal scoring function, it suffices to prove that the ordering induced by s on a pair (x, x') is the same as the one induced by η. Denote by ϕ the df of V with respect to the Lebesgue measure. We have:

$$\forall x, x' \in \mathcal{X} \ , \quad s^*(x) - s^*(x') = \int_{\eta(x')}^{\eta(x)} w(v)\phi(v) \, dv \ ,$$

which gives the result since ϕ and w are nonnegative.

Proof of Proposition 6 and Corollary 7

The proposition is a simple consequence of Neyman-Pearson's lemma formulated in the appropriate setting. Consider the following hypothesis testing problem: given the observation X, test the null assumption $H_0 \ : \ Y = -1$ against the alternative $H_1 \ : \ Y = +1$. Denote by $p = \mathbb{P}\{Y = 1\}$. The optimal test statistic is then given by the likelihood ratio test:

$$\phi^*(x) = \frac{\mathbb{P}\{X = x \mid Y = 1\}}{\mathbb{P}\{X = x \mid Y = -1\}} = \frac{1 - p}{p} \cdot \frac{\eta(x)}{1 - \eta(x)} \ .$$

Denote by $Q(Z, \alpha)$ the quantile of order $1 - \alpha$ for the distribution of Z conditioned on the event $Y = -1$. By Neyman-Pearson's lemma, we have that among all test statistics $\phi(X)$ with fixed type I error $\alpha = \mathbb{P}\{\phi(X) > Q(\phi(X), \alpha) \mid Y = -1\}$, the test defined by the statistic $\phi^*(X)$ maximizes the power $\beta = \mathbb{P}\{\phi(X) > Q(\phi(X), \alpha) \mid Y = 1\}$. Moreover, the class of distributions $\{\mathbb{P}\{X = x \mid Y = \theta\}\}_{\theta \in \{0,1\}}$ is a monotone likelihood ratio family in $\eta(X)$. Indeed, since the function $u \mapsto \frac{1-p}{p} \cdot \frac{u}{1-u}$ is strictly increasing on $(0, 1)$, the test based on the statistic $\phi^*(X)$ is obviously equivalent to the one based $\eta(X)$. Hence η is an optimal scoring function in the sense of the ROC curve. Any element of the class $\mathcal{S}^*$ will also maximize the ROC curve thanks to the invariance property under strictly increasing transforms.

The last statement of Proposition 6 is proved as follows. First, we use the fact that, for any measurable function h, we have:

$$\mathbb{E}(h(X) \mid Y = +1) = \frac{1 - p}{p}\mathbb{E}\left(\frac{\eta(X)}{1 - \eta(X)} \, h(X) \mid Y = -1\right) \ .$$

We apply this with $h(X) = \mathbb{I}\{X \in R^*_\alpha\} - \mathbb{I}\{X \in R_{s,\alpha}\}$ to get:

$$\begin{aligned} &\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(\mathrm{s}, \alpha) \\ &\qquad = \frac{1 - p}{p}\mathbb{E}\left(\frac{\eta(X)}{1 - \eta(X)} h(X) \mid Y = -1\right) \ . \end{aligned}$$

Then we add and substract $\frac{Q^*(\alpha)}{1-Q^*(\alpha)}$ and using the fact that

$$1 - \alpha = \mathbb{P}\{X \in R_{s,\alpha}\} = \mathbb{P}\{X \in R^*_\alpha\} \ ,$$

we get:

$$\begin{aligned} &\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(\mathrm{s}, \alpha) \\ &= \left(\frac{1 - p}{p}\right) \mathbb{E}\left(\left(\frac{\eta(X)}{1 - \eta(X)} - \frac{Q^*(\alpha)}{1 - Q^*(\alpha)}\right) \, h(X) \middle| Y = -1\right) \ . \end{aligned}$$

We remove the conditioning with respect to $Y = -1$ and using then conditioning on X, we obtain:

$$\begin{aligned} &\mathrm{ROC}^*(\alpha) - \mathrm{ROC}(\mathrm{s}, \alpha) \\ &\qquad = \frac{1}{p}\mathbb{E}\left(\left(\frac{\eta(X) - Q^*(\alpha)}{1 - Q^*(\alpha)}\right) \, h(X)\right) \ . \end{aligned}$$

It is then easy to see that this expression corresponds to the statement in the Proposition.

Proof of Proposition 8

In the proof of Proposition 6, we saw that the likelihood ratio test statistic was given by:

$$\phi^*(x) = \frac{\mathbb{P}\{X = x \mid Y = 1\}}{\mathbb{P}\{X = x \mid Y = -1\}} = \frac{1 - p}{p} \cdot \frac{\eta(x)}{1 - \eta(x)} \ .$$

Now consider, for any measurable function m, the following conditional expectation with respect to the random variable X given $Y = 1$:

$$\mathbb{E}(m(\eta(X)) \mid Y = 1) = \mathbb{E}\left(m(\eta(X)) \cdot \phi^*(X) \mid Y = -1\right)$$

which can also be expressed as a conditional expectation with respect to the random variable $Z = \eta(X)$ given $Y = 1$:

$$\mathbb{E}(m(Z) \mid Y = 1) = \mathbb{E}\left(m(Z) \cdot \frac{dG^*}{dH^*}(Z) \middle| Y = -1\right) .$$

We can then proceed to the following identification:

$$\phi^*(X) = \frac{dG^*}{dH^*}(\eta(X))$$

We have obtained the following formula for the likelihood ratio of the random variable $\eta(X)$:

$$\forall u \in (0,1) , \quad \frac{dG^*}{dH^*}(u) = \frac{1-p}{p} \cdot \frac{u}{1-u} ,$$

which gives the result.

Proof of Proposition 10

We recall (see [CLVar]) that:

$$\mathrm{AUC}^* - \mathrm{AUC}(\hat{\eta}) = \frac{\mathbb{E}\left(|\eta(\mathrm{X}) - \eta(\mathrm{X}')| \mathbb{I}\{(\mathrm{X}, \mathrm{X}') \in \Gamma\}\right)}{2\mathrm{p}(1-\mathrm{p})} .$$

where

$$\Gamma = \{(x, x') : \ \mathrm{sgn}(\hat{\eta}(X) - \hat{\eta}(X')) \neq \mathrm{sgn}(\eta(X) - \eta(X'))\}$$

But, one may easily check that:
if $\mathrm{sgn}(\hat{\eta}(X) - \hat{\eta}(X')) \neq \mathrm{sgn}(\eta(X) - \eta(X'))$, then

$$|\eta(X) - \eta(X')| \leq |\eta(X) - \hat{\eta}(X)| + |\eta(X') - \hat{\eta}(X')| ,$$

which gives the first part of the result.

Turning to the second assertion, consider the event

$$\mathcal{E} = \{X \in R^*_\alpha \Delta R_{\hat{\eta},\alpha}\}.$$

Notice first that, after Proposition 6, we have:

$$\begin{aligned} \mathrm{ROC}^*(\alpha) - \mathrm{ROC}(\hat{\eta}, \alpha) &= \frac{\mathbb{E}(|\eta(X) - Q^*(\alpha)| \, \mathbb{I}_{\mathcal{E}})}{p(1 - Q^*(\alpha))} \\ &\leq \frac{c\mathbb{E}(|H^*(\eta(X)) - 1 + \alpha| \, \mathbb{I}_{\mathcal{E}})}{p(1 - Q^*(\alpha))} \end{aligned}$$

by virtue of the finite increments theorem. Now, observing that

$$\mathcal{E} = \{\mathrm{sgn}(H^*(\eta(X)) - 1 + \alpha) \neq \mathrm{sgn}(H_{\hat{\eta}}(\hat{\eta}(X)) - 1 + \alpha)\},$$

we have in a similar fashion as above: if $X \in R^*_\alpha \Delta R_{\hat{\eta},\alpha}$, then

$$|H^*(\eta(X)) - 1 + \alpha| \leq |H^*(\eta(X)) - H_{\hat{\eta}}(\hat{\eta}(X))|,$$

which, combined to the previous bound, proves the second part.

Proof of Proposition 15

We now show that the recursive approximation procedure described in Subsection 3.4 provides a sequence of piecewise constant scoring functions $(s_D)_{D \geq 0}$ with N constant parts which achieves an approximation error rate for the AUC of the order of 2^{-2D}.

For any $\alpha \in (\alpha_{D,k}, \alpha_{D,k+1})$, we have, for any optimal scoring function s^*, by concavity of η:

$$\begin{aligned} \mathrm{ROC}(\mathrm{s}^*, \alpha) - \mathrm{ROC}(\mathrm{s_D}, \alpha) &\leq -\frac{1}{8}(\alpha_{D,k+1} - \alpha_{D,k})^2 \\ &\times \frac{d^2}{d\alpha^2}\mathrm{ROC}(\mathrm{s}^*, \alpha_{\mathrm{D,k}}) . \end{aligned}$$

By assumption, the second derivative of the optimal ROC is bounded and hence, it suffices to check that, for some constant C, we have:

$$\forall k , \quad \alpha_{D,k+1} - \alpha_{D,k} \leq C \cdot 2^{-D} .$$

This inequality follows immediately from a recurrence based on the next lemma.

Lemma 23 *Consider $f : [0,1] \to [0,1]$ a twice differentiable and concave function such that: $m \leq f'' \leq M < 0$. Take x_0, x_1 such that $x_0 < x_1$ and set x_* such that*

$$f'(x_*) = \frac{f(x_1) - f(x_0)}{x_1 - x_0} .$$

Then, we have:

$$x_* - x_0 \leq C(x_1 - x_0)$$

for some constant C which does not depend on x_0, x_1.

PROOF. Set the notations: $\Delta f = f(x_1) - f(x_0)$ and $\Delta x = x_1 - x_0$ As f' is continuous and strictly increasing, we can use the following expression for x_*:

$$x_* = f'^{-1}\left(\frac{\Delta f}{\Delta x}\right) .$$

By applying the theorem of finite increment to f'^{-1} between and $f'(x_1)$ and $\frac{\Delta f}{\Delta x}$, we have

$$x_* - x_0 = x_1 - x_0 + \left(\frac{\Delta f}{\Delta x} - f'(x_1)\right)(f'^{-1})'(c)$$

for some c. But we also have by Taylor's formula that:

$$\frac{\Delta f}{\Delta x} - f'(x_1) = \frac{1}{2}(x_1 - x_0) f''(c')$$

for some c'. This leads to the result, as $m \leq f'' \leq M < 0$ since:

$$(f'^{-1})' = \frac{f''}{f'' \circ (f')^{-1}} .$$

∎

Proof of Proposition 19

The ROC curve of $s_{2,t}$ is a broken line with the extremities of the two linear parts being $(0,0)$, $(\alpha(t), \beta(t))$ and $(1,1)$. Hence, the corresponding AUC can be written as:

$$A_2(t) = \frac{1}{2}\left(1 + \beta(t) - \alpha(t)\right) \ .$$

As the ROC curve is differentiable, the maximum of $A_2(t)$ is obtained at the point t^* such that:

$$d\beta(t^*) = d\alpha(t^*) \ ,$$

and hence $\frac{d}{d\alpha}\mathrm{ROC}^*(\alpha^*) = 1$ for $\alpha^* = \alpha(t^*)$. We use Proposition 8 to get $\alpha^* = \bar{H}^*(p)$ and this ultimately leads to $t^* = p$.

References

[AGH+05] S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization bounds for the area under the ROC curve. *Journal of Machine Learning Research*, 6:393–425, 2005.

[BFOS84] L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

[CLV05] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and scoring using empirical risk minimization. In P. Auer and R. Meir, editors, *Proceedings of COLT 2005*, volume 3559 of *Lecture Notes in Computer Science*, pages 1–15. Springer, 2005.

[CLVar] S. Clémençon, G. Lugosi, and N. Vayatis. Ranking and empirical risk minimization of U-statistics. *The Annals of Statistics*, To appear, To appear.

[CM04] C. Cortes and M. Mohri. Auc optimization vs. error rate minimization. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[CS01] K. Crammer and Y. Singer. Pranking with ranking. In *Proceedings of the conference on Neural Information Processing Systems (NIPS)*, 2001.

[CSS98] W.W. Cohen, R.E. Schapire, and Y. Singer. Learning to order things. In *NIPS '97: Proceedings of the 1997 conference on Advances in neural information processing systems 10*, pages 451–457, Cambridge, MA, USA, 1998. MIT Press.

[CV07] S. Clémençon and N. Vayatis. Ranking the best instances. *Journal of Machine Learning Research*, 8:2671–2699, 2007.

[dEW06] M. desJardins, E. Eaton, and K. Wagstaff. Learning user preferences for sets of objects. In *Proceedings of the Twenty-Third International Conference (ICML 2006)*, pages 273–280, 2006.

[DGL96] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[DL93] R. Devore and G. Lorentz. *Constructive Approximation*. Springer, 1993.

[DS66] D.M.Green and J.A. Swets. *Signal detection theory and psychophysics*. Wiley, 1966.

[Ega75] J.P. Egan. *Signal Detection Theory and ROC Analysis*. Academic Press, 1975.

[FFHO02] C. Ferri, P.A. Flach, and J. Hernández-Orallo. Learning decision trees using the area under the roc curve. In *ICML '02: Proceedings of the Nineteenth International Conference on Machine Learning*, pages 139–146, San Francisco, CA, USA, 2002. Morgan Kaufmann Publishers Inc.

[FISS03] Y. Freund, R. D. Iyer, R. E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, November 2003.

[GKKW02] L. Györfi, M. Köhler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.

[HGBSO98] R. Herbrich, T. Graepel, P. Bollmann-Sdorra, and K. Obermayer. Learning a preference relation for information retrieval. In *Proceedings of the AAAI Workshop Text Categorization and Machine Learning*, 1998.

[HGO00] R. Herbrich, T. Graepel, and K. Obermayer. *Advances in Large Margin Classifiers*, chapter Large margin rank boundaries for ordinal regression, pages 115–132. MIT Press, 2000.

[HM82] J.A. Hanley and J. McNeil. The meaning and use of the area under a ROC curve. *Radiology*, (143):29–36, 1982.

[HT90] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Chapman & Hall/CRC, 1990.

[HT96] D. Hsieh and B.W. Turnbull. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. *The Annals of Statistics*, 24:25–40, 1996.

[PD03] F. Provost and P. Domingos. Tree induction for probability-based ranking. *Machine Learning*, 52(3):199–215, 2003.

[Rak04] A. Rakotomamonjy. Optimizing area under roc curve with svms. In *Proceedings of the First Workshop on ROC Analysis in AI*, 2004.

[vT68] H.L. van Trees. *Detection, Estimation, and Modulation Theory, Part I*. John Wiley, 1968.

[XZW06] F. Xia, W. Zhang, and J. Wang. An effective tree-based algorithm for ordinal regression. *IEEE Intelligent Informatics Bulletin*, 7(1):22–26, December 2006.

[YDMW03] L. Yan, R.H. Dodier, M. Mozer, and R.H. Wolniewicz. Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic. In Tom Fawcett and Nina Mishra, editors, *Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003)*, pages 848–855, 2003.