On Partitioning Rules for Bipartite Ranking

Stéphan Clémençon
LTCI - Telecom/CNRS UMR 5141
Telecom ParisTech, France

Nicolas Vayatis
CMLA - CNRS UMR 8536
ENS Cachan & UniverSud, France

Abstract

The purpose of this paper is to investigate the properties of partitioning scoring rules in the bipartite ranking setup. We focus on ranking rules based on scoring functions. General sufficient conditions for the AUC consistency of scoring functions that are constant on cells of a partition of the feature space are provided. Rate bounds are obtained for cubic histogram scoring rules under mild smoothness assumptions on the regression function. In this setup, it is shown how to penalize the empirical AUC criterion in order to select a scoring rule nearly as good as the one that can be built when the degree of smoothness of the regression function is known.

1 Introduction

In this paper, we consider the ranking problem with classification data, also known as the *bipartite ranking problem* (Freund et al. (2003), Agarwal et al. (2005), Clémençon et al. (2008)). Our perspective follows the scoring approach where ranking rules are based on real-valued scoring functions. We focus here on scoring functions which take discrete values. For the sake of interpretability, it is indeed of practical importance in many ranking applications (medical diagnosis, credit-risk screening, marketing) to segment the population in ordered "strata" with distinct features. Data-dependent partitioning schemes have a long history in the classical statistical problems like density estimation, regression and classification (see Devroye et al. (1996) and references therein). However the ranking problem presents some specific features due to the nature of performance measures which are involved to assess the quality of a ranking/scoring rule. Central tools for the comparison of scoring functions are the ROC curve and its scalar summary known as the AUC (Huang and Ling (2005)). The present paper examines the conditions under which data-dependent partitioning techniques yield an AUC consistent ranking. Note that the AUC corresponds to the L_1-distance in the ROC space.

We first relate partitioning of the feature space to the approximation/estimation issue of the optimal ROC curve by piecewise constants in the L_1-sense. As a preliminary, it is proved that the scoring rule with maximum (empirical) AUC among the ones that are constant on each cell of a given partition may be described as a *plug-in* rule. The deficit of AUC is then related to the approximation error of the regression estimate in the L_1-sense. Consistency results and rate bounds are then established for regular histogram rules under smoothness assumptions on the regression function. In the work presented here, it is important to note that optimal scoring functions are not assumed to be contained in the class of candidate scoring rules, namely the collection of piecewise constant scoring functions. Eventually, we propose a penalization method for the empirical AUC maximization in order to select the size of the cubes forming the partition. The estimator obtained by the penalized criterion adapts to the degree of smoothness of the regression function.

The paper is structured as follows. In Section 2 notations are set out, important concepts of ROC analysis are briefly recalled and preliminary results related to partition-based scoring rules are established. The main results of the paper are stated in Section 3. Consistency of cubic histogram scoring rules is proved under mild conditions, while a rate bound is established in the ideal case when the degree of smoothness θ of the regression function is known. A penalization approach is considered in order to exhibit a data-driven method

Appearing in Proceedings of the 12^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2009, Clearwater Beach, Florida, USA. Volume 5 of JMLR: W&CP 5. Copyright 2009 by the authors.

for selecting a scoring rule that exhibits the same rate bound without the knowledge of θ. Technical proofs are postponed to the Appendix.

2 Background and Preliminaries

The probabilistic framework is exactly the same as the one in standard binary classification. We denote by (X, Y) a pair of random variables where $Y \in \{-1, +1\}$ is a binary label and X models some observation for predicting Y, taking its values in a feature space $\mathcal{X} \subset \mathbb{R}^d$ of high dimension. Here and throughout, $\mathcal{L}$ denotes the joint distribution of (X, Y) and $p = \mathbb{P}\{Y = +1\}$. The probability distribution $\mathcal{L}$ is entirely determined by the pair (μ, η) where μ denotes the marginal distribution of X and the regression function $\eta(x) = \mathbb{P}\{Y = +1 \mid X = x\}$, $x \in \mathcal{X}$. We also introduce $G(dx)$ and $H(dx)$, the conditional distributions of X given $Y = +1$ and $Y = -1$ respectively. Through the paper, these probability measures are assumed to be equivalent. Observe that, with these notations, $\eta(x) = p(dG/dH)(x)/(1-p+p(dG(x)/dH)(x))$ and $\mu(dx) = pG(dx) + (1-p)H(dx)$.

2.1 Bipartite ranking

We start off with a brief description of the bipartite ranking task and recall the basic concepts related to this statistical learning problem.

Based on the observation of i.i.d. examples $\mathcal{D}_n = \{(X_i, Y_i) : 1 \leq i \leq n\}$, the goal is to learn how to order all instances $x \in \mathcal{X}$ in a way that instances X such that $Y = +1$ appear on top in the list with the largest possible probability . Clearly, the simplest way of defining an order relationship on $\mathcal{X}$ is to transport the natural order on the real line to the feature space through a *scoring rule* $s : \mathcal{X} \to \mathbb{R}$. The notion of ROC curve, which we recall below, provides a functional criterion for evaluating the performance of the ordering induced by such a function. Here and throughout, we denote by $F^{-1}(t) = \inf\{u \in \mathbb{R} : F(u) \geq t\}$ the pseudo-inverse of any càd-làg increasing function $F : \mathbb{R} \to \mathbb{R}$ and by $\mathcal{S}$ the set of all scoring functions, *i.e.* the space of real-valued measurable functions on $\mathcal{X}$.

Definition 1 (ROC CURVE) *Let $s \in \mathcal{S}$. The* ROC *curve of the scoring function $s(x)$ is given by*

$$\alpha \in [0,1] \mapsto \mathrm{ROC}(s, \alpha) = 1 - G_s \circ F_s^{-1}(1-\alpha),$$

where $G_s(dx)$ and $H_s(dx)$ denote the conditional distributions of $s(X)$ given $Y = +1$ and given $Y = -1$ respectively.

When $G_s(du)$ and $H_s(du)$ are both continuous distributions, the ROC curve of $s(x)$ is nothing else than the PP-plot:

$$t \mapsto (\mathbb{P}\{s(X) \geq t \mid Y = -1\}, \mathbb{P}\{s(X) \geq t \mid Y = +1\}).$$

It is a well-known result in ROC analysis that increasing transforms of the regression function $\eta(x)$ form the class $\mathcal{S}^*$ of optimal scoring functions in the sense that their ROC curve, namely $\mathrm{ROC}^* = \mathrm{ROC}(\eta, .)$, dominates the ROC curve of any other scoring function $s(x)$ everywhere:

$$\forall \alpha \in [0,1[, \ \ \mathrm{ROC}(s,\alpha) \leq \mathrm{ROC}^*(\alpha).$$

The proof of this fact is based on a simple application of Neyman-Pearson's lemma (see Clémençon and Vayatis (2008b)). It is noteworthy that, when continuous, the curve ROC^* is concave. More generally, for any scoring function $s(x)$, $\mathrm{ROC}(s, .)$ is a concave curve as soon as $G_s(du)$ and $H_s(du)$ are continuous distributions and the likelihood ratio $dG_s/dH_s(s(X))$ is monotone.

In practice, the functional performance measure described above is generally summarized by a scalar feature, the Area Under the ROC Curve (AUC in abbreviated form).

Definition 2 (THE AUC CRITERION) *Let $s(x)$ be a scoring function. The Area Under the* ROC *Curve is given by*

$$\mathrm{AUC}(s) = \int_{\alpha=0}^{1} \mathrm{ROC}(s, \alpha) d\alpha.$$

Of course, $\mathcal{S}^*$ corresponds to the set of scoring functions with maximum AUC. We set:

$$\forall s \in \mathcal{S}^*, \ \ \mathrm{AUC}^* = \mathrm{AUC}(s).$$

In Clémençon et al. (2008), it has been shown that, when $\eta(X)$ has a continuous distribution, the maximal AUC may be related to $\eta(X)$'s dispersion through:

$$\mathrm{AUC}^* = \frac{1}{2} + \frac{\mathbb{E}\left(|\eta(X) - \eta(X')|\right)}{4p(1-p)},$$

where X' is an independent copy of X. The quantity $\mathbb{E}\left(|\eta(X) - \eta(X')|\right)$ is known as the *Gini mean difference* of $\eta(X)$. Hence, the more concentrated is $\eta(X)$, the more difficult the bipartite ranking problem.

The popularity of the AUC criterion mainly arises from the fact that it may be interpreted in a probabilistic manner.

Proposition 3 *The* AUC *criterion may be viewed as the rate of concordant pairs. For any scoring function $s(x)$, we have:*

$$\mathrm{AUC}(s) = \mathbb{P}\{s(X) > s(X') \mid (Y, Y') = (+1, -1)\},$$

where (X', Y') denotes a copy of the pair (X, Y), independent from the latter.

2.2 Piecewise constant scoring functions

Here we focus on the simplest scoring functions, namely real-valued *piecewise constant* functions on the feature space $\mathcal{X}$. Any scoring function $s(x)$ of this type, taking $K \geq 1$ distinct values say, yields a ranking/ordering of all instances $x \in \mathcal{X}$ entirely characterized by a partition $\mathcal{P}$ with K nonempty measurable subsets $C_1, \ldots, C_K$, together with a permutation σ in the symmetric group Σ_K of $\{1, \ldots, K\}$.

Definition 4 ($(\mathcal{P}, \sigma)$-REPRESENTATION) *The $(\mathcal{P}, \sigma)$-representation of a piecewise constant scoring function $s(x)$ taking K distinct values $\lambda_1 > \ldots > \lambda_K$ is given by:*

$$s(x) = \sum_{k=1}^{K} \lambda_k \cdot \mathbb{I}\{x \in C_{\sigma(k)}\}, \qquad (1)$$

where $\mathcal{P} = \{C_k\}_{1 \leq k \leq K}$ is a partition of $\mathcal{X}$ in K non empty cells and σ a permutation of $\{1, \ldots, K\}$.

Reciprocally, a partition $\mathcal{P} = \{C_1, \ldots, C_K\}$ with $\#\mathcal{P} = K$ non empty cells combined with a permutation $\sigma \in \Sigma_K$ defines a scoring function with $(\mathcal{P}, \sigma)$-representation:

$$s_{\mathcal{P},\sigma}(x) = \sum_{k=1}^{K} (K - k + 1) \cdot \mathbb{I}\{x \in C_{\sigma(k)}\}.$$

The ordering induced by (1) is entirely characterized by the pair $(\mathcal{P}, \sigma)$, in the sense that its ROC curve coincides with $\mathrm{ROC}(s_{\mathcal{P},\sigma}, .)$

Remark 1 *(*A GLOBAL LEARNING PROBLEM*) In contrast to binary classification, the bipartite ranking problem is of global nature. Indeed, in classification, a decision rule may be immediately derived from a partition $\mathcal{P}$ of the feature space through a majority vote scheme. Here, the local properties of the regression function on a given cell are useless, since cells of $\mathcal{P}$ have somehow to be compared one to each other.*

The next results provide some basic properties of piecewise constant scoring functions. In order to formulate them precisely, we introduce the following notations. We set

$$\begin{aligned} \alpha(C) &= \mathbb{P}\{X \in C \mid Y = -1\}, \\ \beta(C) &= \mathbb{P}\{X \in C \mid Y = +1\}, \end{aligned}$$

for any a measurable subset $C \subset \mathcal{X}$. In the following proposition, the ROC curve of a piecewise constant scoring function and the corresponding AUC are made explicit.

Proposition 5 *Let $s(x)$ be a piecewise constant scoring function with $(\mathcal{P}, \sigma)$-representation $s(x) = \sum_{k=1}^{K} \lambda_k \cdot \mathbb{I}\{x \in C_{\sigma(k)}\}$. Its* ROC *curve is the stepwise function*

$$\mathrm{ROC}(s, \alpha) = \sum_{k=1}^{K-1} \beta_k(s) \cdot \mathbb{I}\{\alpha \in [\alpha_k(s), \alpha_{k+1}(s)[\}, \qquad (2)$$

where: $\forall k \in \{1, \ldots, K\}$,

$$\alpha_k(s) \;=\; \sum_{l=1}^{k} \alpha(C_{\sigma(l)}) \text{ and } \beta_k(s) = \sum_{l=1}^{k} \beta(C_{\sigma(l)}),$$

with $\alpha_0(s) = \beta_0(s) = 0$ by convention. Its AUC *is:*

$$\mathrm{AUC}(s) \;=\; \sum_{k=1}^{K-1} (\alpha_{k+1}(s) - \alpha_k(s)) \cdot \beta_k(s). \qquad (3)$$

Remark 2 (PIECEWISE LINEAR ROC CURVES) *We point out that in the case of a piecewise constant scoring function (i.e. when the distributions G_s and H_s are degenerate), another usual convention for representing its* ROC *curve consists in plotting the broken line $\widetilde{\mathrm{ROC}}(s, .)$ that connects the knots $\{(\alpha_k, \beta_k) : k = 0, \ldots, K\}$, see Clémençon and Vayatis (2008b,a). With this convention, the area under the* ROC *curve may be expressed as $\int_{\alpha=0}^{1} \widetilde{\mathrm{ROC}}(s, \alpha) d\alpha = \mathbb{P}\{s(X) > s(X') \mid (Y, Y') = (1, -1)\} + \frac{1}{2}\mathbb{P}\{s(X) = s(X') \mid (Y, Y') = (1, -1)\}$.*

The next result result describes the best scoring function in the AUC sense among all piecewise constant scoring functions that may be represented by means of a given partition $\mathcal{P}$. We denote by $\mathcal{S}_{\mathcal{P}}$ the set of scoring functions with a $(\mathcal{P}, \sigma)$-representation for some $\sigma \in S_{\#\mathcal{P}}$.

Theorem 6 *(*AUC OPTIMALITY*) Consider a partition of $\mathcal{X}$ with $K \geq 1$ non empty cells: $\mathcal{P} = \{C_k\}_{1 \leq k \leq K}$. Let $\sigma^* \in \Sigma_K$ such that*

$$\frac{\beta(C_{\sigma^*(1)})}{\alpha(C_{\sigma^*(1)})} \geq \ldots \geq \frac{\beta(C_{\sigma^*(K)})}{\alpha(C_{\sigma^*(K)})}.$$

Then, $s^*_{\mathcal{P}}(x) = s_{\mathcal{P},\sigma^*}(x)$ maximizes the AUC *over $\mathcal{S}_{\mathcal{P}}$:*

$$\mathrm{AUC}(s^*_{\mathcal{P}}) = \max_{s \in \mathcal{S}_{\mathcal{P}}} \mathrm{AUC}(s).$$

In the case where the cells are equivalent with respect to the false positive rate, i.e. $\forall k \in \{1, \ldots, K\}$: $\alpha(C_k) = 1/K$, we also have

$$\forall \alpha \in [0, 1],\ \mathrm{ROC}(s, \alpha) \leq \mathrm{ROC}(s^*_{\mathcal{P}}, \alpha),$$

for all $s \in \mathcal{S}_{\mathcal{P}}$. The latter result also holds when cells are equivalent with respect to the true positive rate.

Remark 3 *(*ON CONCAVITY*) It is noteworthy that σ^* corresponds to the permutation which makes the linear-by-part curve $\widetilde{\mathrm{ROC}}(s_{\mathcal{P},\sigma}, .)$ concave.*

To any partition $\mathcal{P} = \{C_k\}_{1\leq k\leq K}$ of $\mathcal{X}$ also correspond piecewise constant approximants of the regression function, which may serve as scoring functions. For instance,

$$\eta_{\mathcal{P}}(x) = \sum_{k=1}^{K} \frac{p\beta(C_k)}{\mu(C_k)} \cdot \mathbb{I}\{x \in C_k\}, \tag{4}$$

is the best approximant among functions that are constant on the cells C_k in the $L_2(\mu)$-sense, *i.e.* $||\eta_{\mathcal{P}}(X) - \eta(X)||^2_{L_2(\mu)} = \min_{s\in\mathcal{S}_{\mathcal{P}}} \mathbb{E}[(s(X)-\eta(X))^2]$. It follows from the fact that $\mu(C_k) = p\alpha(C_k) + (1-p)\beta(C_k)$ for all k that the *plug-in* scoring function $\eta_{\mathcal{P}}(x)$ yields the same ranking as $s^*_{\mathcal{P}}(x)$. Hence, the next result immediately derives from Theorem 6.

Corollary 7 *(*PLUG-IN SCORING RULE*) As a scoring function, the approximant $\eta_{\mathcal{P}}(x)$ of the regression function is optimal in the* AUC *sense among all scoring rules in $\mathcal{S}_{\mathcal{P}}$:*

$$\mathrm{AUC}(\eta_{\mathcal{P}}) = \max_{s\in\mathcal{S}_{\mathcal{P}}} \mathrm{AUC}(s).$$

2.3 Approximation of the optimal ROC curve by piecewise constants

Consider the following approximant of the curve ROC*, which is piecewise constant with breakpoints from a given subdivision $\pi : \alpha_0 = 0 < \alpha_1 < \ldots < \alpha_{K-1} < \alpha_K = 1$ of $[0,1]$:

$$R_\pi(\alpha) = \sum_{k=0} \mathrm{ROC}^*(\alpha_k) \cdot \mathbb{I}\{\alpha \in [\alpha_k, \alpha_{k+1}[\} \ . \tag{5}$$

Assuming that ROC* is Lipschitz with constant κ (*i.e.* $\forall(\alpha,\alpha')$, $|\mathrm{ROC}^*(\alpha') - \mathrm{ROC}^*(\alpha)| \leq \kappa|\alpha-\alpha'|$) and the meshgrid is such that $\max_{1\leq k\leq K}\{\alpha_k - \alpha_{k-1}\} \leq M/K$ for some constant $M < \infty$, the L_1-error $||R_\pi - \mathrm{ROC}^*||_1 = \int_{\alpha=0}^{1} |R_\pi(\alpha) - \mathrm{ROC}^*(\alpha)|d\alpha$ is less than $M\kappa/K$, which corresponds, in absence of further hypothesis about the curve ROC*, to the optimal approximation rate, in the minimax sense, using stepwise constant approximants with K pieces, see Devore (1998).

Remark 4 *(*ON THE LIPSCHITZ CONDITION*) It is noteworthy that the Lipschitz assumption for* ROC* *is guaranteed, as soon as G_η (respectively, H_η) has a density that is bounded (resp., bounded by below by a strictly positive constant).*

We point out that R_π actually corresponds to the ROC curve of the piecewise constant scoring function $\eta_{\mathcal{P}_\pi}(x)$, where $\mathcal{P}_\pi$ is the partition defined by: $\forall k \in \{1,\ldots,K\}$,

$$C^*_{\pi,k} = \{Q^*(\alpha_k) < \eta(x) < Q^*(\alpha_{k-1})\},$$

denoting by $Q^*(\alpha)$ the $(1-\alpha)$-quantile of the conditional distribution of $\eta(X)$ given $Y=-1$. The bound for the L_1-approximation error mentioned above may be then rewritten:

$$\mathrm{AUC}^* - \mathrm{AUC}(s^*_{\mathcal{P}_\pi}) \leq \frac{M\kappa}{K} \ . \tag{6}$$

The following result describes the performance of the ROC curve of a piecewise constant scoring function as an approximant of the optimal curve ROC* in the ROC space in terms of L_1-distance.

Theorem 8 *Suppose that the random variable $\eta(X)$ is continuous. Consider a piecewise constant scoring function with $(\mathcal{P},\sigma)$-representation $s(x) = \sum_{k=1}^{K} \lambda_k \cdot \mathbb{I}\{x \in C_{\sigma(k)}\}$. Then, we have*

$$\mathrm{AUC}^* - \mathrm{AUC}(s) = \frac{\mathbb{E}[|\eta(X)-\eta(X')|]\mathbb{I}\{(X,X')\in\Gamma_s\}}{2p(1-p)} + \frac{1}{4p(1-p)}\sum_{k=1}^{K} G_k + \frac{1}{2}\sum_{k=1}^{K}\alpha(C_k)\beta(C_k) \ ,$$

where $\Gamma_s \subset \mathcal{X}^2$ denotes the set of couples (x,x') such that $(\eta(x)-\eta(x'))\cdot(s(x)-s(x')) < 0$ and G_k denotes the Gini mean difference of $\eta(X)$ with the expectation restricted to the domain $\{(X,X') \in C_k \times C_k\}$.

Remark 5 (ON RANKING ACCURACY) *We point out that the term G_k involved in the* AUC *deviation measures to which extent $\eta(X)$ may be accurately approximated by a constant over the cell C_k. Observe also that, when applied to the scoring rule $s^*_{\mathcal{P}_\pi}$, the first term on the right hand side of the equation vanishes.*

The first part of the next result relates the deficit of AUC for optimal piecewise constant scoring functions $s^*_{\mathcal{P}}$ to the L_1 error of the corresponding plug-in estimator $\eta_{\mathcal{P}}$ in estimating the regression function η. In the second part, a precise rate bound is given in the case where μ has a bounded support and the partition considered is uniform. For simplicity, we take $\mathcal{X} = [0,1]^d$ and consider the partition $\mathcal{P}_j$, made of dyadic cubes with side length 2^{-j} (in this case $\#\mathcal{P}_j = 2^{jd}$). Assumptions related to the smoothness properties of η with respect to μ are naturally required to establish an approximation rate. As will be seen in the subsequent analysis, this permits to control the bias of grid-based ranking rules. Following in the footsteps of Binev et al. (2005), consider the space $\mathcal{A}_\theta(\mu)$ consisting of all functions $f \in L_2(\mu)$ for which there exists $M < \infty$ such that:

$$\forall j \in \mathbb{N}, \ ||f - f_{\mathcal{P}_j}||_{L_2(\mu)} \leq M \cdot (\#\mathcal{P}_j)^{-\theta},$$

where $f_{\mathcal{P}_j}$ the orthogonal projection of f onto $\mathcal{S}_{\mathcal{P}_j}$, viewed as a subspace of the Hilbert space $L_2(\mu)$. We denote by $|f|_{\mathcal{A}_\theta(\mu)}$ the smallest constant M for which this bound holds for all $j \geq 0$. We point out that, when μ is the Lebesgue measure, this approximation space corresponds to the usual Besov space $\mathcal{B}^{\theta d}_{2,\infty}(\mathcal{X})$[1].

Corollary 9 *Assume that $\eta(X)$ has a continuous distribution.*

(i) For any partition $\mathcal{P} = \{C_k\}_{1\leq k\leq K}$ with $K \geq 2$ non empty cells, we have:

$$\mathrm{AUC}^* - \mathrm{AUC}(s^*_{\mathcal{P}}) \leq \frac{||\eta_{\mathcal{P}}(X) - \eta(X)||_{L_1(\mu)}}{p(p-1)} + \frac{1}{4p(1-p)}\sum_{k=1}^K G_k + \frac{1}{2}\sum_{k=1}^K \alpha(C_k)\beta(C_k) \ .$$

(ii) Assume that $\mathcal{X} = [0,1]^d$. Suppose in addition $\mu(dx)$ has a bounded density with respect to Lebesgue measure and that $\eta(x)$ belongs to $\mathcal{A}_\theta(\mu)$ with $0 < \theta \leq 1$. Then, there exists a constant $c < \infty$ independent from j such that: $\forall j \geq 1$,

$$\mathrm{AUC}^* - \mathrm{AUC}(s^*_{\mathcal{P}_j}) \leq c \cdot (\#\mathcal{P}_j)^{-\theta} \ .$$

In the case $\theta = 1$, the scoring function $s^*_{\mathcal{P}_j}(x)$ achieves the same approximation rate as $s_{\mathcal{P}_\pi}$ when the subdivision π has cardinality 2^{jd}, see Eq. (6), even though the assumptions guaranteeing these rate bounds are not exactly of the same nature. However, we point out that one may exhibit conditions related to $\eta(x)$ and to the smoothness properties of H and G ensuring the Lipschitz property of ROC*, see Remark 5 in Clémençon and Vayatis (2007).

3 Empirical partitioning techniques.

We now turn to the statistical problem. From a practical perspective, the selection of a scoring function $s(x)$ is based on training data $\mathcal{D}_n = \{(X_i, Y_i);\ 1 \leq i \leq n\}$. The relevance of a candidate $s(x)$ is thus evaluated by plotting the empirical version of its ROC curve. We set:

$$\hat{\alpha}_i(s) = \frac{1}{n_-}\sum_{j/\ Y_j=-1} \mathbb{I}\{s(X_j) \geq s(X_i)\} \ ,$$

$$\hat{\beta}_i(s) = \frac{1}{n_+}\sum_{j/\ Y_j=+1} \mathbb{I}\{s(X_j) \geq s(X_i)\} \ ,$$

for all $i \in \{1,\dots,n\}$ where $n_+ = \sum_{i\leq n}\mathbb{I}\{Y_i = +1\} = n - n_-$. The empirical ROC curve of $s(x)$ is the stepwise function given by: $\forall \alpha \in [0,1]$,

$$\widehat{\mathrm{ROC}}(s,\alpha) = \sum_{i=1}^n \hat{\beta}_{\sigma(i)}(s)\cdot \mathbb{I}\{\alpha \in [\hat{\alpha}_{\sigma(i)}(s), \hat{\alpha}_{\sigma(i+1)}(s)[\},$$

where $\sigma \in S_n$ is such that: $\hat{\alpha}_{\sigma(1)} \leq \ldots \leq \hat{\alpha}_{\sigma(n)}$, with the convention that $\hat{\alpha}_{\sigma(0)}(s) = \hat{\beta}_{\sigma(0)}(s) = 0$ and $\hat{\alpha}_{\sigma(n+1)}(s) = \hat{\beta}_{\sigma(n+1)}(s) = 1$.

By definition, the empirical AUC of $s(x)$ is the area under its empirical ROC curve, namely the *rate of concordant pairs*:

$$\begin{aligned}\widehat{\mathrm{AUC}}(s) &= \int_{\alpha=0}^1 \widehat{\mathrm{ROC}}(s,\alpha)d\alpha, \\ &= \frac{1}{n_+n_-}\sum_{i/\ Y_i=+1}\sum_{j/\ Y_j=-1}\mathbb{I}\{s(X_i) > s(X_j)\}.\end{aligned}$$

All results established when considering true ROC curves extend to their empirical versions, replacing G, H and p by their counterparts calculated from the sample $\mathcal{D}_n$. In particular, given a partition $\mathcal{P} = \{C_k\}_{1\leq k\leq K}$ of the feature space $\mathcal{X}$, the ordering of the cells with maximum empirical AUC corresponds to permutations $\widehat{\sigma}^*$ such that,

$$\frac{\widehat{\beta}(C_{\widehat{\sigma}^*(1)})}{\widehat{\alpha}(C_{\widehat{\sigma}^*(1)})} \geq \ldots \geq \frac{\widehat{\beta}(C_{\widehat{\sigma}^*(K)})}{\widehat{\alpha}(C_{\widehat{\sigma}^*(K)})},$$

where for all measurable subset $C \subset \mathcal{X}$:

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

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

It renders the empirical ROC curve concave and corresponds to the same ranking induced by the estimator of the regression function

$$\widehat{\eta}_{\mathcal{P}}(x) = \sum_{k=1}^K \frac{n_+\widehat{\beta}(C_k)}{n_-\widehat{\alpha}(C_k) + n_+\widehat{\beta}(C_k)}\cdot \mathbb{I}\{x \in C_k\},$$

meaning that $\widehat{\eta}_{\mathcal{P}} = \arg\max_{s\in\mathcal{S}_{\mathcal{P}}} \widehat{\mathrm{AUC}}(s)$.

3.1 Histogram ranking rules

The next result shows how general consistency results may be typically established for a large class of piecewise constant scoring rules, when the partitions are fixed in advance.

[1] Recall that the Besov space $\mathcal{B}^{\theta d}_{2,\infty}(\mathcal{X})$ is the set of Borel functions $f : \mathcal{X} \subset \mathbb{R}^d \to \mathbb{R}$ for which there exists $M < \infty$ such that: $\forall h \in \mathcal{X}$, $(\int_{x\in\mathcal{X}:\ x+h\in X}(f(x+h)-f(x))^2dx)^{1/2} \leq M|h|^{\theta d}$, where $|.|$ denotes the euclidian norm on $\mathbb{R}^d$. The smallest constant M for which the latter bound holds is the norm $||f||_{\mathcal{B}^{\theta}_{2,\infty}(\mathcal{X})}$.

Theorem 10 (Histogram scoring rules) *Consider $\mathcal{X} = [0,1]^d$ and $\mathcal{P}_j$ the partition composed with dyadic cubes. Assume that $\eta(X)$ has a continuous distribution and that $\mu(dx)$ has a bounded density with respect to Lebesgue measure. For the piecewise constant scoring function defined as $\widehat{s}^*_{\mathcal{P}_j} = s_{\mathcal{P}_j,\widehat{\sigma}^*}$, we have the following properties:*

(i) (Consistency) *If $j(n)$ tends to infinity so that $n2^{-j(n)d} \to 0$ as $n \to \infty$, the scoring rule $\widehat{s}^*_{\mathcal{P}_{j(n)}}$ is consistent in the* AUC *sense:*

$$\mathbb{E}[\mathrm{AUC}(\widehat{s}^*_{\mathcal{P}_{j(n)}})] \to \mathrm{AUC}^*.$$

(ii) (Rate bound) *Suppose that $\eta \in \mathcal{A}_\theta(\mu)$. If $j(n)$ is picked in a way that $2^{j(n)d} \sim n^{\frac{1}{1+2\theta}}$, then there exists a constant $c < \infty$ such that for all $\delta \in]0,1[$, we have with probability at least $1-\delta$: $\forall n \geq 1$,*

$$\mathrm{AUC}(\widehat{s}^*_{\mathcal{P}_{j(n)}}) - \mathrm{AUC}^* \leq c\sqrt{\frac{\log(n/\delta)}{n^{2\theta/(1+2\theta)}}}. \qquad (7)$$

The main drawback of choosing in advance the side length $2^{j(n)}$ of the cells of the partition lies in the fact that the smoothness class which η belongs to has to be known in order to achieve the optimal rate. In the next subsection, we tackle the problem of building a scoring rule that achieves the same rate without knowing the exact amount of smoothness of η.

Remark 6 (Optimality) *Although no lower bound result is available for this problem, we point out that, up to our knowledge, the best possible rate (up to a logarithmic factor) is of the order $n^{1/3}$ for $\theta = 1$.*

3.2 Selecting the best partition for scoring

Model selection procedures have been successfully developed in the statistical learning setup for binary classification (Massart (2006); Boucheron et al. (2005)). We propose here a similar strategy for selecting the partition $\mathcal{P}_{\widehat{j}}$ among all dyadic partitions $\{\mathcal{P}_j\}_j$ in a data-driven fashion and with best possible rate. In our setup, the model selection procedure takes the following form:

$$\widehat{\mathrm{CPAUC}}(\mathcal{P}_j) = \widehat{\mathrm{AUC}}(\widehat{s}^*_{\mathcal{P}_j}) - \mathrm{pen}(j,n) \ ,$$

where $\mathrm{pen}(j,n)$ is a penalty term. We set

$$\widehat{j}(n) = \arg\max_{j \geq 1} \widehat{\mathrm{CPAUC}}(\mathcal{P}_{j(n)}) \quad \text{and } \widehat{s} = \widehat{s}^*_{\mathcal{P}_{\widehat{j}(n)}}$$

The argument invoked in this analysis could be applied to more flexible partitions and serve as a basis for investigating the pruning stage of the TreeRank algorithm proposed by Clémençon and Vayatis (2008b).

The key to the study of the AUC performance of the selected scoring rule is the following *oracle inequality.*

Proposition 11 (Oracle inequality) *Suppose that the penalty term is picked so that: $\forall j \geq 1$, $pen(j,n) \geq \frac{1}{p(1-p)}\sqrt{\frac{\log(2^{jd+1}!)}{2n}}$. Then we have:*

$$\mathbb{E}[\mathrm{AUC}^* - \mathrm{AUC}(\widehat{s})] \leq \inf_{j \geq 1} \mathcal{B}(j,n), \qquad (8)$$

where for all $j \geq 1$,

$$\mathcal{B}(j,n) = \mathrm{AUC}^* - \mathrm{AUC}(s^*_{\mathcal{P}_j}) + 2pen(j,n) \ .$$

Remark 7 (On the penalty) *The penalty can be rendered independent of the distribution if it is assumed that the proportion p belongs to an interval $[\underline{p}, \bar{p}]$ with $0 < \underline{p} < \bar{p} < 1$.*

Remark 8 (On sharper bounds) *Under the specific noise condition proposed by Clémençon et al. (2008) in the ranking setup, the rate stated above could be refined, using a Bernstein's type inequality for U-statistics (see the proof in the Appendix). Due to space limitations, here we leave this question aside.*

The next result reveals that the scoring rule achieves almost the same rate of convergence. It results from Proposition 11 and Theorem 10 combined with Stirling's formula.

Theorem 12 *(*Model selection*) Suppose that assumptions of part (ii) of Corollary 9 and Proposition 11 are fulfilled. Then, there exists a constant $c > 0$ such that for all $\delta \in]0,1[$, we have, with probability at least $1-\delta$: $\forall n \geq 1$,*

$$\mathrm{AUC}^* - \mathrm{AUC}(\widehat{s}) \leq c\sqrt{\frac{\log(n/\delta)}{n^{2\theta/(1+2\theta)}}} \ . \qquad (9)$$

An alternative to model selection is to use adaptive partitioning algorithms. Owing to space limitations, they will be studied in a future work.

Appendix - Technical proofs

Proof of Theorem 6

The proof is based on the next lemma.

Lemma 13 *Let $\mathcal{P} = \{C_k\}_{1 \leq k \leq K}$ be a partition with $K \geq 2$ non empty cells. Consider $\sigma \in \Sigma_K$, fix $k \in \{1, \ldots, K-1\}$ and let $\tau_k \in \Sigma_K$ be the transposition exchanging k and $k+1$.Then, if $(\sigma(k) - \sigma(k+1)) \cdot (\sigma^*(k) - \sigma^*(k+1)) > 0$, we have*

$$\mathrm{AUC}(s_{\mathcal{P},\sigma}) \geq \mathrm{AUC}(s_{\mathcal{P},\sigma \circ \tau_k}).$$

PROOF. Without loss of generality, one may suppose that $\sigma(k) - \sigma(k+1)$ and $\sigma^*(k) - \sigma^*(k+1)$ are both nonnegative. It follows from the expression of the AUC stated in Proposition 5 that

$$\mathrm{AUC}(s_{\mathcal{P},\sigma}) - \mathrm{AUC}(s_{\mathcal{P},\sigma\circ\tau_k}) = \beta(C_{\sigma(k+1)})\alpha(C_{\sigma(k)}) - \beta(C_{\sigma(k)})\alpha(C_{\sigma(k+1)}) \ ,$$

and the latter quantity is negative by definition of σ^*. □

PROOF OF THE THEOREM. Observing that any permutation σ may be decomposed as $\sigma^* \circ \tau$, where τ is a compound of a finite number of transpositions τ_k, $k \in \{1, \dots K-1\}$, the proof of the first part of the theorem immediately follows from the lemma stated above. The second part straightforwardly results from Eq. (3) in Proposition 5. □

Proof of Theorem 8

Observe first that, for all scoring function s:

$$\mathrm{AUC}(s) = -\mathbb{P}\{s(X) = s(X') \mid (Y, Y') = (1, -1)\} + 1 - \frac{L(s)}{2p(1-p)} \ , \quad (10)$$

where $L(s) = \mathbb{P}\{(s(X) - s(X')) \cdot (Y - Y') < 0\}$. As $L(s)$ may be expressed as the expectation of $\eta(X)(1-\eta(X'))\mathbb{I}\{s(X) < s(X')\} + (1-\eta(X))\eta(X')\mathbb{I}\{s(X) > s(X')\}$ and $\eta(X)$ has a continuous distribution, one may check that

$$\begin{aligned} L(s) - L(\eta) &= \mathbb{E}\left[|\eta(X) - \eta(X')|\mathbb{I}\{(X, X') \in \Gamma_s\}\right] \\ &+ \frac{1}{2}\mathbb{E}[\mathbb{I}\{s(X) = s(X')\}|\eta(X) - \eta(X')|] \\ &- \mathbb{P}\{s(X) = s(X'), (Y, Y') = (1, -1)\}. \end{aligned}$$

Observe in addition that, when $s(x)$ admits a $(\mathcal{P}, \sigma)$-representation, one may write the second and third terms on the right hand side of the equation above as $\frac{1}{2}\sum_{C\in\mathcal{P}} \mathbb{E}(|\eta(X) - \eta(X')|\mathbb{I}\{(X, X') \in C^2\})$ and $\sum_{C\in\mathcal{P}} \alpha(C)\beta(C)$ respectively, which eventually concludes the proof. □

Proof of Corollary 9

Concerning the first assertion, observe that:
if $(X, X') \in \Gamma_s$, then

$$|\eta(X) - \eta(X')| \leq |\eta(X) - \widehat{\eta}(X)| + |\eta(X') - \widehat{\eta}(X')|.$$

Combined to Theorem 8, this establishes (i).

Turning to the second assertion, it follows from the hypothesis on μ and the smoothness assumption with respect to μ for $\eta(x)$ that: $\forall j \geq 1$,

$$||\eta - \eta_{\mathcal{P}_j}||_{L_2(\mu)} \leq |\eta|_{\mathcal{A}_\theta(\mu)} \cdot 2^{-jd\theta}.$$

Observe additionally that, for all $C \in \mathcal{P}_j$, the quantity $G(\eta(X)\mathbb{I}\{X \in C\})$ is bounded by $\mu(C)^2$ and that $\mu(C)$ (and *a fortiori* $(1-p)\alpha(C)$ and $p\beta(C)$) is bounded by $||d\mu/dx||_\infty 2^{-jd}$. Combined to (i), this yields the desired result. □

Proof of Theorem 10

Recall that

$$\mathrm{AUC}^* - \mathrm{AUC}(\widehat{\eta}_{\mathcal{P}_j}) = \frac{1}{p(1-p)}\mathbb{E}[|\widehat{\eta}_{\mathcal{P}_j}(X) - \eta(X)|] + \frac{\sum_{C\in\mathcal{P}_j} G(\eta(X)\mathbb{I}\{X \in C\})}{4p(1-p)} + \frac{1}{2}\sum_{C\in\mathcal{P}_j} \alpha(C)\beta(C) \ .$$

By assumption, for all $C \in \mathcal{P}_j$, $\alpha(C)\beta(C)$ is bounded by $||d\mu/dx||_\infty^2 2^{-2jd}/(p(1-p))$. In addition, following line by line the argument of Theorem 6.2 in Devroye et al. (1996), we obtain that $\widehat{\eta}_{\mathcal{P}_{j(n)}}$ converges to η in the $L_1(\mu)$ sense, which proves (i).

We now turn to the second assertion. With a slight abuse of notation, we set for all measurable $C \subset \mathcal{X}$:

$$\eta(C) = \frac{p\beta(C)}{\mu(C)} \ \ \textit{and} \ \ \widehat{\eta}(C) = \frac{n_+\widehat{\beta}(C)}{n\widehat{\mu}(C)},$$

with $\widehat{\mu}(C) = n^{-1}\sum_{i\leq n} \mathbb{I}\{X_i \in C\}$ and the convention: $0/0 = 0$. It follows from Corollary 9 that:

$$\mathrm{AUC}^* - \mathrm{AUC}(\widehat{\eta}_{\mathcal{P}}) \leq c \cdot \#\mathcal{P}_j^{-\theta} + \frac{||\widehat{\eta}_{\mathcal{P}_j} - \eta_{\mathcal{P}_j}||_{L_1(\mu)}}{p(1-p)}.$$

Therefore, we have

$$\begin{aligned} ||\widehat{\eta}_{\mathcal{P}_j} - \eta_{\mathcal{P}_j}||_{L_1(\mu)} &= \mathbb{E}[\sum_{C\in\mathcal{P}_j} |\eta(C) - \widehat{\eta}(C)|\mathbb{I}\{X \in C\}] \\ &\leq \max_{C\in\mathcal{P}_j}\left(\sqrt{\mu(C)} \cdot |\eta(C) - \widehat{\eta}(C)|\right) \\ &\quad \times \left(\textstyle\sum_{C\in\mathcal{P}_j} \sqrt{\mu(C)}\right) \ . \end{aligned}$$

We have $\sum_{C\in\mathcal{P}_j} \sqrt{\mu(C)} \leq 2^{jd/2}$ by Cauchy-Schwarz inequality and for all $C \in \mathcal{P}_j$:

$$|\eta(C) - \widehat{\eta}(C)| \leq \tfrac{|\widehat{\mu}(C) - \mu(C)|}{\mu(C)} + \tfrac{|p\beta(C) - n_+\widehat{\beta}(C)/n|}{p\beta(C)}.$$

By virtue of the relative deviation results stated in Theorem 5.1 in Boucheron et al. (2005), for all $\delta \in]0,1[$, we have with probability at least $1-\delta$ that $\forall n \geq 1$, $\forall j \geq 1$, the quantities $\max_{C\in\mathcal{P}_j} |\widehat{\mu}(C) - \mu(C)|/\sqrt{\mu(C)}$ and $\max_{C\in\mathcal{P}_j} |n_+\widehat{\beta}(C)/n - p\beta(C)|/\sqrt{p\beta(C)}$ are bounded by $\sqrt{(2d\log(2n+1) + \log(8/\delta))/n}$. Combined with the previous bounds and the fact that $2^{j(n)d} \sim n^{1/(1+2\theta)}$, this yields the claimed rate bound. □

Proof of Proposition 11

We first establish the following result.

Lemma 14 *Assume that the hypotheses of the proposition are fulfilled. Then, $\forall j \ge 1$, we have, for n sufficiently large:*

$$\mathbb{E}[\sup_{s\in\mathcal{S}_{\mathcal{P}_j}} |\widehat{\mathrm{AUC}}(s)-\mathrm{AUC}(s)|] \le \frac{1}{p(1-p)}\sqrt{\frac{\log(2^{jd+1}!)}{2n}} \ .$$

PROOF. We first express the $\widehat{\mathrm{AUC}}(s)$ as:

$$\widehat{\mathrm{AUC}}(s) = \frac{n(n-1)}{2n_+n_-}\widehat{U}_n(s) \ ,$$

where $\widehat{U}_n(s) = \frac{2}{n(n-1)}\sum_{1\le i<j\le n} h_s((X_i,Y_i),(X_j,Y_j))$ is a U-statistic of order 2 with bounded symmetric kernel $h_s((x_1,y_1),(x_2,y_2)) = \mathbb{I}\{(y_1-y_2)(s(x_1)-s(x_2)) > 0\}$ and expectation $U(s) = 2p(1-p)\mathrm{AUC}(s)$. Then the bound may be proved for the process $(\widehat{U}-U)(s)$ by applying Hoeffding's inequality for U-statistics combined with the union bound. There is an implicit multiplicative factor of 2 in the bound to remove the part coming from the estimation of the proportion for n large enough. Owing to space limitations, details are left to the reader. □

By virtue of the definition of $\hat{j}$, we have: $\forall j \ge 1$,

$$\begin{aligned}\mathrm{AUC}^* - \mathrm{AUC}(\widehat{s}) &\le \mathrm{AUC}^* - \mathrm{AUC}(s^*_{\mathcal{P}_j}) \\ &+\mathrm{AUC}(s^*_{\mathcal{P}_j}) - \widehat{\mathrm{AUC}}(s^*_{\mathcal{P}_j}) - \mathrm{AUC}(\widehat{s}) + \widehat{\mathrm{AUC}}(\widehat{s}) \\ &+\mathrm{pen}(j,n) - \mathrm{pen}(\widehat{j},n) \ .\end{aligned}$$

Taking expectations and using Lemma 14 combined with the previous bound, we deduce that: $\forall j \ge 1$,

$$\begin{aligned}\mathbb{E}[\mathrm{AUC}^* - \mathrm{AUC}(\widehat{s}^*_{\mathcal{P}_{\hat{j}}})] &\le \mathrm{AUC}^* - \mathrm{AUC}(s^*_{\mathcal{P}_j}) \\ &+\mathbb{E}[\mathrm{AUC}(s^*_{\mathcal{P}_j}) - \widehat{\mathrm{AUC}}(s^*_{\mathcal{P}_j})] + \mathrm{pen}(j,n) - \mathrm{pen}(\widehat{j},n) \\ &+\frac{1}{p(1-p)}\sqrt{\frac{\log(2^{\widehat{j}d+1}!)}{2n}} \ .\end{aligned}$$

As we choose the penalty so that: $\forall j' \ge 1$,

$$\mathrm{pen}(j',n) \ge \frac{1}{p(1-p)}\sqrt{\frac{\log(2^{j'd+1}!)}{2n}} \ ,$$

we obtain:

$$\begin{aligned}\mathbb{E}[\mathrm{AUC}^* - \mathrm{AUC}(\widehat{s}^*_{\mathcal{P}_{\hat{j}}})] &\le \mathrm{AUC}^* - \mathrm{AUC}(s^*_{\mathcal{P}_j}) \\ &+\mathbb{E}[\mathrm{AUC}(s^*_{\mathcal{P}_j}) - \widehat{\mathrm{AUC}}(s^*_{\mathcal{P}_j})] + \mathrm{pen}(j,n) \ .\end{aligned}$$

Moreover, for n large enough and for any $j \ge 1$:

$$\mathbb{E}[\mathrm{AUC}(s^*_{\mathcal{P}_j}) - \widehat{\mathrm{AUC}}(s^*_{\mathcal{P}_j})] \le \mathrm{pen}(j,n) \ ,$$

since the expected deviation of the empirical AUC from its expectation for a single scoring function is of the order of $n^{-1/2}$. This concludes the proof. □

References

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

Binev, P., A. Cohen, W. Dahmen, R. DeVore, and V. Temlyakov (2005). Universal algorithms for learning theory part i: piecewise constant functions. *Journal of Machine Learning Research 6*, 1297–1321.

Boucheron, S., O. Bousquet, and G. Lugosi (2005). Theory of Classification: A Survey of Some Recent Advances. *ESAIM: Probability and Statistics 9*, 323–375.

Clémençon, S., G. Lugosi, and N. Vayatis (2008). Ranking and empirical risk minimization of U-statistics. *The Annals of Statistics 36*(2), 844–874.

Clémençon, S. and N. Vayatis (2007). Ranking the best instances. *Journal of Machine Learning Research 8*, 2671–2699.

Clémençon, S. and N. Vayatis (2008a). Overlaying classifiers: a practical approach for optimal ranking. In *NIPS '08: Proceedings of the 2008 conference on Advances in neural information processing systems.*

Clémençon, S. and N. Vayatis (2008b). Tree-structured ranking rules and approximation of the optimal ROC curve. In *ALT '08: Proceedings of the 2008 conference on Algorithmic Learning Theory.*

Devore, R. (1998). Nonlinear approximation. *Acta Numerica*, 51–150.

Devroye, L., L. Györfi, and G. Lugosi (1996). *A Probabilistic Theory of Pattern Recognition.* Springer.

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

Huang, J. and C. X. Ling (2005). Using AUC and accuracy in evaluating learning algorithms. *IEEE Trans. on Knowl. and Data Eng. 17*(3).

Massart, P. (2006). *Concentration inequalities and model selection.* Lecture Notes in Mathematics. Springer.