Foundations of the Theory of Performance-Based Ranking

Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck
Montefiore Institute, University of Liège, Liège, Belgium
{S.Pierard,Anais.Halin,Anthony.Cioppa,Adrien.Deliege,M.VanDroogenbroeck}@uliege.be

Abstract

Ranking entities such as algorithms, devices, methods, or models based on their performances, while accounting for application-specific preferences, is a challenge. To address this challenge, we establish the foundations of a universal theory for performance-based ranking. First, we introduce a rigorous framework built on top of both the probability and order theories. Our new framework encompasses the elements necessary to (1) manipulate performances as mathematical objects, (2) express which performances are worse than or equivalent to others, (3) model tasks through a variable called satisfaction, (4) consider properties of the evaluation, (5) define scores, and (6) specify application-specific preferences through a variable called importance. On top of this framework, we propose the first axiomatic definition of performance orderings and performance-based rankings. Then, we introduce a universal parametric family of scores, called ranking scores*, that can be used to establish rankings satisfying our axioms, while considering application-specific preferences. Finally, we show, in the case of two-class classification, that the family of ranking scores encompasses well-known performance scores, including the accuracy, the true positive rate (recall, sensitivity), the true negative rate (specificity), the positive predictive value (precision), and F_1. However, we also show that some other scores commonly used to compare classifiers are unsuitable to derive performance orderings satisfying the axioms.*

1. Introduction

Every day, millions of people are faced with choices to make. Often, these choices are between entities (*e.g.*, algorithms, devices, methods, models, options, procedures, solutions, strategies, *etc.*) considered to be interchangeable, although not necessarily equivalent in terms of performance. One of the main difficulties arises from the uncertainty that people have regarding the use that will be made of the entity to choose. A widespread approach to objectifying these choices is to (1) perform an evaluation to determine (*i.e.*, assume, calculate, estimate, predict, *etc.*) a *performance*, encompassing the necessary uncertainty, for each of these entities; (2) choose a way of comparing these performances with each other; and (3) assume that an entity is preferable to others if it has the best performance. A more general problem is to establish an order of preference between the entities: this is the *performance-based ranking*.

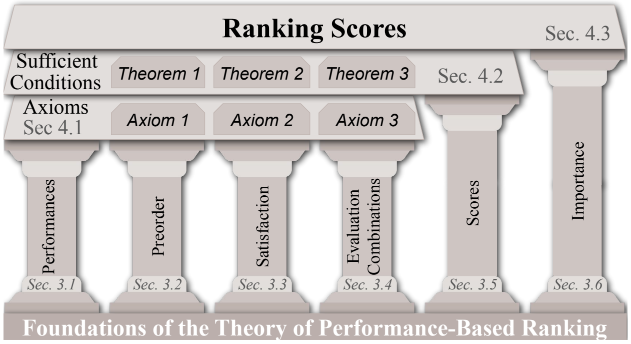

Figure 1. This work establishes the foundations of the theory of performance-based ranking. We do this in two steps. First, we introduce a new mathematical framework with 6 main elements, as depicted here by the pillars. Second, we build on top of it: (1) a set of three axioms for the ordering of performances and for the performance-based ranking of entities, (2) sufficient conditions for them when the performance ordering is induced by a score, and (3) a family of scores, named *ranking scores* that consider the application-specific preferences. This theory is universal in the sense that it is applicable to any task.

The approach of performance-based ranking is common in many fields and has proved its usefulness, especially in scientific communities that organize themselves around competitions [3, 10, 16, 17] for the development of algorithms for specific tasks. Nevertheless, several studies [18, 19] have alerted the scientific community about the ranking methodology used in these competitions.

A critical analysis [18] of common practices for 150 biomedical image analysis challenges reveals that the scores used are justified in only 23% of the cases, and that the rank computation method is reported in only 36% of the cases. Moreover, there are at least 10 different methods for determining the rank of an algorithm based on multiple scores. The properties of these methods are largely unknown.

There remains an obvious lack of theoretical foundations for the performance-based ranking. From our perspective, there is a common and detrimental confusion between the concept of performance and the numerical scores (also called metrics, measures, indicators, criteria, factors, and indices [2, 24]). Moreover, the way of comparing performances is often chosen by intuition [19] or by imitation of what has been chosen in the previously published related works. This inevitably leads to a drastic loss of diversity in the rankings that can be presented. From our perspective, considering the diversity of possible rankings is healthy, and also desirable for those who want to subsequently choose an entity based on their own application-specific preferences.

The lack of theoretical foundations is profound. Loosely speaking, the *performance* of a given entity can be defined as the information necessary to determine the degree of *satisfaction* one has with it, relatively to application-specific preferences tuned by the relative *importance* given to the various cases that can occur when the entity is used. Although such a definition sheds interesting light on the subject, it does not specify the mathematical nature of the performance, the space in which it is, or the operations that are permitted, particularly those that enable the various performances to be compared with one another.

The aim of this paper is to look at the performance-based ranking from a broader angle. As shown in Fig. 1, we establish mathematically rigorous theoretical foundations that can be applied to a wide range of problems. Throughout the paper, we exemplify our theory with the task of two-class classification. More tasks are detailed in Appendix A.2 about: multi-class classification with links to micro- and macro-averaging; regression with links to the mean squared error and the mean absolute error; information retrieval; detection with links to the intersection over-union and the F-score; clustering with a link to the Fowlkes-Mallows index; ranking with a link to Kendall's τ.

Contributions. Our contributions are threefold. (1) First, in Sec. 3, we present a new mathematical framework that encompasses the evaluation of entities, the performances, the space in which they are, the notions of *satisfaction* and *importance*, the scores that characterize numerically the performances, and some operations permitted on performances such as those for combining several performances and those for comparing performances with one another (*is equivalent to*, *is worse than*, *is better than*, *is incomparable with*, *etc.*). (2) Second, we introduce the foundations of the theory of performance-based ranking in Sec. 4. We innovate with an *axiomatic definition for both the ordering of performances and the ranking of entities based on their performances*. We also present a new family of scores, named *ranking scores*, that can be used to induce performance orderings that satisfy our axioms. These scores are parameterized by application-specific preferences, and are universal, *i.e.*, applicable to any task. (3) Third, in Sec. 5, we study the particularization of our theory of performance-based ranking to the popular case of the two-class crisp classification.

2. Related Work

Mathematical foundations for our work can be found in the probability theory, in the order theory, and in statistics. The probability theory [14, 15] provides the tools needed to consider the uncertainty that one has about how an entity will be used. However, to be rigorous, probability measures cannot be used without defining a measurable space. Surprisingly, after reviewing hundreds of papers that use the notion of performance, mainly in the fields of computer vision, medicine and physics, we found none that explicitly gives such a space and explains how to express, based on it, what the performances are and what operations are allowed on them. The order theory [6, 11] provides the basis for defining and manipulating homogeneous binary relations such as *is equivalent to*, *is worse than*, *is better than*, and *is incomparable with* that underpin rankings. Statistics provide tools to compare rankings through rank correlations, in particular Kendall's τ [13] and Spearman's ρ [23].

In a recent attempt to formalize the notion of ranking, for basic vision tasks, Nguyen *et al.* [19] proposed to impose three properties for ranking: (1) reliability ("a small change in parameter values, should not result in a drastic change in rankings"), (2) meaningfulness (evaluated by humans), and (3) mathematical consistency ("use scores that satisfy certain properties"). The mathematical framework introduced in this paper helps to clarify these three requirements. (1) Regarding the reliability, which is also a matter of concern in [18], we argue that one should distinguish between two types of parameters: those involved in the evaluation (*i.e.* the step in which the performance of an entity is determined) and those involved in the ranking of the entities based on their previously determined performances. The case actually discussed in [19] is of the first type. The second type of parameters can be useful to adapt the ranking to application-specific preferences: the *importance*. (2) Regarding the meaningfulness, in the absence of analytical means to ensure the meaningfulness of scores, Nguyen *et al.* [19] suggest testing the ranking procedure on sanity tests with pre-determined desired rankings. In contrast, we propose to start by modeling the task through a variable called *satisfaction* and then to derive meaningful ranking scores, by construction. (3) Concerning the mathematical consistency, we show that it is possible to impose it based on axioms, at a level just below the scores (see Fig. 1).

3. Mathematical Framework

We now present the mathematical framework in which the theory of performance-based rankings will be established in Sec. 4.

To the best of our knowledge, it is the first time that a rigorous mathematical framework is conceived for the universal comparison of performances. All mathematical symbols used in this paper are defined where they first appear. For convenience, we also provide a list of them in Appendix A.1. Our framework involves six components that correspond to the pillars depicted in Fig. 1.

Performance to address uncertainty. First, we introduce the notion of *performance*, which is our main object of interest. To anchor it on a solid mathematical ground, we leverage probability theory to benefit from its established expressivity and rigor, and thus define performances as probability measures. Probability theory is indeed the ideal framework for studying uncertainty and randomness, which naturally pertain to performances, hence our choice.

The essence of performance. Then, we define the three components that really differentiate performances from ordinary probability measures: performances should be comparable/rankable (*e.g.*, we need notions of *better* and *worse* performances) through a *preorder* $\lesssim$; performances should be related to a task, that we model by a random variable called *satisfaction* S; and performances should be related to entities (algorithms, devices, *etc.*) through an *evaluation*, that we model by a function Φ. We show that there exist compatibility conditions to be met between $\lesssim$ and S as well as between $\lesssim$ and Φ, which lead us to formalize the *axioms* at the basis of our theory in Sec. 4.

Performance in practical applications. Finally, we define two components that connect our theory to practical applications: the *scores* X, which are functions mapping performances to numerical values (*e.g.* accuracy, error rate,*etc.*); and the *importance* I, a random variable that encodes applicative preferences about the possible outcomes of the process. Later in Sec. 4, we show how these two components can be defined to satisfy sufficient conditions (through three theorems) to fulfill our three axioms, thus further serving as basis for our performance-based ranking theory and the definition of new universal ranking scores.

3.1. The Performance P as a Probability Measure

For us, a performance is not a real number or a collection of them, as sometimes assumed in the literature. Instead, we choose to design our mathematical framework specifically to compare performances having a probabilistic meaning. Thus, we ground our framework in probability theory [14, 15] and consider that *performances* P are probability measures. For two performances to be comparable, they should be defined on a common measurable space (Ω, Σ), where Ω is the sample space, or universe, and Σ is a σ-algebra on Ω called event space. Without loss of generality, when Ω is finite, one can choose $\Sigma = 2^{\Omega}$. We note $\mathbb{P}_{(\Omega,\Sigma)}$ the set of probability measures on (Ω, Σ).

Example 1. *For the popular case of two-class crisp classification, several choices can be made for Ω. In the simplest setting, this set contains two elements interpreted as "correct result" and "incorrect result". In this case, the performance analysis can only be based on the proportion of correct results,* i.e.*, the accuracy. In another setting, one can choose to have three elements in Ω, making the distinction between the two types of incorrect results, namely* false positive *(*a.k.a. *type I error) and* false negative *(*a.k.a. *type II error). Finally, one could prefer having four elements in Ω, one for each pair of ground-truth and predicted classes (in the frequency-based approach, this corresponds to the normalized contingency table or confusion matrix). This enlarges the flexibility in the analysis of performances.*

3.2. Ordering Performances with a Preorder $\lesssim$

Our framework is not only grounded in probability theory, as explained above, but also in order theory [6, 11]. We aim at being able to decide if a performance is, *e.g.*, worse than, equivalent, or better than another one. Mathematically, these comparisons are done thanks to binary homogeneous relations on $\mathbb{P}_{(\Omega,\Sigma)}$. Indeed, all binary relations used to compare performances should be coherent, *i.e.*, they should all correspond to a common *performance ordering*. Following this path, in this paper, the binary relations $\sim$, $>$, $<$, and $\not\gtreqless$ on $\mathbb{P}_{(\Omega,\Sigma)}$ will be implicitly considered as derived from a common binary relation $\lesssim$ on $\mathbb{P}_{(\Omega,\Sigma)}$ as follows:

- $P_1 \sim P_2$ if and only if (*iif*) $P_1 \lesssim P_2 \wedge P_2 \lesssim P_1$;
- $P_1 > P_2$ *iif* $P_1 \not\lesssim P_2 \wedge P_2 \lesssim P_1$;
- $P_1 < P_2$ *iif* $P_1 \lesssim P_2 \wedge P_2 \not\lesssim P_1$;
- $P_1 \not\gtreqless P_2$ *iif* $P_1 \not\lesssim P_2 \wedge P_2 \not\lesssim P_1$.

With such a construction, if $\lesssim$ is a preorder (*i.e.*, reflexive and transitive), then $\sim$ is an equivalence (*i.e.* reflexive, transitive, and symmetric), $>$ and $<$ are converse strict partial orders (*i.e.*, irreflexive, asymmetric, and transitive), and $\not\gtreqless$ is irreflexive and symmetric. For the proof, see Appendix A.3.4. These are the intuitively expected properties that justify to interpret $\lesssim$ as *worse or equivalent*, $>$ as *better*, $<$ as *worse*, and $\not\gtreqless$ as *incomparable*.

Example 1 (continued)**.** *In the case of two-class classification, two spaces are commonly used to depict the performances as points, at fixed priors: the* Receiver Operating Characteristic *(ROC) space and the* Precision-Recall *(PR) space. On the one hand, ROC users will all intuitively agree that a performance P_1 is better than, equivalent to, or worse than a performance P_2 when both the values of TNR (true negative rate) and the TPR (true positive rate) of P_1 are greater, equal, or smaller than those of P_2, respectively.*

Few of these users will risk deciding when one of the two scores is higher while the other is lower, as this depends on application-specific preferences. On the other hand, PR users will perform a similar intuitive reasoning based on the TPR (recall) and the PPV (precision). A careful comparison reveals that ROC and PR users do not intuitively make the same decisions. This is surprising since, at fixed priors, both spaces show the same thing [7]. Our theory clarifies what are the suitable performance orderings.

3.3. Modeling the Task as a Random Variable S

The random variable *satisfaction*, $S : \Omega \to \mathbb{R}$, is task-specific. The user is responsible for assigning satisfaction values that are meaningful for the task at hand. We argue that a task is ill-defined when the satisfaction is not specified, either explicitly or implicitly. In this paper, we pose $s_{min,\Omega} = \min_{\omega \in \Omega} S(\omega)$ and $s_{max,\Omega} = \max_{\omega \in \Omega} S(\omega)$.

Example 1 (continued)**.** *In the case of two-class classification, we expect that most people will agree that (1) the satisfaction takes the same value for all samples corresponding to incorrect results (*e.g. $S = 0$*), (2) the satisfaction takes the same value for all samples corresponding to correct results (*e.g. $S = 1$*), and that (3) the satisfaction is strictly greater for correct results than for incorrect results.*

3.4. Modeling the Evaluation as a Function Φ

We now have all the elements necessary to introduce the notion of *evaluation*. Let us denote, by $\mathbb{E}$, the set of entities of interest. In our framework, the evaluation is modeled by function $\text{eval} : \mathbb{E} \to \mathbb{P}_{(\Omega,\Sigma)} : \epsilon \mapsto \text{eval}(\epsilon)$. We find it convenient to think in terms of a (thought) random experiment involving an entity ϵ and having outcomes that allow to determine how satisfying the various realizations are, *e.g.*, if the result is correct, how accurate it is, or how many resources are used. We define the performance $P = \text{eval}(\epsilon)$ of an entity ϵ as the distribution of these outcomes.

Example 1 (continued)**.** *In the particular case of classification, a typical random experiment implicitly considered in the literature is in five steps. (1) Draw a sample s at random from a source. (2) Apply the oracle on s to obtain ground-truth class $y(s)$. (3) Apply a descriptor on s to obtain the features (*a.k.a. *attributes) $x(s)$. (4) Feed the classifier with $x(s)$ to obtain the predicted class $\hat{y}(x(s))$. (5) Set the outcome of the experiment to the pair $(y, \hat{y})$.*

Our framework can be further enriched by integrating some knowledge about the function eval. By definition, a performance P is achievable when there exists an entity ϵ whose evaluation leads to it: $\exists \epsilon : \text{eval}(\epsilon) = P$. It often happens that, based solely on the performances of the evaluated entities, we can be sure that some other performances are also achievable, by combining and/or disrupting the entities we have at our disposal. To take this knowledge into account, we define the function $\Phi : 2^{\mathbb{P}_{(\Omega,\Sigma)}} \to 2^{\mathbb{P}_{(\Omega,\Sigma)}}$ that gives the set of performances that are achievable for sure, for some set of achievable performances given in input. Note that Φ is idempotent, *i.e.* $\Phi \circ \Phi = \Phi$.

Consider a random experiment using a black box entity ϵ only once (this is the knowledge we have about eval). Let $\lambda_1, \lambda_2, \dots \lambda_n$ be positive values summing up to one. All other things remaining identical, if one can achieve the performances $P_1, P_2, \dots P_n$ with, respectively, the entities $\epsilon_1, \epsilon_2, \dots \epsilon_n$, then the performance $\sum_i \lambda_i P_i$ is achievable with a hybrid entity that randomly selects an entity among $\epsilon_1, \epsilon_2, \dots \epsilon_n$ with the series of respective selection probabilities $\lambda_1, \lambda_2, \dots \lambda_n$ before running the corresponding entity. In this case, denoting the set of all possible convex combinations by conv, we can take $\Phi = \text{conv}$.

Example 1 (continued)**.** *In the particular case of two-class crisp classification, when the source of samples, the oracle, and the descriptor are kept unchanged,* Fawcett's interpolation *[8] allows interpolating linearly between performances with a hybrid classifier. And, when the oracle, the descriptor, and the classifier are kept unchanged,* Piérard's summarization *[20] allows interpolating linearly between performances with a hybrid source of samples.*

3.5. The Scores as Functions X of Performances

In our framework, *scores*[1] (also called metrics, measures, indicators, criteria, factors, and indices in the literature [2, 24]) are functions associating a real value to performances, that is, $X : \text{dom}(X) \to \mathbb{R} : P \mapsto X(P)$ with $\text{dom}(X) \subseteq \mathbb{P}_{(\Omega,\Sigma)}$. One can define different parametric families of scores, such as:

- *Expected value scores* are parameterized by a random variable V. We define them as $X^E_V : \mathbb{P}_{(\Omega,\Sigma)} \to [\min_{\omega \in \Omega} V(\omega), \max_{\omega \in \Omega} V(\omega)] : P \mapsto X^E_V(P) = \mathbf{E}_P[V]$, where $\mathbf{E}$ denotes the mathematical expectation. The score X^E_S, that we call the *expected satisfaction*, is a universal score in the sense that it exists for all sample spaces Ω. It will be further studied in Sec. 4.
- *Probabilistic scores* are parameterized by two events $E_1, E_2 \in \Sigma$ such that $\emptyset \subsetneq E_1 \subsetneq E_2 \subseteq \Omega$. We define them as $X^P_{E_1|E_2} : \{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(E_2) \neq 0\} \to [0,1] : P \mapsto X^P_{E_1|E_2}(P) = P(E_1|E_2)$. All probabilistic scores can be expressed as a ratio of two expected value scores.

Example 1 (continued)**.** *In the case of two-class classification, the expected satisfaction is both the expected value score X^E_S and the probabilistic score $X^P_{S=1|\Omega}$. This is because S is a $\{0,1\}$-binary random variable. In this case, the expected satisfaction is called* accuracy *and is noted A.*

[1] We choose the term *score* to avoid any possible confusion with the mathematical meaning of the terms *metric*, *measure*, and *indicator*.

3.6. Modeling Application-Specific Preferences as a Random Variable I

It is well known that the ranking of entities does not only have to be specific for the task, but that it should also be sensitive to application-specific preferences. To encode these preferences, we propose to rely on a second random variable that we call *importance*: $I : \Omega \to \mathbb{R}_{\geq 0}$. We require that $I \neq 0$, *i.e.* $\exists \omega : I(\omega) \neq 0$.

Further in this paper, we will describe a new family of scores, called *ranking scores*, which are parameterized by this random variable. We would, however, like to draw the reader's attention to the fact that the importance is something that cannot, in general, be deduced from a score. In particular, the visual inspection of a formula for a given score could be misleading. Therefore, in this paper, we present a technique to analyze the behavior of any score by computing the rank correlations with the ranking scores for which the importances are well-defined.

Example 1 (continued)**.** *In the case of two-class classification, we provide examples of misleading formulas in Appendix A.4, in particular two equivalent formulas for the accuracy and two for the true positive rate. In both cases, by visually inspecting them, one would intuitively draw different conclusions about the importance given to the true negatives, false positives, false negatives, and true positives.*

4. Performance-Based Ranking Theory

In this section, we build our theory in three steps, represented by the three lintels of Fig. 1, on top of the six pillars depicting our mathematical framework. First, we present a universal axiomatic definition of performance orderings and performance-based rankings of entities (Sec. 4.1). Then, we link these axioms with the scores, and give a sufficient condition per axiom (Sec. 4.2). Finally, we account for the application-specific preferences and provide an infinite, diversified, and universal family of scores that induce performance orderings satisfying our axioms (Sec. 4.3).

4.1. Axiomatic Definition

We propose an axiomatic definition (the first as far as we know) of both performance orderings and performance-based rankings. The axioms do not involve any X or I.

4.1.1 Leveraging the Preorder $\lesssim$

We argue that if several entities from a set $\mathbb{E}$ have been ranked, then removing or adding an entity should not affect the relative order of those ranked entities. To guarantee it, our first axiom imposes that the ranking function is based on a preorder $\lesssim$ on $\mathbb{P}_{(\Omega,\Sigma)}$. There is no consensus in the literature on the rank value to consider when several entities have equivalent performances. Thus, instead of setting an arbitrarily chosen value, our axiom specifies bounds for it.

Axiom 1. *The ranking function* $\mathrm{rank}_{\mathbb{E}} : \mathbb{E} \to [1, |\mathbb{E}|] : \epsilon \mapsto \mathrm{rank}_{\mathbb{E}}(\epsilon)$ *satisfies* $|\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) < \mathrm{eval}(\epsilon')\}| + 1 \leq \mathrm{rank}_{\mathbb{E}}(\epsilon) \leq |\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) \lesssim \mathrm{eval}(\epsilon')\}|$*, where* $\lesssim$ *is a preorder on* $\mathbb{P}_{(\Omega,\Sigma)}$.

4.1.2 Leveraging the Satisfaction S

We argue that if the satisfaction that can be obtained with an entity ϵ_1 is for sure less or equal than the satisfaction that can be obtained with an entity ϵ_2, then the performance P_1 of ϵ_1 cannot be better than the performance P_2 of ϵ_2.

Axiom 2. *For* $P_1, P_2 \in \mathbb{P}_{(\Omega,\Sigma)}$ *such that* $P_1(S \leq s) = 1$ *and* $P_2(S \geq s) = 1$ *for some* s*, then* $P_1 \lesssim P_2$ *or* $P_1 \not\lesseqgtr P_2$.

This implies that, thanks to S, we can use the natural order $\leq$ that exists on $\mathbb{R}$ to obtain a preorder on Ω, and the axiom states that the preorder on $\mathbb{P}_{(\Omega,\Sigma)}$ is coherent with it.

Let us now take a look at three implications that are clearly intuitively expected.

Corollary 1. *For any* s*, all performances* P *such that* $P(S = s) = 1$ *are either equivalent or incomparable.*

Corollary 2. *We have* $P(S = s_{min,\Omega}) = 1 \Rightarrow \nexists P' : P' < P$*. In other words,* $P(S = s_{min,\Omega}) = 1$ *means that* P *belongs to the set of the worst performances, among* $\mathbb{P}_{(\Omega,\Sigma)}$.

Corollary 3. *We have* $P(S = s_{max,\Omega}) = 1 \Rightarrow \nexists P' : P' > P$*. In other words,* $P(S = s_{max,\Omega}) = 1$ *means that* P *belongs to the set of the best performances, among* $\mathbb{P}_{(\Omega,\Sigma)}$.

4.1.3 Leveraging the Combinations Φ

We argue that, for any set of achievable performances, it must be impossible to obtain with certainty a performance better than the best of them, or worse than the worst of them, by sequentially combining operations from an arbitrary set of possible operations to perturb (*e.g.*, add noise to their output) the entities corresponding to that initial set. Expressing this requirement in terms of $\lesssim$ leads to our third axiom.

Axiom 3. *Let* P *be a performance, and* Π *be a set of performances on* $\mathbb{P}_{(\Omega,\Sigma)}$ *such that* $P' \lesssim P \vee P \lesssim P' \, \forall P' \in \Pi$.

- $P' \lesssim P \, \forall P' \in \Pi \Rightarrow \overline{P} \lesssim P \, \forall \overline{P} \in \Phi(\Pi)$*;*
- $P' \not\lesssim P \, \forall P' \in \Pi \Rightarrow \overline{P} \not\lesssim P \, \forall \overline{P} \in \Phi(\Pi)$*;*
- $P \lesssim P' \, \forall P' \in \Pi \Rightarrow P \lesssim \overline{P} \, \forall \overline{P} \in \Phi(\Pi)$*;*
- *and* $P \not\lesssim P' \, \forall P' \in \Pi \Rightarrow P \not\lesssim \overline{P} \, \forall \overline{P} \in \Phi(\Pi)$.

With the definitions for $\sim$, $>$, $<$, and $\not\lesseqgtr$ given in Sec. 3.2, the following corollary can be derived.

Corollary 4. *Let* $P \in \mathbb{P}_{(\Omega,\Sigma)}$ *and* $\Pi \subseteq \mathbb{P}_{(\Omega,\Sigma)}$ *such that* $P' \lesssim P \vee P \lesssim P' \, \forall P' \in \Pi$*. We have:*

- $P' \sim P \forall P' \in \Pi \Rightarrow \overline{P} \sim P \forall \overline{P} \in \Phi(\Pi)$;
- $P' > P \forall P' \in \Pi \Rightarrow \overline{P} > P \forall \overline{P} \in \Phi(\Pi)$;
- $P' < P \forall P' \in \Pi \Rightarrow \overline{P} < P \forall \overline{P} \in \Phi(\Pi)$;
- $P' \not\lesseqqgtr P \forall P' \in \Pi \Rightarrow \overline{P} \not\lesseqqgtr P \forall \overline{P} \in \Phi(\Pi)$.

4.1.4 Consistency

The three axioms are consistent in the sense that they do not contradict each others. A trivial preorder $\lesssim$ for which all axioms are satisfied is the one such that all performances are equivalent, regardless of what S and Φ are.

4.2. Sufficient Conditions for Score-Based Rankings

We can now connect the axioms and scores and give sufficient conditions to satisfy our axioms. The proofs for the three following theorems are given in Appendix A.5.

The 1st theorem explains how performance orderings $\lesssim$ can be induced from scores X, which allows capitalizing on the natural order $\leq$ on $\mathbb{R}$ to obtain a preorder $\lesssim$ on $\mathbb{P}_{(\Omega,\Sigma)}$.

Theorem 1 (Sufficient condition for Axiom 1)**.** *A binary relation $\lesssim_X$ on $\mathbb{P}_{(\Omega,\Sigma)}$ induced by a score X as $P_1 \lesssim_X P_2$* iif *either $P_1 = P_2$ or $P_1 \in \mathrm{dom}(X)$ and $P_2 \in \mathrm{dom}(X)$ and $X(P_1) \leq X(P_2)$, is a preorder satisfying Axiom 1.*

The 2nd theorem makes the connection between the properties of the scores X and the satisfaction S.

Theorem 2 (Sufficient condition for Axiom 2)**.** *If a score X satisfies $\min_{\omega \in E} S(\omega) \leq X(P) \leq \max_{\omega \in E} S(\omega)$ for all events $E \in \Sigma$ and all performances $P \in \mathrm{dom}(X)$ such that $P(E) = 1$, then the ordering $\lesssim_X$ satisfies Axiom 2.*

The 3rd theorem makes the connection between the properties of the scores X and the function Φ.

Theorem 3 (Sufficient condition for Axiom 3)**.** *If a score X is such that $\Pi \subseteq \mathrm{dom}(X) \Rightarrow \Phi(\Pi) \subseteq \mathrm{dom}(X)$ and $\min_{P \in \Pi} X(P) \leq X(\overline{P}) \leq \max_{P \in \Pi} X(P)$ for all $\Pi \subseteq \mathrm{dom}(X)$ and all $\overline{P} \in \Phi(\Pi)$, then the ordering $\lesssim_X$ satisfies Axiom 3.*

4.3. The Ranking Scores: Solutions for $\Phi = \mathrm{conv}$

We now aim to define scores that satisfy our theorems and thus our axioms. We provide such scores in the case of $\Phi = \mathrm{conv}$, shown particularly relevant in our catalog of practical examples (see Appendix A.2). Beyond that, we include the application-specific preferences modeled by the random variable I. We introduce a new family of scores, the *ranking scores* R_I, that are parameterized by the *importance*, a non-negative random variable $I \neq 0$:

$$R_I : \mathrm{dom}(R_I) \to [s_{min,\Omega}, s_{max,\Omega}] :$$
$$P \mapsto R_I(P) = \frac{\mathbf{E}_P[IS]}{\mathbf{E}_P[I]} = \frac{\sum_{\omega \in \Omega} I(\omega) S(\omega) P(\{\omega\})}{\sum_{\omega \in \Omega} I(\omega) P(\{\omega\})}$$

where $\mathrm{dom}(R_I) = \{P \in \mathbb{P}_{(\Omega,\Sigma)} : \mathbf{E}_P[I] \neq 0\}$ and the second equality holds in the common case where Ω is finite and $\Sigma = 2^{\Omega}$. The expected satisfaction X_S^E corresponds to R_I when all samples are equally important.

The scores R_I and the performance orderings $\lesssim_{R_I}$ satisfy the conditions of Theorem 1, 2, and 3 for $\Phi = \mathrm{conv}$ (proofs in Appendix A.6.1). Thus, the performance orderings $\lesssim_{R_I}$ induced by the ranking scores satisfy all our axioms. Hence, the chosen name. Remarkably, these scores are universal: they can be used for performance-based ranking, regardless of the sample space Ω.

Properties. Let us now give some selected properties (proofs are in Appendix A.6.1). The first one clarifies how the importance can be interpreted. Consider any random experiment for which the set of possible outcomes is Ω, their distribution being given by P. One can choose to *filter* the outcomes as follows: run the experiment, and if the outcome is ω, end the experiment with probability $I(\omega) / \sum_{\omega' \in \Omega} I(\omega')$, otherwise restart everything. The new distribution of outcomes is given by the operation $\mathrm{filter}_I(P)$.

Property 1. *The ranking scores can be decomposed into an operation on performances considering the importance I and a score (the expected satisfaction) considering the satisfaction S. We have $R_I = X_S^E \circ \mathrm{filter}_I$, with*

$$\mathrm{filter}_I : \mathrm{dom}(R_I) \to \mathbb{P}_{(\Omega,\Sigma)} :$$
$$P \mapsto \left(\Sigma \to [0,1] : E \mapsto \frac{\sum_{\omega \in E} P(\{\omega\}) I(\omega)}{\sum_{\omega \in \Omega} P(\{\omega\}) I(\omega)} \right).$$

So far, the satisfaction was fixed. But what if one hesitates about the objective of the task, *i.e.* with its modeling through S? The next property clarifies what really matters.

Property 2. *Linearly transforming the satisfaction results in the same linear transformation of the ranking score. It is something that does not affect the ordering.*

The next property is about the scale invariance of I.

Property 3. $R_{kI} = R_I, \forall k \neq 0$. *We can thus restrict the study of ranking scores to the case in which the total importance $\sum_{\omega \in \Sigma} I(\omega)$ is constant.*

We can go further in the case of a binary satisfaction.

Property 4. *For a binary satisfaction, the performance ordering induced by a ranking score is insensitive to the uniform scaling of the importance given to the unsatisfying ($S^{-1}(0)$) or to the satisfying ($S^{-1}(1)$) samples.*

Most scores obtained by averaging or integrating ranking scores are not themselves ranking scores, and are unsuited for ranking. Thanks to Theorem 2, we know that the performance orderings induced by them satisfy Axiom 2. However, most often, they do not satisfy Axiom 3. Yet, two exceptions occur in the case of a binary satisfaction.

Property 5. *If I is the arithmetic mean of I_1 and I_2, then R_I is the f-mean of R_{I_1} and R_{I_2} with $f : x \mapsto x^{-1}$,* i.e.*, the harmonic mean, when $S(\omega) = 1 \Rightarrow I_1(\omega) = I_2(\omega)$.*

Property 6. *If I is the arithmetic mean of I_1 and I_2, then R_I is the f-mean of R_{I_1} and R_{I_2} with $f : x \mapsto (1-x)^{-1}$, when $S(\omega) = 0 \Rightarrow I_1(\omega) = I_2(\omega)$.*

For any homogeneous binary relation $\mathcal{R}$ (*e.g.* $\leq$, $<$, $=$, $>$, $\geq$, $\ldots$) on $\mathbb{R}$, let us define the set $\phi_{\mathcal{R}}(P) = \{P' \in \mathbb{P}_{(\Omega,\Sigma)} : R_I(P')\mathcal{R}R_I(P)\}$. Based on these sets, the following property helps to understand why the ranking scores are suitable when $\Phi = \mathrm{conv}$. This is indeed related to the fact that the scores R_I are pseudolinear functions.

Property 7. *With the performance orderings $\lesssim_{R_I}$ induced by the ranking scores R_I and for any given performance $P \in \mathbb{P}_{(\Omega,\Sigma)}$, the worse set $\phi_<(P)$ (*a.k.a. *the strictly lower contour set), the worse or equivalent set $\phi_\leq(P)$ (the lower contour set), the equivalent set $\phi_=(P)$, the better or equivalent set $\phi_\geq(P)$ (the upper contour set) and the better set $\phi_>(P)$ (the strict upper contour set) are all convex.*

5. Observations for Two-Class Classification

We now examine the particular case of two-class crisp classification. First, we compare the classical formulation of this task with ours. Then, we examine more closely what our ranking theory teaches us for this task.

5.1. Classical vs. our Formulation

Limitation of the classical formulation. Usual two-class classification settings first define a set $\mathbb{C} = \{c_-, c_+\}$ in which c_- and c_+ are named *negative class* and *positive class*. Then, pairs $(y, \hat{y}) \in \mathbb{C}^2$ composed of a "ground truth" y and a "prediction" $\hat{y}$ are interpreted as: a *true negative* $tn = (c_-, c_-)$, a *false positive* $fp = (c_-, c_+)$ (type I error), a *false negative* $fn = (c_+, c_-)$ (type II error), and a *true positive* $tp = (c_+, c_+)$. Finally, scores are defined as formulas involving these elements, upon which rankings of entities are based. However, there is no mathematical guarantee of the meaningfulness of such rankings.

Advantage of our formulation. In our framework, the two-class classification settings are as follows. We consider the sample space $\Omega = \{tn, fp, fn, tp\}$ and the event space $\Sigma = 2^\Omega$. The samples tn, fp, fn, and tp are interpreted as in the classical case. The natural choice for the satisfaction is such that $S(fp) = S(fn) = 0$ and $S(tn) = S(tp) = 1$. After modeling application-specific preferences through a choice of importances I, we can derive rankings scores R_I to rank entities based on their performances, with the certainty that the ranking is mathematically valid.

Link between the two formulations. Fortunately, the two formulations can be easily connected, which allows translating concepts from one to the other whenever necessary. Indeed, to go from our formulation to the classical one, we only need to define a "ground-truth" random variable $Y : \Omega \to \mathbb{C}$ such that $Y(tn) = Y(fp) = c_-$ and $Y(fn) = Y(tp) = c_+$, as well as the "prediction" random variable $\hat{Y} : \Omega \to \mathbb{C}$ such that $\hat{Y}(tn) = \hat{Y}(fn) = c_-$ and $\hat{Y}(fp) = \hat{Y}(tp) = c_+$. Conversely, moving from the classical formulation to ours just requires considering $\Omega = \mathbb{C}^2$ and $S = \mathbf{1}_{Y=\hat{Y}}$. This link is represented in Fig. A.7.1.

5.2. What Does our Ranking Theory Teach us?

We now examine some common scores used in the literature for two-class classification, from the ranking standpoint. Methodologically, we choose the set of numerical quantities that were listed in a recent review [2], and focus on those that are scores (this excludes the base measures and the 1st level measures as they are called in that paper).

We consider three sets of performances: (1) all performances $\mathbb{P}_{(\Omega,\Sigma)}$; (2) all performances for fixed and unbalanced priors (we set arbitrarily a positive prior of 0.2); (3) all performances for fixed and balanced priors. For each score X and each set of performances Π, we report in Tab. 1 the result of three tests on X as well as the minimum and maximum rank correlations X has with the ranking scores.

The tests are the following. The 1st test determines if $\lesssim_X$ satisfies Axiom 2 when the set $\mathbb{P}_{(\Omega,\Sigma)}$ is restricted to Π. The 2nd and 3rd tests relate to Theorem 3 (and thus to Axiom 3). The 2nd test determines if, for any subset Π' of Π, $\max_{P' \in \Pi'} X(P')$ is greater or equal to $X(P)$ for all performances $P \in \mathrm{conv}(\Pi')$. The 3rd test determines if, for any subset Π' of Π, $\min_{P' \in \Pi'} X(P')$ is less or equal to $X(P)$ for all performances $P \in \mathrm{conv}(\Pi')$.

For the rank correlations, we first try to determine analytically if the score is monotonically increasing or decreasing with one of the ranking scores. If it is the case, we report a maximum correlation of 1 or a minimum correlation of -1, respectively. The proofs can be found in Appendices A.7.3 and A.7.4. Otherwise, we report empirical values obtained by optimizing Kendall's τ. We consider a uniform distribution of performances within Π. Kendall's τ is computed using a function provided in SciPy [25], with its default parameters, and fed with the values of R_I and X for about 6,550 performances regularly placed in Π. Note that Kendall's τ is not a continuous function of I when estimated on a finite set of performances. We designed a custom optimizer to estimate τ, detailed in Appendix A.7.2.

The result of our review is given in Tab. 1, where the scores have been grouped in three categories: in green are the scores satisfying the three tests in all cases. In orange

Table 1. Properties of some common scores defined in the literature for two-class crisp classification. The symbol †indicates a value that has been obtained theoretically, the others have been obtained empirically. Our conclusion is that, for the purpose of ranking, the scores in green can always be used, those in orange should only be used when the priors are fixed, and those in black cannot be used even when the priors are fixed. See Sec. 5 for the detailed description.

	without any constraint: all performances					with constraint: positive prior = 0.2					with constraint: positive prior = 0.5				
score	1st test	2nd test	3rd test	τ_{min}	τ_{max}	1st test	2nd test	3rd test	τ_{min}	τ_{max}	1st test	2nd test	3rd test	τ_{min}	τ_{max}
Accuracy	**V**	**V**	**V**	**0.469**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.157**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.505**	$\mathbf{1}^{\dagger}$
F-score for $\beta = 0.5$	**V**	**V**	**V**	**0.079**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.451**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.352**	$\mathbf{1}^{\dagger}$
F-score for $\beta = 1.0$	**V**	**V**	**V**	**0.161**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.352**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.194**	$\mathbf{1}^{\dagger}$
F-score for $\beta = 2.0$	**V**	**V**	**V**	**0.079**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.194**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.072**	$\mathbf{1}^{\dagger}$
Negative Predictive Value (NPV)	**V**	**V**	**V**	**0.000**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.503**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.503**	$\mathbf{1}^{\dagger}$
Positive Predictive Value (PPV)	**V**	**V**	**V**	**0.000**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.503**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.503**	$\mathbf{1}^{\dagger}$
True Negative Rate (TNR)	**V**	**V**	**V**	**0.000**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.000**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.000**	$\mathbf{1}^{\dagger}$
True Positive Rate (TPR)	**V**	**V**	**V**	**0.000**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.000**	$\mathbf{1}^{\dagger}$	**V**	**V**	**V**	**0.000**	$\mathbf{1}^{\dagger}$
Balanced Accuracy	V	X	X	0.486	0.713	V	V	V	0.504	$1^{\dagger}$	V	V	V	0.505	$1^{\dagger}$
Cohen's κ	X	X	X	0.476	0.697	V	V	V	0.503	$1^{\dagger}$	V	V	V	0.505	$1^{\dagger}$
Informedness	V	X	X	0.486	0.713	V	V	V	0.504	$1^{\dagger}$	V	V	V	0.505	$1^{\dagger}$
Positive Likelihood Ratio (PLR)	V	X	X	0.420	0.677	V	V	V	0.491	$1^{\dagger}$	V	V	V	0.491	$1^{\dagger}$
Probability of True Negative (PTN)	X	V	V	-0.007	0.818	V	V	V	0.000	$1^{\dagger}$	V	V	V	0.000	$1^{\dagger}$
Probability of True Positive (PTP)	X	V	V	-0.006	0.818	V	V	V	0.000	$1^{\dagger}$	V	V	V	0.000	$1^{\dagger}$
Chance in Cohen's κ	X	X	X	0.194	0.498	X	V	V	-0.157	0.849	V	V	V	$0^{\dagger}$	$0^{\dagger}$
Error Rate	X	V	V	$-1^{\dagger}$	-0.469	X	V	V	$-1^{\dagger}$	-0.157	X	V	V	$-1^{\dagger}$	-0.505
False Discovery Rate (FDR)	X	V	V	$-1^{\dagger}$	0.000	X	V	V	$-1^{\dagger}$	-0.503	X	V	V	$-1^{\dagger}$	-0.503
False Negative Rate (FNR)	X	V	V	$-1^{\dagger}$	0.000	X	V	V	$-1^{\dagger}$	0.000	X	V	V	$-1^{\dagger}$	0.000
False Omission Rate (FOR)	X	V	V	$-1^{\dagger}$	0.000	X	V	V	$-1^{\dagger}$	-0.503	X	V	V	$-1^{\dagger}$	-0.503
False Positive Rate (FPR)	X	V	V	$-1^{\dagger}$	0.000	X	V	V	$-1^{\dagger}$	0.000	X	V	V	$-1^{\dagger}$	0.000
Geometric mean of TNR and TPR	V	X	X	0.461	0.653	V	X	V	0.503	0.831	V	X	V	0.503	0.830
Markedness	V	X	X	0.486	0.713	V	X	X	0.418	0.887	V	X	X	0.503	0.913
Matthews Correlation Coefficient (MCC)	V	X	X	0.503	0.746	V	X	X	0.458	0.944	V	X	X	0.503	0.963
Negative Likelihood Ratio (NLR)	X	X	X	-0.677	-0.418	X	V	V	$-1^{\dagger}$	-0.491	X	V	V	$-1^{\dagger}$	-0.491
Odds Ratio (OR)	V	X	X	0.499	0.671	V	X	X	0.503	0.894	V	X	X	0.503	0.892
Rate of positive predictions	X	V	V	-0.469	0.469	X	V	V	-0.849	0.157	X	V	V	-0.504	0.505
Sensitivity Index Estimate (d')	V	X	X	0.502	0.786	V	X	X	0.503	0.926	V	X	X	0.503	0.924

are those that satisfy the three tests only for fixed priors. In black are the others. We draw 4 conclusions. (1) Performance orderings satisfying our axioms can be induced by several classical scores for two-class classification (in green). (2) There exist scores that are commonly used in the literature (in orange) that cannot be used for ranking classifiers, unless the priors are fixed. (3) There exist scores that are often used to compare classifiers (*e.g.*, the geometric mean of TNR and TPR, the markedness, Matthews Correlation Coefficient, the Odds Ratio) but that cannot be used to rank, even when the priors are fixed. (4) In all studied cases where our axioms are satisfied, there is a perfect correlation with a ranking score. This shows that our family of scores covers at least a broad part of the needs. Only one exception occurred (in gray): the accuracy achievable by chance, as considered by Cohen in the definition of his κ [4], when the classes are balanced. In this case, the score takes a constant value, that is why it satisfies our axioms.

6. Conclusion

We present a mathematical framework for a theory of performance-based ranking, that comes with several practical benefits. Notably, it provides a universal language to properly define tasks and characterize applications. Also, separating these concepts allows evaluating entities (algorithms, devices, *etc*.) independently of application-specific preferences, as an entity can be used in various use cases, and various entities can be candidates for a same use case.

Importantly, our axioms are crucial for organizers of challenges to ensure sound rankings, to avoid, *e.g*., that the relative order between two methods changes when a new method appears and thus prevents drawing perennial conclusions on which is better. Our axioms are minimal requirements to satisfy, and we provide practical theorems to help check if that is the case.

Besides, practitioners often face many scores in the literature for a task, but none comes with an analysis of their usability for ranking. We prove that our theory can help find appropriate scores (we propose an infinite family of them) or legitimate existing ones. As the chosen performance score for ranking may have a major impact on the growth of a research field, our framework clarifies best practices.

We can further deepen our axiomatic framework and infinite family of scores for ranking classifiers. In [21], we particularize extensively this framework to binary classification and present the *Tile*, a visualization tool that organizes these scores (among which the precision PPV, the true positive rate TPR, the true negative rate TNR, the scores F_β, and the accuracy A) in a single plot. Finally, in [12], we provide a comprehensive guide to using the Tile according to four practical scenarios. For that purpose, we present different Tile flavors on a real example, analyzing and ranking 74 segmentation classifiers. We now wish to build upon this trilogy of papers to reach various research communities and impact significantly their way of establishing performance-based rankings.

Acknowledments. S. Piérard and A. Halin are funded by the SPW EER, Wallonia, Belgium (grant 2010235, ARIAC by DIGITALWALLONIA4.AI); A. Cioppa and A. Deliège (grant T.0065.22) by the F.R.S.-FNRS.

References

[1] Davide Ballabio, Francesca Grisoni, and Roberto Todeschini. Multivariate comparison of classification performance measures. *Chemom. Intell. Lab. Syst.*, 174:33–44, 2018. 17

[2] Gurol Canbek, Seref Sagiroglu, Tugba Taskaya Temizel, and Nazife Baykal. Binary classification performance measures/metrics: A comprehensive visualized roadmap to gain new insights. In *Int. Conf. Comput. Sci. Eng. (UBMK)*, pages 821–826, Antalya, Turkey, 2017. 2, 4, 7

[3] Anthony Cioppa, Silvio Giancola, Vladimir Somers, Floriane Magera, Xin Zhou, Hassan Mkhallati, Adrien Deliège, Jan Held, Carlos Hinojosa, Amir M. Mansourian, Pierre Miralles, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdullah Kamal, Adrien Maglo, Albert Clapés, Amr Abdelaziz, Artur Xarles, Astrid Orcesi, Atom Scott, Bin Liu, Byoungkwon Lim, Chen Chen, Fabian Deuser, Feng Yan, Fufu Yu, Gal Shitrit, Guanshuo Wang, Gyusik Choi, Hankyul Kim, Hao Guo, Hasby Fahrudin, Hidenari Koguchi, Håkan Ardö, Ibrahim Salah, Ido Yerushalmy, Iftikar Muhammad, Ikuma Uchida, Ishay Be'ery, Jaonary Rabarisoa, Jeongae Lee, Jiajun Fu, Jianqin Yin, Jinghang Xu, Jongho Nang, Julien Denize, Junjie Li, Junpei Zhang, Juntae Kim, Kamil Synowiec, Kenji Kobayashi, Kexin Zhang, Konrad Habel, Kota Nakajima, Licheng Jiao, Lin Ma, Lizhi Wang, Luping Wang, Menglong Li, Mengying Zhou, Mohamed Nasr, Mohamed Abdelwahed, Mykola Liashuha, Nikolay Falaleev, Norbert Oswald, Qiong Jia, Quoc-Cuong Pham, Ran Song, Romain Hérault, Rui Peng, Ruilong Chen, Ruixuan Liu, Ruslan Baikulov, Ryuto Fukushima, Sergio Escalera, Seungcheon Lee, Shimin Chen, Shouhong Ding, Taiga Someya, Thomas B. Moeslund, Tianjiao Li, Wei Shen, Wei Zhang, Wei Li, Wei Dai, Weixin Luo, Wending Zhao, Wenjie Zhang, Xinquan Yang, Yanbiao Ma, Yeeun Joo, Yingsen Zeng, Yiyang Gan, Yongqiang Zhu, Yujie Zhong, Zheng Ruan, Zhiheng Li, Zhijian Huang, and Ziyu Meng. SoccerNet 2023 challenges results. *Sports Eng.*, 27(2):1–18, 2024. 1

[4] Jacob Cohen. A coefficient of agreement for nominal scales. *Educ. Psychol. Meas.*, 20(1):37–46, 1960. 8

[5] Andrew R. Conn, Katya Scheinberg, and Luis N. Vicente. *Introduction to Derivative-Free Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2009. 31

[6] Brooke A. Davey and Hilary A. Priestley. *Introduction to Lattices and Order*. Camb. Univ. Press, second edition, 2002. 2, 3

[7] Jesse Davis and Mark Goadrich. The relationship between precision-recall and ROC curves. In *Int. Conf. Mach. Learn. (ICML)*, pages 233–240, Pittsburgh, Pennsylvania, 2006. 4

[8] Tom Fawcett. An introduction to ROC analysis. *Pattern Recognit. Lett.*, 27(8):861–874, 2006. 4

[9] Edward B. Fowlkes and Colin L. Mallows. A method for comparing two hierarchical clusterings. *J. Am. Stat. Assoc.*, 78(383):553–569, 1983. 17

[10] Nil Goyette, Pierre-Marc Jodoin, Fatih Porikli, Janusz Konrad, and Prakash Ishwar. A novel video dataset for change detection benchmarking. *IEEE Trans. Image Process.*, 23 (11):4663–4679, 2014. 1

[11] George Grätzer. *General Lattice Theory*. Birkhäuser Basel, second edition, 2003. 2, 3

[12] Anaïs Halin, Sébastien Piérard, Anthony Cioppa, and Marc Van Droogenbroeck. A hitchhiker's guide to understanding performances of two-class classifiers. *arXiv*, abs/2412.04377, 2024. 8, 14

[13] Maurice George Kendall. A new measure of rank correlation. *Biometrika*, 30(1/2):81, 1938. 2, 13, 18, 31

[14] Andrey Nikolaevich Kolmogorov. Grundbegriffe der wahrscheinlichkeitsrechnung. In *Grundbegriffe der Wahrscheinlichkeitsrechnung*, page 62 pages. Springer Berl. Heidelb., 1933. 2, 3

[15] Andrey Nikolaevich Kolmogorov. *Foundations of the theory of probability*. Chelsea Publishing Company, 1950. 2, 3

[16] Matej Kristan, Jiri Matas, Ale Leonardis, Michael Felsberg, Luka Cehovin, Gustavo Fernandez, Toma Vojir, Gustav Hager, Georg Nebehay, Roman Pflugfelder, Abhinav Gupta, Adel Bibi, Alan Lukezic, Alvaro Garcia-Martin, Amir Saffari, Alfredo Petrosino, and Andrés Solis Montero. The visual object tracking VOT2015 challenge results. In *IEEE Int. Conf. Comput. Vis. Work. (ICCV Work.)*, pages 564–586, Santiago, Chile, 2015. 1

[17] Matej Kristan, Jiri Matas, Pavel Tokmakov, Michael Felsberg, Luka Čehovin Zajc, Alan Lukežič, Khanh-Tung Tran, Xuan-Son Vu, Johanna Bjorklund, Hyung Jin Chang, and Gustavo Fernandez. The second visual object tracking segmentation VOTS2024 challenge results. In *Eur. Conf. Comput. Vis. Work. (ECCV Work.)*, Milan, Italy, 2024. 1

[18] Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P. Bradley, Aaron Carass, Carolin Feldmann, Alejandro F. Frangi, Peter M. Full, Bram van Ginneken, Allan Hanbury, Katrin Honauer, Michal Kozubek, Bennett A. Landman, Keno März, Oskar Maier, Klaus Maier-Hein, Bjoern H. Menze, Henning Müller, Peter F. Neher, Wiro Niessen, Nasir Rajpoot, Gregory C. Sharp, Korsuk Sirinukunwattana, Stefanie Speidel, Christian Stock, Danail Stoyanov, Abdel Aziz Taha, Fons van der Sommen, Ching-Wei Wang, Marc-André Weber, Guoyan Zheng, Pierre Jannin, and Annette Kopp-Schneider. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat. Commun.*, 9(1):1–13, 2018. 1, 2

[19] Tran Thien Dat Nguyen, Hamid Rezatofighi, Ba-Ngu Vo, Ba-Tuong Vo, Silvio Savarese, and Ian Reid. How trustworthy are performance evaluations for basic vision tasks? *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(7):8538–8552, 2023. 1, 2

[20] Sébastien Piérard and Marc Van Droogenbroeck. Summarizing the performances of a background subtraction algorithm measured on several videos. In *IEEE Int. Conf. Image*

Process. (ICIP), pages 3234–3238, Abu Dhabi, United Arab Emirates, 2020. 4

[21] Sébastien Piérard, Anaïs Halin, Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. The Tile: A 2D map of ranking scores for two-class classification. *arXiv*, abs/2412.04309, 2024. 8, 14

[22] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Inf. Process. & Manag.*, 45(4):427–437, 2009. 14

[23] Charles Spearman. The proof and measurement of association between two things. *Am. J. Psychol.*, 15(1):72–101, 1904. 2, 31

[24] Putnam P. Texel. Measure, metric, and indicator: An object-oriented approach for consistent terminology. *Proc. IEEE Southeastcon*, pages 1–5, 2013. 2, 4

[25] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C. J. Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, Aditya Vijaykumar, Alessandro Pietro Bardelli, Alex Rothberg, Andreas Hilboll, Andreas Kloeckner, Anthony Scopatz, Antony Lee, Ariel Rokem, C. Nathan Woods, Chad Fulton, Charles Masson, Christian Häggström, Clark Fitzgerald, David A. Nicholson, David R. Hagen, Dmitrii V. Pasechnik, Emanuele Olivetti, Eric Martin, Eric Wieser, Fabrice Silva, Felix Lenders, Florian Wilhelm, G. Young, Gavin A. Price, Gert-Ludwig Ingold, Gregory E. Allen, Gregory R. Lee, Hervé Audren, Irvin Probst, Jörg P. Dietrich, Jacob Silterra, James T. Webber, Janko Slavič, Joel Nothman, Johannes Buchner, Johannes Kulick, Johannes L. Schönberger, José Vinícius de Miranda Cardoso, Joscha Reimer, Joseph Harrington, Juan Luis Cano Rodríguez, Juan Nunez-Iglesias, Justin Kuczynski, Kevin Tritz, Martin Thoma, Matthew Newville, Matthias Kümmerer, Maximilian Bolingbroke, Michael Tartre, Mikhail Pak, Nathaniel J. Smith, Nikolai Nowaczyk, Nikolay Shebanov, Oleksandr Pavlyk, Per A. Brodtkorb, Perry Lee, Robert T. McGibbon, Roman Feldbauer, Sam Lewis, Sam Tygier, Scott Sievert, Sebastiano Vigna, Stefan Peterson, Surhud More, Tadeusz Pudlik, Takuya Oshima, Thomas J. Pingel, Thomas P. Robitaille, Thomas Spura, Thouis R. Jones, Tim Cera, Tim Leslie, Tiziano Zito, Tom Krauss, Utkarsh Upadhyay, Yaroslav O. Halchenko, and Yoshiki Vázquez-Baeza. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods*, 17(3):261–272, 2020. 7

A. Supplementary Material

Contents

A.1. List of Symbols

A.1.1 Mathematical Symbols

- $\mathbf{1}_U$: the 0-1 indicator function of subset U
- $\mathbb{R}$: the real numbers
- $\mathcal{R}$: a relation
- conv: the set of convex combinations
- $\vee$: the *inclusive disjunction* (*i.e.*, logical or)
- $\wedge$: the *conjunction* (*i.e.*, logical and)
- $\circ$: the composition of functions, *i.e.* $(g \circ f)(x) = g(f(x))$
- **E**: the mathematical expectation

A.1.2 Symbols Related to Our Mathematical Framework

We organize these symbols according to the 6 pillars depicted in Fig. 1, which correspond to the 6 subsections of Sec. 3.

Symbols related to the 1$^{\text{st}}$ pillar (Sec. 3.1)

- Ω: the sample space (universe)
- ω: a sample (*i.e.*, an element of Ω)
- Σ: the event space (a σ-algebra on Ω, *e.g.* 2^Ω)
- E: an event (*i.e.*, an element of Σ)
- (Ω, Σ): the measurable space
- $\mathbb{P}_{(\Omega,\Sigma)}$: all performances on (Ω, Σ)
- Π: a set of performances ($\Pi \subseteq \mathbb{P}_{(\Omega,\Sigma)}$)
- P: a performance (*i.e.*, an element of $\mathbb{P}_{(\Omega,\Sigma)}$)

Symbols related to the 2$^{\text{nd}}$ pillar (Sec. 3.2)

- $\lesssim$: binary relation *worse or equivalent* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $\gtrsim$: binary relation *better or equivalent* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $\sim$: binary relation *equivalent* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $>$: binary relation *better* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $<$: binary relation *worse* on $\mathbb{P}_{(\Omega,\Sigma)}$
- $\not\lesseqqgtr$: binary relation *incomparable* on $\mathbb{P}_{(\Omega,\Sigma)}$

Symbols related to the 3$^{\text{rd}}$ pillar (Sec. 3.3)

- S: the random variable *Satisfaction*
- $s_{min,\Omega}$: the minimum satisfaction value
- $s_{max,\Omega}$: the maximum satisfaction value

Symbols related to the 4$^{\text{th}}$ pillar (Sec. 3.4)

- $\mathbb{E}$: the set of entities to rank
- ϵ: an entity, *i.e.* an element of $\mathbb{E}$
- eval: the performance *evaluation* function
- Φ: some performances that are for sure achievable

Symbols related to the 5$^{\text{th}}$ pillar (Sec. 3.5)

- X: a score
- $\mathrm{dom}(X)$: the domain of the score X
- X^E_V: the *expected value score* parameterized by the random variable V
- $X^P_{E_1|E_2}$: the *probabilistic score* parameterized by the events E_1 and E_2

Symbols related to the 6th pillar (Sec. 3.6)

- I: the random variable *Importance*

A.1.3 Symbols used for Operations on Performances

- filter_I: the *filtering* operation

A.1.4 Symbols used in the Performance Ordering and Performance-Based Ranking Theory

- $\mathrm{rank}_{\mathbb{E}}$: the *ranking* function, w.r.t. the set of entities $\mathbb{E}$
- $\lesssim_X$: the ordering induced by the score X (*cf*. Theorem 1)
- R_I: the *ranking score* parameterized by the importance I
- τ: the rank correlation coefficient of Kendall [13]

A.1.5 Symbols used for the Particular Case of Two-Class Crisp Classifications

Particularization of the mathematical framework

- tn: the sample *true negative*
- fp: the sample *false positive*, *a.k.a.* type I error
- fn: the sample *false negative*, *a.k.a.* type II error
- tp: the sample *true positive*

Extensions to the mathematical framework

- ROC: the *Receiver Operating Characteristic* space, *i.e.* $FPR \times TPR$
- PR: the *Precision-Recall* space, *i.e.* $TPR \times PPV$
- Y: the random variable for the ground truth
- $\hat{Y}$: the random variable for the prediction
- $\mathbb{C}$: the set of classes
- c: a class (*i.e.*, an element of $\mathbb{C}$)
- c_-: the negative class
- c_+: the positive class

Scores

- A: the *accuracy*
- TNR: the *true negative rate*
- FPR: the *false positive rate*
- TPR: the *true positive rate*
- PPV: the *positive predictive value*
- F_β: the F-scores
- π_+: the *prior of the positive class*
- π_-: the *prior of the negative class*

A.2. How to Use our Framework: a Little Catalog of Problems

Throughout the paper, we have exemplified our theory with the problem of two-class classification. This section aims at showing the universality of our theory. It presents a little catalog of other problems, together with discussions on how to use our framework for them. These discussions are introductions. As shown by Sec. 5 and two recent works [12, 21], an in-depth analysis and particularization of our theory to the various problems (*e.g.*, to highlight their distinctive features, to review the current ranking practices from the literature and the consistency of popular scores, and to establish practical tools tailored to different user needs) may require substantial work that is out of the scope of this supplementary material.

In the following, we adopt a systematic approach: for each problem, we start by specifying a thought random experiment for the evaluation (this is an arbitrary choice since the random experiment is not unique for each problem; do not hesitate to use a different random experiment, the important is to specify it explicitly!), then we discuss the possible choices for the sample space Ω (and what the set of all performances is), for the modeling of tasks with the satisfaction S, for the modeling of the knowledge we have about the evaluation with the function Φ, and for the modeling of the application-specific preferences with the importance I.

A.2.1 Multi-Class Classification (with a Note on Micro- and Macro-Averaging)

Let us consider the following thought experiment to evaluate classifiers predicting classes in a finite and non-empty set $\mathbb{C}$.

Random Experiment 1. *(1) Draw a sample $s \in \mathbb{S}$ at random from a given* source $\mathcal{S}$. *(2) Apply the* oracle $\mathcal{O}$ *on s to obtain the ground-truth class $y(s) \in \mathbb{C}$. (3) Apply a* descriptor $\mathcal{D}$ *on s to obtain the features (*a.k.a. *attributes) $x(s) \in \mathbb{X}$. (4) Feed the* classifier $\mathcal{C}$ *with $x(s)$ to obtain the predicted class $\hat{y}(x(s)) \in \mathbb{C}$. (5) Set the outcome of the experiment to the pair $(y, \hat{y}) \in \mathbb{C}^2$.*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \mathbb{C}^2$. By definition, the function eval gives, for any classifier $\mathcal{C} : \mathbb{X} \to \mathbb{C}$ (the evaluated entity, either deterministic or not), the distribution of outcomes resulting from this random experiment: $P_{\mathcal{C}} = \mathrm{eval}(\mathcal{C})$. Note that it is implicit that the performances are specific for some given source $\mathcal{S}$ (*e.g.*, evaluation dataset), oracle $\mathcal{O}$, and descriptor $\mathcal{D}$. By convenience, one can manipulate the ground-truth and predicted classes with, respectively, the random variables Y and $\hat{Y}$ defined in such a way that $\omega = (Y(\omega), \hat{Y}(\omega)) \, \forall \omega \in \Omega$.

Choice for S. Several classification tasks can be distinguished, as the following two examples show. (1) One can consider that all erroneous classifications are unsatisfactory and that correct classifications are satisfactory. For this task, the satisfaction is then binary and given by $S = \mathbf{1}_{Y=\hat{Y}}$. The expected value of the satisfaction, which is a particular case of ranking scores, is then equal to the multi-class accuracy. (2) One can also consider the similarity $\mathrm{sim} : \mathbb{C}^2 \to \mathbb{R}$ between classes, and choose the satisfaction accordingly: $S(\omega) = \mathrm{sim}(Y(\omega); \hat{Y}(\omega))$. A wide variety of tasks can be considered by tuning sim. In general, the expected value of the satisfaction is different from the multi-class accuracy.

Choice for Φ. As the classifier is used once and only once during the execution of the evaluation, we know that if the performances P_1 and P_2 are achievable by some classifiers $\mathcal{C}_1$ and $\mathcal{C}_2$, then any performance $\overline{P} = \lambda_1 P_1 + \lambda_2 P_2$ (with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_1 + \lambda_2 = 1$) is achievable by a classifier $\overline{\mathcal{C}}$ obtained by a non-deterministic combination of $\mathcal{C}_1$ and $\mathcal{C}_2$ that chooses them with respective probabilities λ_1 and λ_2. Thus, $\Phi = \mathrm{conv}$ makes sense, and all ranking scores can be used to rank classifiers. However, it would be possible to go further, by considering other functions Φ that would include the knowledge that we can predict the performance achievable by composing the classifier with any of the $|\mathbb{C}|^{|\mathbb{C}|}$ functions $f : \mathbb{C} \to \mathbb{C}$. This would lead to other performance orderings suitable for ranking classifiers.

Choice for I. When $\Phi = \mathrm{conv}$, we have demonstrated that all rankings induced by the ranking scores R_I satisfy all our three axioms. This leaves a great flexibility for the users to fine-tune the ranking w.r.t. their application-specific preferences through the random variable I. As we have seen, the only constraints are that $I \neq 0$ and $I(\omega) \geq 0 \, \forall \omega \in \Omega$.

Note on micro- and macro-averaging. Micro- and macro-averaging are commonly used techniques to build scores for multi-class classification from scores for two-class classification [22]. We warn that they have pitfalls. In general, micro- and macro-averaging scores suitable for ranking two-class classifiers do not lead to scores suitable for ranking multi-class classifiers. The accuracy put aside, the performance orderings induced from the micro-averaged versions of the scores put in green in Tab. 1 are incompatible with $S = \mathbf{1}_{Y=\hat{Y}}$: our 2nd axiom is not satisfied. Moreover, the accuracy put again aside, the performance orderings induced from the macro-averaged versions of the scores put in green in Tab. 1 are incompatible with $\Phi = \mathrm{conv}$: our 3rd axiom is not satisfied. A solution consists in using directly ranking scores defined for multi-class classification.

A.2.2 Regression (with a Note on the Mean Squared Error and the Mean Absolute Error)

Let us consider the following thought experiment to evaluate regressors.

Random Experiment 2. *(1) Draw a sample $s \in \mathbb{S}$ at random from a given* source $\mathcal{S}$. *(2) Apply the* oracle $\mathcal{O}$ *on s to obtain the ground-truth value $y(s) \in \mathbb{R}$. (3) Apply a* descriptor $\mathcal{D}$ *on s to obtain the features (*a.k.a. *attributes) $x(s) \in \mathbb{X}$. (4) Feed the* regressor $\mathcal{R}$ *with $x(s)$ to obtain the predicted value $\hat{y}(x(s)) \in \mathbb{R}$. (5) Set the outcome of the experiment to the pair $(y, \hat{y}) \in \mathbb{R}^2$.*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \mathbb{R}^2$. By definition, the function eval gives, for any regressor $\mathcal{R} : \mathbb{X} \rightarrow \mathbb{R}$ (the evaluated entity, either deterministic or not), the distribution of outcomes resulting from this random experiment. It is the performance $P_{\mathcal{R}} = \mathrm{eval}(\mathcal{R})$ of $\mathcal{R}$. By convenience, one can manipulate the ground-truth and predicted values with, respectively, the random variables Y and $\hat{Y}$ defined in such a way that $\omega = (Y(\omega), \hat{Y}(\omega))\, \forall \omega \in \Omega$.

Choice for S. Clearly, it is not a good idea to choose $S = \mathbf{1}_{Y=\hat{Y}}$, similarly as one can do in classification. In practice, a regressor as no chance to predict the same value as the oracle, so this unfortunate choice for S would lead to performances P such that $P(S = 0) = 1$. In other words, all performances observed about real regressors would belong to the set of the worst performances (see Corollary 2), and their ranking would be of little interest (see Corollary 1). A better option consists in specifying a tolerance $\epsilon > 0$ and choosing $S = \mathbf{1}_{|Y-\hat{Y}|\leq\epsilon}$. An even more flexible option, which takes advantage of the fact that satisfaction values do not necessarily have to be positive, is to choose $S = f(|Y - \hat{Y}|)$ with any arbitrarily chosen monotonically decreasing function f. The plethora of choices that can be made for S makes it clear that there is an infinity of tasks related to the regression problem.

Choice for Φ. As the regressor is used once and only once during the execution of the evaluation, we know that if the performances P_1 and P_2 are achievable by some regressors $\mathcal{R}_1$ and $\mathcal{R}_2$, then any performance $\overline{P} = \lambda_1 P_1 + \lambda_2 P_2$ (with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_1 + \lambda_2 = 1$) is achievable by a regressor $\overline{\mathcal{R}}$ obtained by a non-deterministic combination of $\mathcal{R}_1$ and $\mathcal{R}_2$ that chooses them with respective probabilities λ_1 and λ_2. Thus, $\Phi = \mathrm{conv}$ makes sense, and all ranking scores can be used to rank regressors. However, it would be possible to go further, by considering other functions Φ that would include the knowledge that we can predict the performance achievable by adding noise, or applying a transformation on the output of the regressor. This would lead to other performance orderings suitable for ranking regressors.

Choice for I. When $\Phi = \mathrm{conv}$, we have demonstrated that all rankings induced by the ranking scores R_I satisfy all our three axioms. This leaves a great flexibility for the users to fine-tune the ranking w.r.t. their application-specific preferences through the random variable I. As we have seen, the only constraints are that $I \neq 0$ and $I(\omega) \geq 0\, \forall \omega \in \Omega$.

Note on the mean squared error and the mean absolute error. If we choose $S = -|Y - \hat{Y}|^2$, the ranking score R_I corresponding to uniform importance values, *i.e.*, the expected satisfaction X_S^E, yields a ranking that minimizes the *mean squared error* (MSE). If we choose $S = -|Y - \hat{Y}|$, the ranking score R_I corresponding to uniform importance values, *i.e.*, the expected satisfaction X_S^E, yields a ranking that minimizes the *mean absolute error* (MAE).

A.2.3 Information Retrieval

Let us consider the following thought experiment to evaluate information retrieval systems. We denote by $\mathbb{Q}$ the set of all possible queries.

Random Experiment 3. *(1) Draw a query $q \in \mathbb{Q}$ at random from a given source $\mathcal{S}$. (2) Apply the oracle $\mathcal{O}$ on q to obtain the ground-truth set of results $\mathbb{Y}$. (3) Apply the evaluated information retrieval system $\mathcal{S}$ on q to obtain the predicted set of results $\hat{\mathbb{Y}}$. (4) If $\mathbb{Y} = \emptyset$ and $\hat{\mathbb{Y}} = \emptyset$, restart the experiment, otherwise draw a result r at random in $\mathbb{Y} \cup \hat{\mathbb{Y}}$. (5) Choose the outcome as follows: fp if $r \notin \mathbb{Y}$ and $r \in \hat{\mathbb{Y}}$, fn if $r \in \mathbb{Y}$ and $r \notin \hat{\mathbb{Y}}$, and tp if $r \in \mathbb{Y}$ and $r \in \hat{\mathbb{Y}}$.*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \{fp, fn, tp\}$. By definition, the function eval gives, for any retrieval system $\mathcal{S}$ defined on $\mathbb{Q}$ (the evaluated entity, either deterministic or not), the distribution of outcomes resulting from this random experiment: $P_{\mathcal{S}} = \mathrm{eval}(\mathcal{S})$.

Choice for S. Intuitively, everyone certainly agrees that $S(fp) < S(tp)$ and $S(fn) < S(tp)$. But we expect different opinions regarding whether the outcome (sample) fp gives less, equal, or more satisfaction than fn.

Choice for Φ. This random experiment is very interesting as, during its execution, the evaluated entity (the information retrieval system $\mathcal{S}$) can be used multiple times. In such a case, we have to discuss whether $\Phi = \mathrm{conv}$ is adequate. Let us consider two systems $\mathcal{S}_1, \mathcal{S}_2$ and their respective performances $P_1 = \mathrm{eval}(\mathcal{S}_1)$, $P_2 = \mathrm{eval}(\mathcal{S}_2)$. It is possible to show that the performance of a retrieval system $\overline{\mathcal{S}}$ obtained by a non-deterministic combination of $\mathcal{S}_1$ and $\mathcal{S}_2$, that chooses them with respective probabilities λ_1 and λ_2, is some interpolated performance $\overline{P} = \mu_1 P_1 + \mu_2 P_2$ (with $\mu_1 \geq 0$, $\mu_2 \geq 0$, and $\mu_1 + \mu_2 = 1$). Unless being in very particular cases, $\mu \neq \lambda$. In other words, we know that the performances that are convex combinations of achievable performances are also achievable (for any λ, there exists μ), but we do not know in general how to achieve them (for most μ it is not possible to determine λ). This contrasts with the other kinds of problems discussed in this catalog. In fact, the question we raise here is not specific to the information retrieval problem: it is peculiar to the random experiment that we have chosen for it. By slightly modifying the thought experiment, the question vanishes: instead of restarting the experiment when $\mathbb{Y} \cup \hat{\mathbb{Y}} = \emptyset$, one could yield a fourth outcome (and add it to the sample space Ω). By doing so, the evaluated entity $\mathcal{S}$ is used only once and $\Phi = \mathrm{conv}$ makes sense for sure.

Choice for I. If $\Phi = \mathrm{conv}$ is considered as adequate, then we have demonstrated that all rankings induced by the ranking scores R_I satisfy all our three axioms. This leaves a great flexibility for the users to fine-tune the ranking w.r.t. their application-specific preferences through the random variable I. As we have seen, the only constraints are that $I \neq 0$ and $I(\omega) \geq 0 \, \forall \omega \in \Omega$.

A.2.4 Detection (with a Note about the Intersection-over-Union and the F-Score)

Different types of detections are present in the literature. An example of spatial detection aims at predicting the axis-aligned bounding boxes around all the objects that match some given properties (*i.e.*, a semantic class) in input images. Examples of temporal detections include the detection of events in video streams and in audio recordings. By definition, such detection problems are called *action spotting* when the temporal window is small, and *activity detection* otherwise. Let us consider the following, generic, thought experiment to evaluate detectors.

Random Experiment 4. *(1) Draw an input at random from a given* source $\mathcal{S}$ *(*e.g.*, dataset). (2) Apply the* oracle $\mathcal{O}$ *on it to obtain a set* $\mathbb{Y}$ *of ground-truth detections. (3) Also apply the* detector $\mathcal{D}$ *on it to obtain a set* $\hat{\mathbb{Y}}$ *of predicted detections. (4) If* $\mathbb{Y} = \emptyset$ *and* $\hat{\mathbb{Y}} = \emptyset$*, then end the experiment with the outcome* $\circledast$*. Otherwise: (5) Apply a matching criterion between* $\mathbb{Y}$ *and* $\hat{\mathbb{Y}}$ *such that to any detection in* $\mathbb{Y}$ *should be associated at most a detection in* $\hat{\mathbb{Y}}$ *and vice versa. (6) Draw a detection* d *at random in* $\mathbb{Y} \cup \hat{\mathbb{Y}}$*. (7) Give as outcome* fp*,* fn *or* tp *depending on whether* d *is a prediction, a ground truth, or both (*i.e.*, a match).*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \{\circledast, fp, fn, tp\}$. By definition, the function eval gives, for any detector (the evaluated entity), the distribution of outcomes resulting from this random experiment. It is the performance $P_{\mathcal{D}} = \mathrm{eval}(\mathcal{D})$ of $\mathcal{D}$.

Choice for S. Intuitively, everyone certainly agrees that $\circledast$ and tp give entire satisfaction. Moreover, we expect agreement on $S(fp) < S(tp)$ and $S(fn) < S(tp)$. But we expect different opinions regarding whether fp gives less, equal, or more satisfaction than fn.

Choice for Φ. As the detector is used once and only once during the execution of the evaluation, we know that if the performances P_1 and P_2 are achievable by some detectors $\mathcal{D}_1$ and $\mathcal{D}_2$, then any performance $\overline{P} = \lambda_1 P_1 + \lambda_2 P_2$ (with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_1 + \lambda_2 = 1$) is achievable by a detector $\overline{\mathcal{D}}$ obtained by a non-deterministic combination of $\mathcal{D}_1$ and $\mathcal{D}_2$ that chooses them with respective probabilities λ_1 and λ_2. Thus, $\Phi = \mathrm{conv}$ makes sense, and all ranking scores can be used to rank detectors.

Choice for I. If $\Phi = \mathrm{conv}$ is considered as adequate, then we have demonstrated that all rankings induced by the ranking scores R_I satisfy all our three axioms. This leaves a great flexibility for the users to fine-tune the ranking w.r.t. their application-specific preferences through the random variable I. As we have seen, the only constraints are that $I \neq 0$ and $I(\omega) \geq 0 \, \forall \omega \in \Omega$.

Note about the Intersection-over-Union and the F-score. Traditionally, in the literature, as soon as one has symbols fp, fn, and tp, regardless of their very fine meaning, one defines quantities $IoU = \frac{P(tp)}{P(fp)+P(fn)+P(tp)}$ and $F_1 = \frac{2P(tp)}{P(fp)+P(fn)+2P(tp)}$, and name them *Intersection-over-Union* (or *Jaccard*) and *F-one*, respectively. The exact meaning of these quantities is not well standardized. In particular, the random experiment supporting the evaluation, if it exists, is rarely specified explicitly. For this reason, we cannot give the guarantee that these quantities are suitable to rank detectors in all works in which they have been used. However, with the random experiment given here-above, with $\Phi = \mathrm{conv}$, and with $S = \mathbf{1}_{\{\otimes,tp\}}$, we can guarantee that IoU and F_1 are suitable to rank detectors because they are equal to the ranking scores with, respectively, $I = \mathbf{1}_{\{fp,fn,tp\}}$ and $I = \mathbf{1}_{\{fp,tp\}} + \mathbf{1}_{\{fn,tp\}}$. Thus, the performance orderings induced by them fulfill our three axioms.

A.2.5 Clustering (with a Note about Fowlkes-Mallows Index)

Let us consider the following thought experiment to evaluate clustering methods. These methods aim to place in different clusters (groups) dissimilar objects and in the same cluster (group) objects that are similar to each other. We denote by $\mathbb{E}$ the set of elements that these methods have to deal with. For the sake of simplicity, we do not consider hierarchical clustering.

Random Experiment 5. *(1) Apply both the* clustering method $\mathcal{C}$ *and the* oracle $\mathcal{O}$ *on* $\mathbb{E}$ *to obtain, respectively, the predicted and ground-truth clusterings. (2) Randomly draw two distinct elements,* ϵ_1 *and* ϵ_2*, from* $\mathbb{E}$*. (3) Consider that the pair* $(\epsilon_1; \epsilon_2)$ *is a negative or a positive, in a given clustering, when* ϵ_1 *and* ϵ_2 *are in different clusters or in the same cluster, respectively. (4) Choose the outcome as follows:* tn *when* $(\epsilon_1; \epsilon_2)$ *is negative in both the predicted and ground-truth clusterings,* fp *when* $(\epsilon_1; \epsilon_2)$ *is negative in the ground-truth clustering and positive in the predicted clustering,* fn *when* $(\epsilon_1; \epsilon_2)$ *is positive in the ground-truth clustering and negative in the predicted clustering, and* tp *when* $(\epsilon_1; \epsilon_2)$ *is positive in both the predicted and ground-truth clusterings.*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \{tn, fp, fn, tp\}$. By definition, the function eval gives, for any clustering method (the evaluated entity), the distribution of outcomes resulting from this random experiment. It is the performance $P_{\mathcal{C}} = \mathrm{eval}(\mathcal{C})$ of $\mathcal{C}$.

Choice for S. When S is chosen such that $S(fp) = S(fn) = 0$ and $S(tn) = S(tp) = 1$, we are in the same setting as the one we studied for the two-class classification in Sec. 5. This is indeed not because we use the same symbols for the elements of Ω —this is just a convention—, but because, in both settings, we have $|\Omega| = 4$, $|S = 0| = 2$, and $|S = 1| = 2$. This implies that the performance orderings that satisfy our three axioms for ranking two-class classifiers can also be used for ranking clustering methods, and vice versa.

Choice for Φ. As the clustering method is used once and only once during the execution of the evaluation, we know that if the performances P_1 and P_2 are achievable by some clustering methods $\mathcal{C}_1$ and $\mathcal{C}_2$, then any performance $\overline{P} = \lambda_1 P_1 + \lambda_2 P_2$ (with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_1 + \lambda_2 = 1$) is achievable by a clustering method $\overline{\mathcal{C}}$ obtained by a non-deterministic combination of $\mathcal{C}_1$ and $\mathcal{C}_2$ that chooses them with respective probabilities λ_1 and λ_2. Thus, $\Phi = \mathrm{conv}$ makes sense, and all ranking scores can be used to rank clustering methods.

Choice for I. When $\Phi = \mathrm{conv}$, we have demonstrated that all rankings induced by the ranking scores R_I satisfy all our three axioms. This leaves a great flexibility for the users to fine-tune the ranking w.r.t. their application-specific preferences through the random variable I. As we have seen, the only constraints are that $I \neq 0$ and $I(\omega) \geq 0 \, \forall \omega \in \Omega$.

Note about Fowlkes-Mallows index. The score FMI known as *Fowlkes-Mallows index* [9] and *cosine coefficient* [1], which is commonly used for clustering methods and defined as the geometric mean of the positive predictive value PPV (*a.k.a.* precision) and true positive rate TPR (*a.k.a.* sensitivity and recall), does not satisfy our 3rd axiom: the performance ordering induced by FMI is incompatible with $\Phi = \mathrm{conv}$. More precisely, the clustering method $\mathcal{C}$ obtained by randomly choosing between some methods $\mathcal{C}_1$ or $\mathcal{C}_2$ can be such that $FMI(\mathrm{eval}(\mathcal{C})) < \min(FMI(\mathrm{eval}(\mathcal{C}_1)), FMI(\mathrm{eval}(\mathcal{C}_2)))$, while it makes no sense to say that $\mathcal{C}$ can be worse than $\mathcal{C}_1$ or $\mathcal{C}_2$. From this perspective, it is advisable to use any ranking score instead of FMI, for example, those in green in Tab. 1.

A.2.6 Ranking (with a Note about Kendall's τ)

Let us consider the following thought experiment to evaluate ranking methods. We denote by $\mathbb{E}$ the set of elements that these methods have to rank. For the sake of simplicity, we prefer to deal only with the case with no tie hereafter.

Random Experiment 6. *(1) Apply both the* ranking method $\mathcal{R}$ *and the* oracle $\mathcal{O}$ *on* $\mathbb{E}$ *to obtain, respectively, the predicted and ground-truth sequences of elements. (2) Randomly draw two distinct elements,* ϵ_1 *and* ϵ_2*, from* $\mathbb{E}$*. (3) Four cases can* *occur depending on whether ϵ_1 is before or after ϵ_2 in the predicted sequence and whether ϵ_1 is before or after ϵ_2 in the ground-truth sequence. Nevertheless, two outcomes are enough: choose $\smiley$ if ϵ_1 and ϵ_2 appear in the same order in both sequences, $\frownie$ otherwise.*

Choice for Ω and $\mathbb{P}_{(\Omega,\Sigma)}$. Our theory applies with $\Omega = \{\smiley, \frownie\}$. By definition, the function eval gives, for any ranking method (the evaluated entity), the distribution of outcomes resulting from this random experiment. It is the performance $P_{\mathcal{R}} = \mathrm{eval}(\mathcal{R})$ of $\mathcal{R}$. The probability of drawing a *discordant pair* is given by $P(\{\frownie\})$, and the probability of drawing a *concordant pair* is given by $P(\{\smiley\})$.

Choice for S. Clearly, $S(\frownie) < S(\smiley)$ is wanted.

Choice for Φ. As the ranking method is used once and only once during the execution of the evaluation, we know that if the performances P_1 and P_2 are achievable by some ranking methods $\mathcal{R}_1$ and $\mathcal{R}_2$, then any performance $\overline{P} = \lambda_1 P_1 + \lambda_2 P_2$ (with $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, and $\lambda_1 + \lambda_2 = 1$) is achievable by a ranking method $\overline{\mathcal{R}}$ obtained by a non-deterministic combination of $\mathcal{R}_1$ and $\mathcal{R}_2$ that chooses them with respective probabilities λ_1 and λ_2. Thus, $\Phi = \mathrm{conv}$ makes sense, and all ranking scores can be used to rank ranking methods.

Choice for I. Because $|\Omega| = 2$ and $S(\frownie) \neq S(\smiley)$, we are in a particular case in which all ranking scores rank the ranking methods in the same way. From this point of view, fine-tuning I is useless.

Note about Kendall's τ. When $S(\frownie) = -1$ and $S(\smiley) = 1$, the expected value of the satisfaction is given by $X_S^E(P) = 1 - 2P(\{\frownie\}) = P(\{\smiley\}) - P(\{\frownie\}) = \tau(P)$. In other words, with the task corresponding to this choice for the satisfaction, Kendall's correlation coefficient τ [13] is the ranking score corresponding to uniform importance values.

A.3. Supplementary Material about Sec. 3.2

This section is devoted to reminders about the order theory.

A.3.1 Reminders of Classical Definitions.

Let $\mathcal{R}$ be a homogeneous binary relation on $\mathbb{P}_{(\Omega,\Sigma)}$. It is said:
- *reflexive iif* $P\mathcal{R}P\,\forall P$;
- *irreflexive iif* $\nexists P : P\mathcal{R}P$;
- *transitive iif* $P_1\mathcal{R}P_2 \wedge P_2\mathcal{R}P_3 \Rightarrow P_1\mathcal{R}P_3\,\forall P_1, P_2, P_3$;
- *symmetric iif* $P_1\mathcal{R}P_2 \Leftrightarrow P_2\mathcal{R}P_1\,\forall P_1, P_2$;
- *asymmetric iif* $\nexists(P_1, P_2) : P_1\mathcal{R}P_2 \wedge P_2\mathcal{R}P_1$;
- and *antisymmetric iif* $P_1\mathcal{R}P_2 \wedge P_2\mathcal{R}P_1 \Rightarrow P_1 = P_2$.

Two homogeneous binary relations $\mathcal{R}_a$ and $\mathcal{R}_b$ on $\mathbb{P}_{(\Omega,\Sigma)}$ are said *converse iif* $P_1\mathcal{R}_aP_2 \Leftrightarrow P_2\mathcal{R}_bP_1\,\forall P_1, P_2$.

A relation $\mathcal{R}$ is:
- an *equivalence iif* it is reflexive, transitive, and symmetric;
- a *preorder iif* it is reflexive and transitive;
- and an *order iif* it is reflexive, transitive, and antisymmetric.

An order $\mathcal{R}$ is said *total iif* $\nexists(P_1, P_2) : P_1 \not\mathcal{R} P_2 \wedge P_2 \not\mathcal{R} P_1$. It is said *partial* otherwise.

A.3.2 The 4 Cases in the Comparison of Two Performances with a Preorder $\lesssim$.

Let us now consider a preorder $\lesssim$ and derive the homogeneous binary relations $\sim$, $>$, $<$, $\not\lesseqqgtr$ as follows:

$$P_1 \sim P_2 \Leftrightarrow P_1 \lesssim P_2 \wedge P_2 \lesssim P_1 \tag{1}$$
$$P_1 > P_2 \Leftrightarrow P_1 \not\lesssim P_2 \wedge P_2 \lesssim P_1 \tag{2}$$
$$P_1 < P_2 \Leftrightarrow P_1 \lesssim P_2 \wedge P_2 \not\lesssim P_1 \tag{3}$$
$$P_1 \not\lesseqqgtr P_2 \Leftrightarrow P_1 \not\lesssim P_2 \wedge P_2 \not\lesssim P_1\,. \tag{4}$$

Indeed, we have:

$$P_1 \lesssim P_2 \Leftrightarrow P_1 < P_2 \vee P_1 \sim P_2\,. \tag{5}$$

Similarly, one can derive other binary relations taking unions of $\sim$, $>$, $<$, or $\not\lesseqqgtr$. For example,

$$P_1 \gtrsim P_2 \Leftrightarrow P_1 > P_2 \vee P_1 \sim P_2\,. \tag{6}$$

A.3.3 Implications of the Transitivity of $\lesssim$.

We can easily check, for each $\mathcal{R}_{ab} \in \{\lesssim, \not\lesssim\}$, each $\mathcal{R}_{ba} \in \{\lesssim, \not\lesssim\}$, each $\mathcal{R}_{bc} \in \{\lesssim, \not\lesssim\}$, each $\mathcal{R}_{cb} \in \{\lesssim, \not\lesssim\}$, each $\mathcal{R}_{ca} \in \{\lesssim, \not\lesssim\}$, and each $\mathcal{R}_{ac} \in \{\lesssim, \not\lesssim\}$, if there exists (P_a, P_b, P_c) such that $P_a\mathcal{R}_{ab}P_b$, $P_b\mathcal{R}_{ba}P_a$, $P_b\mathcal{R}_{bc}P_c$, $P_c\mathcal{R}_{cb}P_b$, $P_c\mathcal{R}_{ca}P_a$, and $P_a\mathcal{R}_{ac}P_c$. Because of the assumed transitivity of $\lesssim$, there are only 29 possible cases out of the 2^6:

1. $P_a \not\lesseqqgtr P_b$, $P_b \not\lesseqqgtr P_c$, $P_a \not\lesseqqgtr P_c$
2. $P_a \not\lesseqqgtr P_b$, $P_b \not\lesseqqgtr P_c$, $P_a > P_c$
3. $P_a \not\lesseqqgtr P_b$, $P_b \not\lesseqqgtr P_c$, $P_a < P_c$
4. $P_a \not\lesseqqgtr P_b$, $P_b \not\lesseqqgtr P_c$, $P_a \sim P_c$
5. $P_a \not\lesseqqgtr P_b$, $P_b > P_c$, $P_a \not\lesseqqgtr P_c$
6. $P_a \not\lesseqqgtr P_b$, $P_b > P_c$, $P_a > P_c$
7. $P_a \not\lesseqqgtr P_b$, $P_b < P_c$, $P_a \not\lesseqqgtr P_c$
8. $P_a \not\lesseqqgtr P_b$, $P_b < P_c$, $P_a < P_c$
9. $P_a \not\lesseqqgtr P_b$, $P_b \sim P_c$, $P_a \not\lesseqqgtr P_c$
10. $P_a > P_b$, $P_b \not\lesseqqgtr P_c$, $P_a \not\lesseqqgtr P_c$
11. $P_a > P_b$, $P_b \not\lesseqqgtr P_c$, $P_a > P_c$
12. $P_a > P_b$, $P_b > P_c$, $P_a > P_c$
13. $P_a > P_b$, $P_b < P_c$, $P_a \not\lesseqqgtr P_c$
14. $P_a > P_b$, $P_b < P_c$, $P_a > P_c$
15. $P_a > P_b$, $P_b < P_c$, $P_a < P_c$

16. $P_a > P_b, P_b < P_c, P_a \sim P_c$
17. $P_a > P_b, P_b \sim P_c, P_a > P_c$
18. $P_a < P_b, P_b \not\lesseqgtr P_c, P_a \not\lesseqgtr P_c$
19. $P_a < P_b, P_b \not\lesseqgtr P_c, P_a < P_c$
20. $P_a < P_b, P_b > P_c, P_a \not\lesseqgtr P_c$
21. $P_a < P_b, P_b > P_c, P_a > P_c$
22. $P_a < P_b, P_b > P_c, P_a < P_c$
23. $P_a < P_b, P_b > P_c, P_a \sim P_c$
24. $P_a < P_b, P_b < P_c, P_a < P_c$
25. $P_a < P_b, P_b \sim P_c, P_a < P_c$
26. $P_a \sim P_b, P_b \not\lesseqgtr P_c, P_a \not\lesseqgtr P_c$
27. $P_a \sim P_b, P_b > P_c, P_a > P_c$
28. $P_a \sim P_b, P_b < P_c, P_a < P_c$
29. $P_a \sim P_b, P_b \sim P_c, P_a \sim P_c$

From this list, we can derive some rules for manipulating the binary relations $\sim$, $>$, $<$, and $\not\lesseqgtr$. First, we can see that $\sim$, $>$, and $<$ are transitive:

$$P_1 \sim P_2 \wedge P_2 \sim P_3 \Rightarrow P_1 \sim P_3 \tag{7}$$
$$P_1 > P_2 \wedge P_2 > P_3 \Rightarrow P_1 > P_3 \tag{8}$$
$$P_1 < P_2 \wedge P_2 < P_3 \Rightarrow P_1 < P_3\,. \tag{9}$$

Second, we can also see how $\sim$ can be combined with the other 3 relations:

$$(P_1 \sim P_2 \wedge P_2 > P_3) \vee (P_1 > P_2 \wedge P_2 \sim P_3) \Rightarrow P_1 > P_3 \tag{10}$$
$$(P_1 \sim P_2 \wedge P_2 < P_3) \vee (P_1 < P_2 \wedge P_2 \sim P_3) \Rightarrow P_1 < P_3 \tag{11}$$
$$(P_1 \sim P_2 \wedge P_2 \not\lesseqgtr P_3) \vee (P_1 \not\lesseqgtr P_2 \wedge P_2 \sim P_3) \Rightarrow P_1 \not\lesseqgtr P_3\,. \tag{12}$$

And, third, we can see how $\not\lesseqgtr$ can be combined with $>$ and $<$:

$$(P_1 \not\lesseqgtr P_2 \wedge P_2 > P_3) \vee (P_1 > P_2 \wedge P_2 \not\lesseqgtr P_3) \Rightarrow P_1 > P_3 \vee P_1 \not\lesseqgtr P_3 \tag{13}$$
$$(P_1 \not\lesseqgtr P_2 \wedge P_2 < P_3) \vee (P_1 < P_2 \wedge P_2 \not\lesseqgtr P_3) \Rightarrow P_1 < P_3 \vee P_1 \not\lesseqgtr P_3\,. \tag{14}$$

A.3.4 Properties of $\sim$, $>$, $<$, $\not\lesseqgtr$, $\lesssim$, and $\gtrsim$.

Lemma 1. *When $\lesssim$ is a preorder, $\sim$ is reflexive.*

Proof. This results from the reflexivity of $\lesssim$ and from Eq. (1): $P \sim P \Leftrightarrow P \lesssim P \wedge P \lesssim P \Leftrightarrow true$. □

Lemma 2. *When $\lesssim$ is a preorder, $\sim$ in transitive.*

Proof. This results from the transitivity of $\lesssim$ (*cf*. Eq. (7)). □

Lemma 3. *When $\lesssim$ is a preorder, $\sim$ is symmetric.*

Proof. This results from the fact that the conjunction is symmetric and from Eq. (1): $P_1 \sim P_2 \Leftrightarrow P_1 \lesssim P_2 \wedge P_2 \lesssim P_1 \Leftrightarrow P_2 \lesssim P_1 \wedge P_1 \lesssim P_2 \Leftrightarrow P_2 \sim P_1$. □

Lemma 4. *When $\lesssim$ is a preorder, $>$ and $<$ are converse.*

Proof. This results from the fact that the conjunction is symmetric and from Eqs. (2) and (3): $P_1 > P_2 \Leftrightarrow P_1 \not\lesssim P_2 \wedge P_2 \lesssim P_1 \Leftrightarrow P_2 \lesssim P_1 \wedge P_1 \not\lesssim P_2 \Leftrightarrow P_2 < P_1$. □

Lemma 5. *When $\lesssim$ is a preorder, $>$ and $<$ are irreflexive.*

Proof. For $>$, this results from the reflexivity of $\lesssim$ and from Eq. (2): $P > P \Leftrightarrow P \not\lesssim P \wedge P \lesssim P \Leftrightarrow false \wedge true = false$. For $<$, the proof is similar. □

Lemma 6. *When $\lesssim$ is a preorder, $>$ and $<$ are asymmetric.*

Proof. For $>$, this is because $P_1 > P_2 \wedge P_2 > P_1 \Leftrightarrow (P_1 \not\lesssim P_2 \wedge P_2 \lesssim P_1) \wedge (P_2 \not\lesssim P_1 \wedge P_1 \lesssim P_2) \Leftrightarrow (P_1 \not\lesssim P_2 \wedge P_1 \lesssim P_2) \wedge (P_2 \not\lesssim P_1 \wedge P_2 \lesssim P_1) \Leftrightarrow false \wedge false \Leftrightarrow false$. For $<$, the proof is similar. □

Lemma 7. *When $\lesssim$ is a preorder, $>$ and $<$ are transitive.*

Proof. This results from the transitivity of $\lesssim$ (*cf*. Eqs. (8) and (9)). □

Lemma 8. *When $\lesssim$ is a preorder, $\not\lesseqqgtr$ is irreflexive.*

Proof. This results from the reflexivity of $\lesssim$ and from Eq. (4): $P \not\lesseqqgtr P \Leftrightarrow P \not\lesssim P \wedge P \not\lesssim P \Leftrightarrow false \wedge false \Leftrightarrow false$. □

Lemma 9. *When $\lesssim$ is a preorder, $\not\lesseqqgtr$ is symmetric.*

Proof. This results from the fact that the conjunction is symmetric and from Eq. (4): $P_1 \not\lesseqqgtr P_2 \Leftrightarrow P_1 \not\lesssim P_2 \wedge P_2 \not\lesssim P_1 \Leftrightarrow P_2 \not\lesssim P_1 \wedge P_1 \not\lesssim P_2 \Leftrightarrow P_2 \not\lesseqqgtr P_1$. □

Lemma 10. *When $\lesssim$ is a preorder, $\lesssim$ and $\gtrsim$ are converse.*

Proof. From Eqs. (5) and (6), as $>$ and $<$ are converse, we have $P_1 \gtrsim P_2 \Leftrightarrow P_1 > P_2 \vee P_1 \sim P_2 \Leftrightarrow P_2 < P_1 \vee P_2 \sim P_1 \Leftrightarrow P_2 \lesssim P_1$. □

Lemma 11. *When $\lesssim$ is a preorder, $\lesssim$ and $\gtrsim$ are reflexive.*

Proof. For $\lesssim$, it is by definition of preorders. For $\gtrsim$, from Eqs. (1), (2) and (6), we have $P \gtrsim P \Leftrightarrow P > P \vee P \sim P \Leftrightarrow (P \not\lesssim P \wedge P \lesssim P) \vee (P \lesssim P \wedge P \lesssim P) \Leftrightarrow (false \wedge true) \vee (true \wedge true) \Leftrightarrow true$. □

Lemma 12. *When $\lesssim$ is a preorder, $\lesssim$ and $\gtrsim$ are transitive.*

Proof. For $\lesssim$, it is by definition of preorders. For $\gtrsim$, as $\lesssim$ and $\gtrsim$ are converse, $P_1 \gtrsim P_2 \wedge P_2 \gtrsim P_3 \Leftrightarrow P_3 \lesssim P_2 \wedge P_2 \lesssim P_1 \Rightarrow P_3 \lesssim P_1 \Leftrightarrow P_1 \gtrsim P_3$. □

A.4. Supplementary Material about Sec. 3.6

A.4.1 The Visual Inspection of Formulas, to Determine the Importance given by Scores, can be Misleading!

Let us consider the example of two-class classification, with $P(\{tn\})$, $P(\{fp\})$, $P(\{fn\})$, and $P(\{tp\})$ denoting, respectively, the probability (or proportion) of true negatives, false positives, false negatives, and true positives. Here are two classical scores, the accuracy and the true positive rate:

$$A = P(\{tn\}) + P(\{tp\}) \qquad\qquad TPR = \frac{P(\{tp\})}{P(\{fn\}) + P(\{tp\})}$$

The formula for the accuracy gives the illusion that the same importance is given to $\{tn\}$ and $\{tp\}$ and that no importance at all is given to $\{fp\}$ and $\{fn\}$. For the true positive rate, the formula might give the impression that the same importance is given to $\{fn\}$ and $\{tp\}$ and that no importance at all is given to $\{tn\}$ and $\{fp\}$. In fact, the visual inspection of formulas like these is not reliable at all to judge the importance given by a score to the various events. To see it, consider rewriting the previous equations as

$$A = (1-\alpha)P(\{tn\}) - \alpha P(\{fp\}) - \alpha P(\{fn\}) + (1-\alpha)P(\{tp\}) + \alpha \qquad \forall \alpha$$

$$TPR = \frac{-\alpha P(\{tn\}) - \alpha P(\{fp\}) - \alpha P(\{fn\}) + (1-\alpha)P(\{tp\}) + \alpha}{-\beta P(\{tn\}) - \beta P(\{fp\}) + (1-\beta)P(\{fn\}) + (1-\beta)P(\{tp\}) + \beta} \qquad \forall \alpha, \beta$$

A visual inspection of such formulas would lead, indeed, to other illusions about the events that have no importance. This observation, however, should not stop us from thinking in terms of importance. In fact, we do it in this paper, but we do it in a mathematical framework that allows to do it rigorously.

A.4.2 So, How can we Determine the Importance given by Scores?

One cannot determine the application-specific preferences implicitly considered by any score X. However, one can determine those that are consistent, or have the best consistency, with X. We detail these notions hereafter.

For a given set of performances. Consider a score X and a ranking score R_I. If, for a given set $\Pi \subseteq \mathbb{P}_{(\Omega,\Sigma)}$, the scores X and R_I are linked by a strict monotonic relationship on $\Pi \cap \mathrm{dom}(X) \cap \mathrm{dom}(R_I)$ and if $\Pi \cap \mathrm{dom}(X) = \Pi \cap \mathrm{dom}(R_I)$, then the performance orderings $\lesssim_X$ and $\lesssim_{R_I}$ (induced by X and R_I in the way specified in Theorem 1) are identical on Π. We say that the score X *is consistent with* the application-specific preferences I, on this set. A score can be consistent with different importance values (*e.g.*, as consequence of Properties 3 and 4).

For a given distribution of performances. Let $\mathring{\mathbb{P}} = \{P \in \mathbb{P}_{(\Omega,\Sigma)} : P(\{\omega\}) > 0 \forall \omega \in \Omega)\}$. All ranking scores are defined on this set. Consider a score X and the set $\Pi = \mathrm{dom}(X) \cap \mathring{\mathbb{P}}$. We say that the score X *has the best consistency with* the application-specific preferences I when I maximizes the rank correlation between X and R_I, on Π, for the given distribution of performances. See Appendix A.7.2 for computational details.

A.5. Supplementary Material about Sec. 4.2

A.5.1 Proof of Theorem 1.

For convenience, we provide a reminder of Theorem 1 and Axiom 1 below.

Theorem 1 (Sufficient condition for Axiom 1)**.** *A binary relation $\lesssim_X$ on $\mathbb{P}_{(\Omega,\Sigma)}$ induced by a score X as $P_1 \lesssim_X P_2$* iif *either $P_1 = P_2$ or $P_1 \in \mathrm{dom}(X)$ and $P_2 \in \mathrm{dom}(X)$ and $X(P_1) \leq X(P_2)$, is a preorder satisfying Axiom 1.*

Axiom 1. *The ranking function $\mathrm{rank}_{\mathbb{E}} : \mathbb{E} \rightarrow [1, |\mathbb{E}|] : \epsilon \mapsto \mathrm{rank}_{\mathbb{E}}(\epsilon)$ satisfies $|\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) < \mathrm{eval}(\epsilon')\}| + 1 \leq \mathrm{rank}_{\mathbb{E}}(\epsilon) \leq |\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) \lesssim \mathrm{eval}(\epsilon')\}|$, where $\lesssim$ is a preorder on $\mathbb{P}_{(\Omega,\Sigma)}$.*

Proof. To establish that $\lesssim_X$ is a preorder, we have to show that it is (1) reflexive and (2) transitive.

(1) The reflexivity of $\lesssim_X$ is trivial to establish, since $P_1 = P_2 \Rightarrow P_1 \lesssim_X P_2$.

(2) The transitivity of $\lesssim_X$ can be shown as follows. $P_1 \lesssim_X P_2 \wedge P_2 \lesssim_X P_3$ implies that:
- either $P_1 \in \mathrm{dom}(X)$, $P_2 \in \mathrm{dom}(X)$, $P_3 \in \mathrm{dom}(X)$, and $X(P_1) \leq X(P_2) \wedge X(P_2) \leq X(P_3) \Rightarrow X(P_1) \leq X(P_3) \Rightarrow P_1 \lesssim_X P_3$;
- or $P_1 \notin \mathrm{dom}(X)$, $P_2 \notin \mathrm{dom}(X)$, $P_3 \notin \mathrm{dom}(X)$, and $P_1 = P_2 \wedge P_2 = P_3 \Rightarrow P_1 = P_3 \Rightarrow P_1 \lesssim_X P_3$.

We conclude that, in all cases, $P \lesssim_X P$ and $P_1 \lesssim_X P_2 \wedge P_2 \lesssim_X P_3 \Rightarrow P_1 \lesssim_X P_3$. The orderings $\lesssim_X$ induced by scores X are thus preorders. □

Summary. If the homogeneous binary relations $\sim$, $>$, $<$, and $\not\lesseqgtr$ on $\mathbb{P}_{(\Omega,\Sigma)}$ are derived from the ordering $\lesssim_X$ as explained above, and if the $\lesssim_X$ is derived from the score X, then the comparison between performances P_1 and P_2 can be summarized as follows.

	$P_1 \in \mathrm{dom}(X)$	$P_1 \notin \mathrm{dom}(X)$
$P_2 \in \mathrm{dom}(X)$	$X(P_1) < X(P_2) \Leftrightarrow P_1 < P_2$ $X(P_1) = X(P_2) \Leftrightarrow P_1 \sim P_2$ $X(P_1) > X(P_2) \Leftrightarrow P_1 > P_2$	$P_1 \not\lesseqgtr P_2$
$P_2 \notin \mathrm{dom}(X)$	$P_1 \not\lesseqgtr P_2$	$P_1 = P_2 \Leftrightarrow P_1 \sim P_2$ $P_1 \neq P_2 \Leftrightarrow P_1 \not\lesseqgtr P_2$

A.5.2 Proof of Theorem 2

For convenience, we provide a reminder of Theorem 2 and Axiom 2 below.

Theorem 2 (Sufficient condition for Axiom 2)**.** *If a score X satisfies $\min_{\omega \in E} S(\omega) \leq X(P) \leq \max_{\omega \in E} S(\omega)$ for all events $E \in \Sigma$ and all performances $P \in \mathrm{dom}(X)$ such that $P(E) = 1$, then the ordering $\lesssim_X$ satisfies Axiom 2.*

Axiom 2. *For $P_1, P_2 \in \mathbb{P}_{(\Omega,\Sigma)}$ such that $P_1(S \leq s) = 1$ and $P_2(S \geq s) = 1$ for some s, then $P_1 \lesssim P_2$ or $P_1 \not\lesseqgtr P_2$.*

Proof. Axiom 2 is satisfied when $P_1 \notin \mathrm{dom}(X)$ or $P_2 \notin \mathrm{dom}(X)$.
- Either $P_1 = P_2 \Leftrightarrow P_1 \sim P_2 \Rightarrow P_1 \lesssim P_2$,
- or $P_1 \neq P_2 \Leftrightarrow P_1 \not\lesseqgtr P_2$.

Axiom 2 is also satisfied when $P_1 \in \mathrm{dom}(X)$ and $P_2 \in \mathrm{dom}(X)$.
- On the one hand, the axiom stipulates that the event $E_1 = \{\omega \in \Omega : S(\omega) \leq s\}$ and the performance P_1 are such that $P_1(E_1) = 1$. Trivially, we have $\max_{\omega \in E_1} S(\omega) \leq s$. On the other hand, the theorem stipulates that, as $P_1(E_1) = 1$, $X(P_1) \leq \max_{\omega \in E_1} S(\omega)$. Putting all together, we have $X(P_1) \leq s$.
- On the one hand, the axiom stipulates that the event $E_2 = \{\omega \in \Omega : S(\omega) \geq s\}$ and the performance P_2 are such that $P_2(E_2) = 1$. Trivially, we have $s \leq \min_{\omega \in E_2} S(\omega)$. On the other hand, the theorem stipulates that, as $P_2(E_2) = 1$, $\min_{\omega \in E_2} S(\omega) \leq X(P_2)$. Putting all together, we have $s \leq X(P_2)$.
- As we have established that $X(P_1) \leq s$ and $s \leq X(P_2)$, we have $X(P_1) \leq X(P_2) \Leftrightarrow P_1 \lesssim P_2$.

□

A.5.3 Proof of Theorem 3

For convenience, we provide a reminder of Theorem 3 and Axiom 3 below.

Theorem 3 (Sufficient condition for Axiom 3)**.** *If a score X is such that $\Pi \subseteq \operatorname{dom}(X) \Rightarrow \Phi(\Pi) \subseteq \operatorname{dom}(X)$ and $\min_{P\in\Pi} X(P) \leq X(\overline{P}) \leq \max_{P\in\Pi} X(P)$ for all $\Pi \subseteq \operatorname{dom}(X)$ and all $\overline{P} \in \Phi(\Pi)$, then the ordering $\lesssim_X$ satisfies Axiom 3.*

Axiom 3. *Let P be a performance, and Π be a set of performances on $\mathbb{P}_{(\Omega,\Sigma)}$ such that $P' \lesssim P \vee P \lesssim P' \, \forall P' \in \Pi$.*
- $P' \lesssim P \, \forall P' \in \Pi \Rightarrow \overline{P} \lesssim P \, \forall \overline{P} \in \Phi(\Pi)$*;*
- $P' \not\lesssim P \, \forall P' \in \Pi \Rightarrow \overline{P} \not\lesssim P \, \forall \overline{P} \in \Phi(\Pi)$*;*
- $P \lesssim P' \, \forall P' \in \Pi \Rightarrow P \lesssim \overline{P} \, \forall \overline{P} \in \Phi(\Pi)$*;*
- *and* $P \not\lesssim P' \, \forall P' \in \Pi \Rightarrow P \not\lesssim \overline{P} \, \forall \overline{P} \in \Phi(\Pi)$.

Proof. We take $\lesssim = \lesssim_X$ and $\Pi \neq \emptyset$.
Remainder of the conditions. The first condition of Theorem 3 is:

$$\Pi \subseteq \operatorname{dom}(X) \Rightarrow \Phi(\Pi) \subseteq \operatorname{dom}(X)\,. \tag{15}$$

The second condition of Theorem 3 is:

$$\min_{P\in\Pi} X(P) \leq X(\overline{P}) \leq \max_{P\in\Pi} X(P) \qquad \forall \, \Pi \subseteq \operatorname{dom}(X) \qquad \forall \, \overline{P} \in \Phi(\Pi)\,. \tag{16}$$

The condition of Axiom 3 is that P is comparable to all performances in the set Π:

$$P' \lesssim P \vee P \lesssim P' \qquad \forall \, P' \in \Pi\,. \tag{17}$$

On the domain of X. By Theorem 1, this last condition implies that

$$P \in \operatorname{dom}(X)\,, \tag{18}$$

and

$$\Pi \subseteq \operatorname{dom}(X)\,. \tag{19}$$

Taking Eq. (15) and Eq. (19) together, we have

$$\Phi(\Pi) \subseteq \operatorname{dom}(X) \qquad \Leftrightarrow \qquad \overline{P} \in \operatorname{dom}(X) \qquad \forall \, \overline{P} \in \Phi(\Pi)\,. \tag{20}$$

Proof that $P' \lesssim P \, \forall P' \in \Pi \Rightarrow \overline{P} \lesssim P \, \forall \overline{P} \in \Phi(\Pi)$. On the one hand, we have, by Theorem 1,

$$\begin{aligned} P' \lesssim P \, \forall P' \in \Pi \Leftrightarrow & X(P') \leq X(P) \qquad \forall \, P' \in \Pi \\ \Leftrightarrow & \max_{P'\in\Pi} X(P') \leq X(P)\,. \end{aligned}$$

On the other hand, Eq. (16) implies that

$$X(\overline{P}) \leq \max_{P'\in\Pi} X(P') \qquad \forall \, \overline{P} \in \Phi(\Pi)\,.$$

Considering the last two equations together, we obtain

$$\begin{aligned} & X(\overline{P}) \leq \max_{P'\in\Pi} X(P') \leq X(P) \qquad \forall \, \overline{P} \in \Phi(\Pi) \\ \Rightarrow & X(\overline{P}) \leq X(P) \qquad \forall \, \overline{P} \in \Phi(\Pi)\,. \end{aligned}$$

By Theorem 1, we have thus $\overline{P} \lesssim P$.
Proof that $P' \not\lesssim P \, \forall P' \in \Pi \Rightarrow \overline{P} \not\lesssim P \, \forall \overline{P} \in \Phi(\Pi)$. On the one hand, we have, by Eq. (17) and Theorem 1,

$$\begin{aligned} P' \not\lesssim P \, \forall \, P' \in \Pi \Rightarrow & P < P' \, \forall \, P' \in \Pi \\ \Leftrightarrow & X(P) < X(P') \qquad \forall \, P' \in \Pi \\ \Leftrightarrow & X(P) < \min_{P'\in\Pi} X(P') \end{aligned}$$

On the other hand, Eq. (16) implies that

$$\min_{P' \in \Pi} X(P') \leq X(\overline{P}) \qquad \forall\, \overline{P} \in \Phi(\Pi)$$

Considering the last two equations together, we obtain

$$\begin{aligned} X(P) &< \min_{P' \in \Pi} X(P') \leq X(\overline{P}) \qquad \forall\, \overline{P} \in \Phi(\Pi) \\ \Rightarrow X(P) &< X(\overline{P}) \qquad \forall\, \overline{P} \in \Phi(\Pi)\,. \end{aligned}$$

By Theorem 1, we have thus $P < \overline{P}$, and by Eq. (17), $\overline{P} \not\lesssim P$.

Proof that $P \lesssim P' \,\forall P' \in \Pi \Rightarrow P \lesssim \overline{P} \,\forall \overline{P} \in \Phi(\Pi)$. On the one hand, we have, by Theorem 1,

$$\begin{aligned} P \lesssim P' \,\forall P' \in \Pi \Leftrightarrow & X(P) \leq X(P') \qquad \forall\, P' \in \Pi \\ \Leftrightarrow & X(P) \leq \min_{P' \in \Pi} X(P')\,. \end{aligned}$$

On the other hand, Eq. (16) implies that

$$\min_{P' \in \Pi} X(P') \leq X(\overline{P}) \qquad \forall\, \overline{P} \in \Phi(\Pi)\,.$$

Considering the last two equations together, we obtain

$$\begin{aligned} X(P) &\leq \min_{P' \in \Pi} X(P') \leq X(\overline{P}) \qquad \forall\, \overline{P} \in \Phi(\Pi) \\ \Rightarrow X(P) &\leq X(\overline{P}) \qquad \forall\, \overline{P} \in \Phi(\Pi)\,. \end{aligned}$$

By Theorem 1, we have thus $P \lesssim \overline{P}$.

Proof that $P \not\lesssim P' \,\forall P' \in \Pi \Rightarrow P \not\lesssim \overline{P} \,\forall \overline{P} \in \Phi(\Pi)$. On the one hand, we have, by Eq. (17) and Theorem 1,

$$\begin{aligned} P \not\lesssim P' \,\forall\, P' \in \Pi \Rightarrow & P' < P \,\forall\, P' \in \Pi \\ \Leftrightarrow & X(P') < X(P) \qquad \forall\, P' \in \Pi \\ \Leftrightarrow & \max_{P' \in \Pi} X(P') < X(P) \end{aligned}$$

On the other hand, Eq. (16) implies that

$$X(\overline{P}) \leq \max_{P' \in \Pi} X(P') \qquad \forall\, \overline{P} \in \Phi(\Pi)$$

Considering the last two equations together, we obtain

$$\begin{aligned} X(\overline{P}) &\leq \max_{P' \in \Pi} X(P') < X(P) \qquad \forall\, \overline{P} \in \Phi(\Pi) \\ \Rightarrow X(\overline{P}) &< X(P) \qquad \forall\, \overline{P} \in \Phi(\Pi)\,. \end{aligned}$$

By Theorem 1, we have thus $\overline{P} < P$, and by Eq. (17), $P \not\lesssim \overline{P}$.

□

A.6. Supplementary Material about Sec. 4.3

A.6.1 All Ranking Scores can be Used to Rank Performances (for $\Phi = \mathrm{conv}$)

To show that all ranking scores can be used to rank performances, for $\Phi = \mathrm{conv}$, we show that these scores satisfy the conditions of Theorems 1, 2, and 3.

All ranking scores satisfy the conditions of Theorem 1, and thus Axiom 1. For convenience, we provide a reminder of Theorem 1 and Axiom 1 below.

Theorem 1 (Sufficient condition for Axiom 1)**.** *A binary relation $\lesssim_X$ on $\mathbb{P}_{(\Omega,\Sigma)}$ induced by a score X as $P_1 \lesssim_X P_2$* iif *either $P_1 = P_2$ or $P_1 \in \mathrm{dom}(X)$ and $P_2 \in \mathrm{dom}(X)$ and $X(P_1) \leq X(P_2)$, is a preorder satisfying Axiom 1.*

Axiom 1. *The ranking function* $\mathrm{rank}_{\mathbb{E}} : \mathbb{E} \to [1, |\mathbb{E}|] : \epsilon \mapsto \mathrm{rank}_{\mathbb{E}}(\epsilon)$ *satisfies* $|\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) < \mathrm{eval}(\epsilon')\}| + 1 \leq \mathrm{rank}_{\mathbb{E}}(\epsilon) \leq |\{\epsilon' \in \mathbb{E} : \mathrm{eval}(\epsilon) \lesssim \mathrm{eval}(\epsilon')\}|$, *where $\lesssim$ is a preorder on $\mathbb{P}_{(\Omega,\Sigma)}$.*

Theorem 4. *All ranking scores satisfy the conditions of Theorem 1.*

Proof. For all ranking scores R_I, it is possible to induce an ordering $\lesssim_{R_I}$ satisfying the requirements of Theorem 1. □

All ranking scores satisfy the conditions of Theorem 2, and thus Axiom 2. For convenience, we provide a reminder of Theorem 2 and Axiom 2 below.

Theorem 2 (Sufficient condition for Axiom 2)**.** *If a score X satisfies $\min_{\omega \in E} S(\omega) \leq X(P) \leq \max_{\omega \in E} S(\omega)$ for all events $E \in \Sigma$ and all performances $P \in \mathrm{dom}(X)$ such that $P(E) = 1$, then the ordering $\lesssim_X$ satisfies Axiom 2.*

Axiom 2. *For $P_1, P_2 \in \mathbb{P}_{(\Omega,\Sigma)}$ such that $P_1(S \leq s) = 1$ and $P_2(S \geq s) = 1$ for some s, then $P_1 \lesssim P_2$ or $P_1 \not\lesseqqgtr P_2$.*

Theorem 5. *All ranking scores satisfy the conditions of Theorem 2.*

Proof. We take $X = R_I$. When $P(E) = 1$, we have

$$R_I(P) = \frac{\sum_{\omega \in \Omega} I(\omega) S(\omega) P(\{\omega\})}{\sum_{\omega \in \Omega} I(\omega) P(\{\omega\})} = \frac{\sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\})}{\sum_{\omega \in E} I(\omega) P(\{\omega\})}$$

with $\sum_{\omega \in \Omega} I(\omega) P(\{\omega\}) > 0$ when $P \in \mathrm{dom}(R_I)$.

- Let $M = \max_{\omega \in E} S(\omega)$. We have

$$\begin{aligned}
& S(\omega) \leq M \qquad \forall \omega \in E \\
\Leftrightarrow & S(\omega) - M \leq 0 \qquad \forall \omega \in E \\
\Rightarrow & \sum_{\omega \in E} I(\omega) \left[S(\omega) - M \right] P(\{\omega\}) \leq 0 \qquad \text{as } I(\omega) \geq 0 \text{ and } P(\{\omega\}) \geq 0 \\
\Leftrightarrow & \sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\}) \leq \sum_{\omega \in E} I(\omega) M P(\{\omega\}) \\
\Leftrightarrow & \sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\}) \leq M \underbrace{\sum_{\omega \in E} I(\omega) P(\{\omega\})}_{>0} \\
\Leftrightarrow & \frac{\sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\})}{\sum_{\omega \in E} I(\omega) P(\{\omega\})} \leq M \\
\Leftrightarrow & R_I(P) \leq M
\end{aligned}$$

- Let $m = \min_{\omega \in E} S(\omega)$. We have

$$\begin{aligned}
&S(\omega) \geq m \qquad \forall \omega \in E \\
\Leftrightarrow &S(\omega) - m \geq 0 \qquad \forall \omega \in E \\
\Rightarrow &\sum_{\omega \in E} I(\omega)\left[S(\omega) - m\right] P(\{\omega\}) \geq 0 \qquad \text{as } I(\omega) \geq 0 \text{ and } P(\{\omega\}) \geq 0 \\
\Leftrightarrow &\sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\}) \geq \sum_{\omega \in E} I(\omega) m P(\{\omega\}) \\
\Leftrightarrow &\sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\}) \geq m \underbrace{\sum_{\omega \in E} I(\omega) P(\{\omega\})}_{>0} \\
\Leftrightarrow &\frac{\sum_{\omega \in E} I(\omega) S(\omega) P(\{\omega\})}{\sum_{\omega \in E} I(\omega) P(\{\omega\})} \geq m \\
\Leftrightarrow &R_I(P) \geq m
\end{aligned}$$

- Putting all together, when $P(E) = 1$, we have $m \leq R_I(P) \leq M$, and so,

$$\min_{\omega \in E} S(\omega) \leq R_I(P) \leq \max_{\omega \in E} S(\omega)\,. \tag{21}$$

□

All ranking scores satisfy the conditions of Theorem 3, and thus Axiom 3 (for $\Phi = \text{conv}$). For convenience, we provide a reminder of Theorem 3 and Axiom 3 below.

Theorem 3 (Sufficient condition for Axiom 3)**.** *If a score X is such that $\Pi \subseteq \text{dom}(X) \Rightarrow \Phi(\Pi) \subseteq \text{dom}(X)$ and $\min_{P \in \Pi} X(P) \leq X(\overline{P}) \leq \max_{P \in \Pi} X(P)$ for all $\Pi \subseteq \text{dom}(X)$ and all $\overline{P} \in \Phi(\Pi)$, then the ordering $\lesssim_X$ satisfies Axiom 3.*

Axiom 3. *Let P be a performance, and Π be a set of performances on $\mathbb{P}_{(\Omega,\Sigma)}$ such that $P' \lesssim P \vee P \lesssim P' \, \forall P' \in \Pi$.*

- *$P' \lesssim P \, \forall P' \in \Pi \Rightarrow \overline{P} \lesssim P \, \forall \overline{P} \in \Phi(\Pi)$;*
- *$P' \not\gtrsim P \, \forall P' \in \Pi \Rightarrow \overline{P} \not\gtrsim P \, \forall \overline{P} \in \Phi(\Pi)$;*
- *$P \lesssim P' \, \forall P' \in \Pi \Rightarrow P \lesssim \overline{P} \, \forall \overline{P} \in \Phi(\Pi)$;*
- *and $P \not\gtrsim P' \, \forall P' \in \Pi \Rightarrow P \not\gtrsim \overline{P} \, \forall \overline{P} \in \Phi(\Pi)$.*

Theorem 6. *All ranking scores satisfy the conditions of Theorem 3 (for $\Phi = \text{conv}$).*

Proof. The proof is in two parts.

- First, let us show that $\Pi \subseteq \text{dom}(R_I) \Rightarrow \text{conv}(\Pi) \subseteq \text{dom}(R_I)$. For any $\overline{P} \in \text{conv}(\Pi)$ there exists a weighting function $\lambda_{\Pi,\overline{P}} : \Pi \to \mathbb{R}_{\geq 0} : P \mapsto \lambda_{\Pi,\overline{P}}(P)$ such that $\sum_{P \in \Pi} \lambda_{\Pi,\overline{P}}(P) = 1$ and $\sum_{P \in \Pi} \lambda_{\Pi,\overline{P}}(P) P = \overline{P}$. For all $\overline{P} \in \text{conv}(\Pi)$, we have:

$$\begin{aligned}
\Pi \subseteq \text{dom}(R_I) &\Leftrightarrow \sum_{\omega \in \Omega} I(\omega) P(\{\omega\}) \neq 0 \qquad \forall P \in \Pi \\
&\Rightarrow \sum_{P \in \Pi} \lambda_{\Pi,\overline{P}}(P) \sum_{\omega \in \Omega} I(\omega) P(\{\omega\}) \neq 0 \\
&\Leftrightarrow \sum_{\omega \in \Omega} I(\omega) \sum_{P \in \Pi} \lambda_{\Pi,\overline{P}} P(\{\omega\}) \neq 0 \\
&\Leftrightarrow \sum_{\omega \in \Omega} I(\omega) \overline{P}(\{\omega\}) \neq 0 \\
&\Leftrightarrow \overline{P} \in \text{dom}(R_I)\,.
\end{aligned}$$

- Second, let us show that, for all $\overline{P} \in \text{conv}(\Pi)$, $\min_{P\in\Pi} R_I(P) \leq R_I(\overline{P}) \leq \max_{P\in\Pi} R_I(P)$. Let us pose $l = \min_{P\in\Pi} R_I(P)$ and $u = \max_{P\in\Pi} R_I(P)$. We have:

$$\begin{aligned}
& l \leq R_I(P) \leq u \qquad \forall P \in \Pi \\
\Leftrightarrow & l \leq \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P(\{\omega\})} \leq u \qquad \forall P \in \Pi \\
\Leftrightarrow & l \sum_{\omega\in\Omega} I(\omega)P(\{\omega\}) \leq \sum_{\omega\in\Omega} I(\omega)S(\omega)P(\{\omega\}) \leq u \sum_{\omega\in\Omega} I(\omega)P(\{\omega\}) \qquad \forall P \in \Pi \\
\Rightarrow & l \sum_{\omega\in\Omega} I(\omega)\overline{P}(\{\omega\}) \leq \sum_{\omega\in\Omega} I(\omega)S(\omega)\overline{P}(\{\omega\}) \leq u \sum_{\omega\in\Omega} I(\omega)\overline{P}(\{\omega\}) \\
\Leftrightarrow & l \leq \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)\overline{P}(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)\overline{P}(\{\omega\})} \leq u \\
\Leftrightarrow & l \leq R_I(\overline{P}) \leq u
\end{aligned}$$

□

A.6.2 On the Properties of Ranking Scores.

Proof of Property 1. Let us demonstrate that we have $R_I(P) = X_S^E(P')$ with $P' = \text{filter}_I(P)$. For all $\omega \in \Omega$, we have:

$$P'(\{\omega\}) = \frac{P(\{\omega\})I(\omega)}{\sum_{\omega'\in\Omega} P(\{\omega'\})I(\omega')}$$

Thus,

$$\begin{aligned}
X_S^E(P') &= \sum_{\omega\in\Omega} P'(\{\omega\})S(\omega) \\
&= \sum_{\omega\in\Omega} \frac{P(\{\omega\})I(\omega)}{\sum_{\omega'\in\Omega} P(\{\omega'\})I(\omega')} S(\omega) \\
&= \frac{\sum_{\omega\in\Omega} P(\{\omega\})I(\omega)S(\omega)}{\sum_{\omega'\in\Omega} P(\{\omega'\})I(\omega')} \\
&= R_I(P)
\end{aligned}$$

□

Proof of Property 2. Let $S' = \alpha S + \beta$ with $\alpha, \beta \in \mathbb{R}$.

$$\frac{\sum_{\omega\in\Omega} P(\{\omega\})S'(\omega)I(\omega)}{\sum_{\omega\in\Omega} P(\{\omega\})I(\omega)} = \alpha \frac{\sum_{\omega\in\Omega} P(\{\omega\})S(\omega)I(\omega)}{\sum_{\omega\in\Omega} P(\{\omega\})I(\omega)} + \beta$$

□

Proof of Property 3. $R_{kI} = \frac{\sum_{\omega\in\Omega} kI(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} kI(\omega)P(\{\omega\})} = \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P(\{\omega\})} = R_I$ □

Proof of Property 4. Let us consider a binary satisfaction, that is $S(\omega) \in \{0, 1\}\, \forall \omega \in \Omega$. Let us define the events $E_0 = \{\omega \in \Omega : S(\omega) = 0\}$ and $E_1 = \{\omega \in \Omega : S(\omega) = 1\}$. If $I' = (\mathbf{1}_{S=0}\alpha_0 + \mathbf{1}_{S=1}\alpha_1)I$ with $\alpha_0 > 0$ and $\alpha_1 > 0$, then

$$\begin{aligned}
R_I(P) &= \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P(\{\omega\})} \\
&= \frac{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\}) + \sum_{\omega\in E_1} I(\omega)P(\{\omega\})}
\end{aligned}$$

and

$$\begin{aligned}
R_{I'}(P) &= \frac{\sum_{\omega\in\Omega} I'(\omega)S(\omega)P(\{\omega\})}{\sum_{\omega\in\Omega} I'(\omega)P(\{\omega\})}\\
&= \frac{\sum_{\omega\in E_1} I'(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I'(\omega)P(\{\omega\}) + \sum_{\omega\in E_1} I'(\omega)P(\{\omega\})}\\
&= \frac{\sum_{\omega\in E_1} \alpha_1 I(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} \alpha_0 I(\omega)P(\{\omega\}) + \sum_{\omega\in E_1} \alpha_1 I(\omega)P(\{\omega\})}\\
&= \frac{\alpha_1 \sum_{\omega\in E_1} I(\omega)P(\{\omega\})}{\alpha_0 \sum_{\omega\in E_0} I(\omega)P(\{\omega\}) + \alpha_1 \sum_{\omega\in E_1} I(\omega)P(\{\omega\})}
\end{aligned}$$

Thus, $R_{I'} = \frac{\alpha_1 R_I}{\alpha_0(1-R_I)\alpha_1 R_I}$ and $\frac{\partial R_{I'}}{\partial R_I} = \frac{\alpha_0\alpha_1}{(\alpha_0(1-R_I)\alpha_1 R_I)^2} > 0$. This leads immediately to the conclusion that $\lesssim_{R_{I'}} = \lesssim_{R_I}$.
□

Proof of Properties 5 and 6. Let us consider a binary satisfaction and the events $E_0 = \{\omega \in \Omega : S(\omega) = 0\}$ and $E_1 = \{\omega \in \Omega : S(\omega) = 1\}$. Let I_1 and I_2 be two random variables and $I = \lambda_1 I_1 + \lambda_2 I_2$ with $\lambda_1, \lambda_2 \in \mathbb{R}$ such that $\lambda_1 + \lambda_2 = 1$.

- When the random variables I_1 and I_2 are such that $I_1(\omega) = I_2(\omega) = I(\omega)\, \forall\omega \in E_1$, if we take $f : x \mapsto x^{-1}$,

$$\begin{aligned}
&\lambda_1 f\left(R_{I_1}(P)\right) + \lambda_2 f\left(R_{I_1}(P)\right)\\
=&\lambda_1\left(1 + \frac{\sum_{\omega\in E_0} I_1(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I_1(\omega)P(\{\omega\})}\right) + \lambda_2\left(1 + \frac{\sum_{\omega\in E_0} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I_2(\omega)P(\{\omega\})}\right)\\
=&\lambda_1\left(1 + \frac{\sum_{\omega\in E_0} I_1(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})}\right) + \lambda_2\left(1 + \frac{\sum_{\omega\in E_0} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})}\right)\\
=&(\lambda_1 + \lambda_2) + \frac{\lambda_1 \sum_{\omega\in E_0} I_1(\omega)P(\{\omega\}) + \lambda_2 \sum_{\omega\in E_0} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})}\\
=&1 + \frac{\sum_{\omega\in E_0}(\lambda_1 I_1 + \lambda_2 I_2)(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})}\\
=&1 + \frac{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})}{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})}\\
=&f\left(R_I(P)\right)
\end{aligned}$$

- When the random variables I_1 and I_2 are such that $I_1(\omega) = I_2(\omega) = I(\omega)\, \forall\omega \in E_0$, if we take $f : x \mapsto (1-x)^{-1}$,

$$\begin{aligned}
&\lambda_1 f\left(R_{I_1}(P)\right) + \lambda_2 f\left(R_{I_1}(P)\right)\\
=&\lambda_1\left(1 + \frac{\sum_{\omega\in E_1} I_1(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I_1(\omega)P(\{\omega\})}\right) + \lambda_2\left(1 + \frac{\sum_{\omega\in E_1} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I_2(\omega)P(\{\omega\})}\right)\\
=&\lambda_1\left(1 + \frac{\sum_{\omega\in E_1} I_1(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})}\right) + \lambda_2\left(1 + \frac{\sum_{\omega\in E_1} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})}\right)\\
=&(\lambda_1 + \lambda_2) + \frac{\lambda_1 \sum_{\omega\in E_1} I_1(\omega)P(\{\omega\}) + \lambda_2 \sum_{\omega\in E_1} I_2(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})}\\
=&1 + \frac{\sum_{\omega\in E_1}(\lambda_1 I_1 + \lambda_2 I_2)(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})}\\
=&1 + \frac{\sum_{\omega\in E_1} I(\omega)P(\{\omega\})}{\sum_{\omega\in E_0} I(\omega)P(\{\omega\})}\\
=&f\left(R_I(P)\right)
\end{aligned}$$

□

Proof of Property 7. Let $\mathcal{R} \in \{<, \leq, =, \geq, >\}$ and $v = R_I(P)$. For all $P' \in \mathbb{P}_{(\Omega,\Sigma)}$, we have:

$$\begin{aligned}
& R_I(P')\mathcal{R}R_I(P) && (22)\\
\Leftrightarrow & R_I(P')\mathcal{R}v && (23)\\
\Leftrightarrow & \frac{\sum_{\omega\in\Omega} I(\omega)S(\omega)P'(\{\omega\})}{\sum_{\omega\in\Omega} I(\omega)P'(\{\omega\})}\mathcal{R}v && (24)\\
\Leftrightarrow & \left[\sum_{\omega\in\Omega} I(\omega)S(\omega)P'(\{\omega\})\right]\mathcal{R}\left[v\sum_{\omega\in\Omega} I(\omega)P'(\{\omega\})\right] && (25)\\
\Leftrightarrow & \sum_{\omega\in\Omega} I(\omega)\left[S(\omega) - v\right]P'(\{\omega\})\mathcal{R}0 && (26)
\end{aligned}$$

This is either a linear equality or a linear inequality constraint. Thus,

$$\begin{aligned}
\phi_{\mathcal{R}}(P) &= \left\{P' \in \mathbb{P}_{(\Omega,\Sigma)} : R_I(P')\mathcal{R}R_I(P)\right\} && (27)\\
&= \left\{P' \in \mathbb{P}_{(\Omega,\Sigma)} : \sum_{\omega\in\Omega} I(\omega)\left[S(\omega) - v\right]P'(\{\omega\})\mathcal{R}0\right\} && (28)
\end{aligned}$$

is a convex subset of $\mathbb{P}_{(\Omega,\Sigma)}$. □

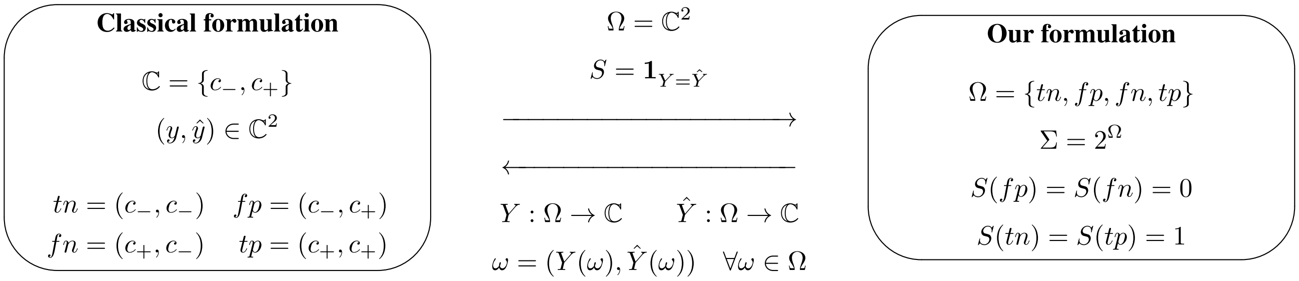
Classical formulation
$\mathbb{C} = \{c_-, c_+\}$
$(y, \hat{y}) \in \mathbb{C}^2$
$tn = (c_-, c_-)$
$fp = (c_-, c_+)$
$fn = (c_+, c_-)$
$tp = (c_+, c_+)$
$\Omega = \mathbb{C}^2$
$S = \mathbf{1}_{Y=\hat{Y}}$
$Y : \Omega \to \mathbb{C}$
$\hat{Y} : \Omega \to \mathbb{C}$
$\omega = (Y(\omega), \hat{Y}(\omega)) \quad \forall \omega \in \Omega$
Our formulation
$\Omega = \{tn, fp, fn, tp\}$
$\Sigma = 2^\Omega$
$S(fp) = S(fn) = 0$
$S(tn) = S(tp) = 1$

Figure A.7.1. Passages between two formulations (left: classical, right: ours) for the performance analysis of two-class classification problems.

A.7. Supplementary Material about Sec. 5.2

A.7.1 Link between Classical Formulation and Ours.

Fig. A.7.1 shows the connections between the classical formulation of the two-class classification task and our formulation, as explained in Sec. 5.

A.7.2 Custom Optimization Algorithm to Estimate Kendall's τ.

For any score X, our algorithm aims at determining the minimum and maximum values that a rank correlation between X and our ranking scores R_I can take over all possible importances I. Note that this algorithm is not specific to Kendall's τ [13] and could also be used with any other rank correlation, for example Spearman's ρ [23].

Variables. Leveraging Properties 3 and 4, we know that the rank-correlation between X and a ranking score R_{I_1} is equal to the rank-correlation between X and another ranking score R_{I_2} if $\frac{I_1(tp)}{I_1(tn)+I_1(tp)} = \frac{I_2(tp)}{I_2(tn)+I_2(tp)}$ and $\frac{I_1(fn)}{I_1(fp)+I_1(fn)} = \frac{I_2(fn)}{I_2(fp)+I_2(fn)}$. For this reason, we consider only two variables: $a = \frac{I(tp)}{I(tn)+I(tp)} \in [0, 1]$ and $b = \frac{I(fn)}{I(fp)+I(fn)} \in [0, 1]$.

Objective function. We optimize the function $\tau(a, b)$ that gives the rank correlation between X and R_{I^*} with $I^*(tp) = 1-a$, $I^*(tp) = 1 - b$, $I^*(tp) = b$, and $I^*(tp) = a$. In practice, this is an estimation based on a finite set of performances on which X and R_{I^*} are applied. Note that $\tau(a, b)$ is not a continuous function when estimated on a finite set of performances. The chosen optimization technique circumvents the difficulties related to that.

Optimization technique. We implemented a custom coarse-to-fine grid-based direct search [5]: we compute $\tau(a, b)$ on a coarse grid over the unit square, locate the maximum on the grid, center a smaller square and a finer grid around that point, and iterate until the square is small enough.

A.7.3 Scores Perfectly Correlated with a Ranking Score, for all Performances

- **The accuracy:** $A = R_I$ with $I(tn) = 1/2$, $I(fp) = 1/2$, $I(fn) = 1/2$, and $I(tp) = 1/2$.
- **The F-score with $\beta = 0.5$:** $F_{0.5} = R_I$ with $I(tn) = 0$, $I(fp) = 4/5$, $I(fn) = 1/5$, and $I(tp) = 1$.
- **The F-score with $\beta = 1.0$:** $F_1 = R_I$ with $I(tn) = 0$, $I(fp) = 1/2$, $I(fn) = 1/2$, and $I(tp) = 1$.
- **The F-score with $\beta = 2.0$:** $F_2 = R_I$ with $I(tn) = 0$, $I(fp) = 1/5$, $I(fn) = 4/5$, and $I(tp) = 1$.
- **The negative predictive value:** $NPV = R_I$ with $I(tn) = 1$, $I(fp) = 0$, $I(fn) = 1$, and $I(tp) = 0$.
- **The positive predictive value:** $PPV = R_I$ with $I(tn) = 0$, $I(fp) = 1$, $I(fn) = 0$, and $I(tp) = 1$.
- **The true negative rate:** $TNR = R_I$ with $I(tn) = 1$, $I(fp) = 1$, $I(fn) = 0$, and $I(tp) = 0$.
- **The true positive rate:** $TPR = R_I$ with $I(tn) = 0$, $I(fp) = 0$, $I(fn) = 1$, and $I(tp) = 1$.

A.7.4 Scores Perfectly Correlated with a Ranking Score, for the Performances Corresponding to Given Class Priors $\pi_- \neq 0$ and $\pi_+ \neq 0$

- **The balanced accuracy:** $BA = R_I$ with $I(tn) = \pi_+$, $I(fp) = \pi_+$, $I(fn) = \pi_-$, and $I(tp) = \pi_-$.
- **Cohen's kappa:** $\kappa = \frac{R_I - 2\pi_-\pi_+}{\pi_-^2+\pi_+^2}$ with $I(tn) = \frac{\pi_+^2}{\pi_-^2+\pi_+^2}$, $I(fp) = \frac{1}{2}$, $I(fn) = \frac{1}{2}$, and $I(tp) = \frac{\pi_-^2}{\pi_-^2+\pi_+^2}$. Thus, $\frac{\partial\kappa}{R_I} > 0$.

- **The informedness (*a.k.a.* Youden's J):** $J_Y = 2R_I - 1$ with $I(tn) = \pi_+$, $I(fp) = \pi_+$, $I(fn) = \pi_-$, and $I(tp) = \pi_-$. Thus, $\frac{\partial J_Y}{R_I} > 0$.
- **The negative likelihood ratio:** $NLR = \frac{1-R_I}{R_I}$ with $I(tn) = 1$, $I(fp) = 0$, $I(fn) = 1$, and $I(tp) = 0$. Thus, $\frac{\partial NLR}{R_I} < 0$.
- **The positive likelihood ratio:** $PLR = \frac{R_I}{1-R_I}$ with $I(tn) = 0$, $I(fp) = 1$, $I(fn) = 0$, and $I(tp) = 1$. Thus, $\frac{\partial PLR}{R_I} > 0$.
- **The probability of the elementary event *true negative*:** $PTN = \pi_- R_I$ with $I(tn) = 1$, $I(fp) = 1$, $I(fn) = 0$, and $I(tp) = 0$. Thus, $\frac{\partial PTN}{R_I} > 0$.
- **The probability of the elementary event *true positive*:** $PTP = \pi_+ R_I$ with $I(tn) = 0$, $I(fp) = 0$, $I(fn) = 1$, and $I(tp) = 1$. Thus, $\frac{\partial PTP}{R_I} > 0$.