Assessing binary classifiers using only positive and unlabeled data

Marc Claesen†
MARC.CLAESEN@ESAT.KULEUVEN.BE
ESAT, KU Leuven, Dept. of Computer Science, Celestijnenlaan 200A, 3001 Heverlee, Belgium

Jesse Davis
JESSE.DAVIS@CS.KULEUVEN.BE
ESAT, KU Leuven, Dept. of Computer Science, Celestijnenlaan 200A, 3001 Heverlee, Belgium

Frank De Smet
FRANK.DESMET@CM.BE
CS, KU Leuven, Dept. of Computer Science, Kasteelpark Arenberg 10, box 2446, 3001 Leuven, Belgium

Bart De Moor†
BART.DEMOOR@ESAT.KULEUVEN.BE
† KU Leuven, Dept. of Electrical Engineering; STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics; iMinds, Department of Medical IT; Kasteelpark Arenberg 10, box 2446, 3001 Leuven, Belgium

Abstract
Assessing the performance of a learned model is a crucial part of machine learning. Most evaluation metrics can only be computed with labeled data. Unfortunately, in many domains we have many more unlabeled than labeled examples. Furthermore, in some domains only positive and unlabeled examples are available, in which case most standard metrics cannot be computed at all. In this paper, we propose an approach that is able to estimate several widely used metrics including ROC and PR curves using only positive and unlabeled data. We provide theoretical bounds on the quality of our estimates. Empirically, we demonstrate that even given only a small number of positive examples and unlabeled data, we are able to reliably estimate both ROC and PR curves.

1. Introduction
Evaluating the quality of a predictive model is a critical step in the learning process. In many applications (e.g., medicine), a model must be rigorously evaluated before it can be deployed (Pencina et al., 2008; Khorana et al., 2008). Typically, evaluations often either report summary metrics, such as accuracy, F1 score, or area under the receiver operator characteristic (ROC) curve or visually show a model’s performance under different operating conditions by using ROC or precision-recall curves. The commonality between these evaluation approaches is that they all require constructing contingency tables (also called confusion matrices). A contingency table shows how a model’s predicted labels relate to an example’s ground truth label. Figure 1 depicts a contingency table for a binary classification task.

However, computing a contingency table requires labeled examples. Yet, for many problems only a few labeled examples and many unlabeled ones are available (Yarowsky, 1995; Weston et al., 2005) as acquiring labels can be time-consuming, costly, unreliable, and in some cases impossible. This setting is known as semi-supervised learning, (Blum & Mitchell, 1998; Chapelle et al., 2006; Raina et al., 2007). PU learning is a special case of semi-supervised learning in which only positive and unlabeled data is given (Liu et al., 2003; Yu et al., 2004; Denis et al., 2005; Elkan & Noto, 2008; Mordelet & Vert, 2014; Claesen et al., 2015).

The majority of the research in semi-supervised learning and PU learning focuses on developing learning algorithms that cope with partial labeling during training as opposed to evaluating algorithms when the test set is partially labeled. Typically, the empirical evaluations assume that all test labels are known (Goldman & Zhou, 2000; Nigam et al., 2000; Chawla & Karakoulas, 2005; Calvo et al., 2007; Chapelle et al., 2008; Mordelet & Vert, 2014; Claesen et al., 2015). While such an evaluation can corroborate the superiority of a new learning method, it does not provide a means for assessing the quality of a model if the test set is only partially labeled. Alternatively, in the PU setting it is sometimes assumed that all unlabeled instances are negative (Mordelet & Vert, 2011; Sifrim et al., 2013). However, we will show that this assumption could lead to a poor es-

| true label is positive | true label is negative |
|------------------------|-----------------------|
| model predicts positive | TP        | FP        |
| model predicts negative | FN        | TN        |

Figure 1. A contingency table for a binary classification problem. The entries TP, FP, TN and FN are counts of true positives, false positives, true negatives, and false negatives, respectively.
Assessing binary classifiers using only positive and unlabeled data

timate of a model’s performance in certain circumstances.

This paper describes how to incorporate the unlabeled data in the model evaluation process. We show how to compute contingency tables based on only positive and unlabeled examples, where the unlabeled set contains both positive and negative examples. Our approach is based on looking at the ranking of examples produced by a model. Concretely, we establish important relationships between contingency tables and rank distributions, which allow us to compute contingency tables based only on the positions of known positives in an overall ranking. Theoretically, we give a lower and upper bound on the false positive rate at each rank. While our analysis focuses on the PU setting, we indicate how to incorporate negative examples. Empirically, we demonstrate the efficacy of our approach by estimating ROC and PR curves from real-world data.

2. Background and definitions

We first briefly review the relevant background on model evaluation and the issues associated with partial labeling.

2.1. Rank distributions and contingency tables

We focus on binary decision problems, where the goal is to classify examples as either positive or negative. Most learned models (e.g., SVM, logistic regression, naïve Bayes) assign a numeric decision value to each example during prediction. Intuitively, the decision value represents the confidence that the instance belongs to the positive class. Typically, higher decision values imply higher confidence that the instance belongs to the positive class.

A ranking \( R \) is an ordered list of instances sorted by their corresponding decision values in decreasing order. The rank of an instance is defined by its position in \( R \), such that confident positive predictions (characterized by high decision values) are ranked close to the top of \( R \).1 Within a ranking \( R \), we treat \( P \subset R \) as the subset of examples with positive labels, and its complement \( \overline{P} = R - P \) as the subset of examples with negative labels. \(|P|\) denotes the cardinality of the set \( P \) and \( \text{rank}(R, x) \) denotes the rank of an instance \( x \) in \( R \).

Given a cutoff rank \( r \), a decision rule to obtain binary predictions is to predict the positive class for the \( r \) top ranked instances and the negative class for the rest. Based on this classifier, we can compute both the true positive rate (TPR), which is the fraction of positive examples that are correctly labeled as positive, and false positive rate (FPR), which is the fraction of negative examples that are incor-

directly labeled as positive, as:

\[
\text{TPR}(P, r) = \Pr(\text{rank}(R, x) \leq r \mid x \in P),
\]

(1)

\[
\text{FPR}(P, r) = \Pr(\text{rank}(R, x) \leq r \mid x \notin P),
\]

(2)

\[
= \frac{|\{ x \in P : \text{rank}(R, x) \leq r \}|}{|P|},
\]

\[
= \text{TPR}(R - P, r).
\]

(3)

The corresponding contingency table is straightforward to compute based on TPR(\( P, r \)) and FPR(\( P, r \)) and the number of positives \(|P|\) and negatives \(|R - P|\):

\[
\text{TP}(P, r) = \text{TPR}(P, r) \cdot |P|,
\]

(4)

\[
\text{FN}(P, r) = |P| - \text{TP}(P, r),
\]

(5)

\[
\text{FP}(P, r) = \text{FPR}(P, r) \cdot |R - P|,
\]

(6)

\[
\text{TN}(P, r) = |R - P| - \text{FP}(P, r).
\]

The rank distribution of a set of instances \( P \) within an overall ranking \( R \) is defined as the distribution of their corresponding ranks within \( R \). The rank cumulative distribution function (CDF) of a set of instances \( P \) is defined as the (empirical) CDF of their ranks, i.e. \( \forall r \in \{1, \ldots, |R|\} \):

\[
\text{F}(P, r) = \Pr(\text{rank}(R, x) \leq r \mid x \in P).
\]

(7)

The concept of rank CDF is illustrated in Figure 2. Note that \( \text{F}(P, r) \equiv \text{TPR}(P, r) \) (Eqs (1) and (7)), that is, the rank CDF of \( P \) at rank \( r \) in an overall ranking \( R \) can be interpreted directly as a true positive rate, when labeling the top \( r \) as positive.

![Figure 2. Rank CDF of two sets of positives \( P_1 = \{B, D, A, C\} \) and \( P_2 = \{E, G, F\} \) within an overall ranking \( R = \{B, E, D, G, H, A, F, C, I\} \), with \(|P_1| = 4\) and \(|P_2| = 3\).](image)

We use two convenience functions to partition sets of ranks:

\[
\text{head}(X, r) = \{ \text{rank}(R, x) \leq r : x \in X \},
\]

\[
\text{tail}(X, r) = \{ \text{rank}(R, x) > r : x \in X \},
\]

such that \( \text{head}(X, r) \cap \text{tail}(X, r) = \emptyset, \text{head}(X, r) \cup \text{tail}(X, r) = X \) and \(|\text{head}(X, r)| = \text{F}(X, r) \cdot |X|\).
2.2. ROC and PR curves

Receiver operator characteristic (ROC) curves depict how a model’s true positive rate (shown on the y-axis) varies as a function of its false positive rate (shown on the x-axis). Each cutoff rank \( r \in \{1, \ldots, |\mathcal{R}|\} \) corresponds to a single point (i.e., (FPR, TPR) pair) in ROC space (Equations (2) and (3)). An (empirical) ROC curve for a given ranking \( \mathcal{R} \) and set of positives \( \mathcal{P} \subset \mathcal{R} \) is constructed by computing \( \text{FPR}(\mathcal{P}, r) \) and \( \text{TPR}(\mathcal{P}, r) \) at each rank \( r \) and interpolating by drawing a straight line between two points corresponding to consecutive ranks. ROC curves are used extensively in machine learning for performance assessment of classifiers (Bradley, 1997). The advantage of ROC analysis is that it illustrates the performance of a model over its entire operating range.

The area under an ROC curve (AUROC) is a commonly used summary statistic which typically ranges between 0.5 (random model) and 1 (perfect model). AUROC is equal to the probability that a random positive is ranked higher than a random negative, which is equivalent to the Wilcoxon test of ranks (Hanley & McNeil, 1982). It is a popular criterion to optimize during hyperparameter search (Lessmann et al., 2008; Claesen et al., 2014).

Precision-Recall (PR) curves (Raghavan et al., 1989; Davis & Goadrich, 2006) are an alternative to ROC curves that show how a model’s precision (y-axis) varies as a function of recall (x-axis). Recall is equivalent to TPR and precision is the fraction of examples classified as positive that are truly positive (\( \text{TPR} / (\text{TP} + \text{FN}) \)). PR curves are widely used when there is a skew in the class distributions (i.e., many more negative than positive examples) (Goadrich et al., 2004; Bunescu et al., 2005; Davis & Domingos, 2009; Kok & Domingos, 2010; Claesen et al., 2015).

2.3. The implications of partial labeling

In the partial labeling setting, the overall ranking \( \mathcal{R} \) consists of disjoint sets of known positives \( \mathcal{P}_L \), known negatives \( \mathcal{N}_L \), and unlabeled instances \( \mathcal{U} \):

\[
\mathcal{R} = \mathcal{P}_L \cup \mathcal{N}_L \cup \mathcal{U},
\]

\[
\emptyset = \mathcal{P}_L \cap \mathcal{N}_L = \mathcal{P}_L \cap \mathcal{U} = \mathcal{N}_L \cap \mathcal{U}.
\]

The unlabeled set \( \mathcal{U} \) consists of latent positives \( \mathcal{P}_U \) and latent negatives. We focus on the PU learning setting, in which only positive labels are available, i.e. \( \mathcal{N}_L = \emptyset \), though our approach can include known negatives too.

Computing contingency tables and performance curves based on an overall ranking \( \mathcal{R} \) as described in Sections 2.1 and 2.2 requires a known set of all positives \( \mathcal{P}_U = \mathcal{P}_L \cup \mathcal{P}_U \). Since \( \mathcal{P}_U \) is unknown, they cannot be computed directly. Section 3 details how to tackle this problem theoretically and Section 4 describes a practical approach.

The fraction of latent positives in the unlabeled set plays a crucial role in our work, denoted by \( \beta \):

\[
\beta = \Pr(y = 1 \mid x \in \mathcal{U}) = |\mathcal{P}_U|/|\mathcal{U}|. \tag{9}
\]

We assume an estimate \( \hat{\beta} \approx \beta \) is available. This can be obtained via direct statistical estimates based on training data (Elkan & Noto, 2008) or domain knowledge, e.g. in disease screening \( \hat{\beta} \) is the prevalence of undiagnosed patients of the disease (Holmberg, 1996; Rathmann et al., 2003).

2.4. Performance curves from partially labeled data

If only a few labeled instances of both classes are available, they can be used to construct a rough estimate of either the ROC or PR curve. However, if only positive labels are available, even a rough approximation of the ROC or PR curve cannot be estimated directly.

In PU learning, a simple approach is to treat the full unlabeled set as negative, i.e., assume \( \hat{\beta} = 0 \). This approach is inherently pessimistic: at any cutoff rank, the FPR is overestimated and the TPR is underestimated because all latent positives are treated as negative. Hence, this method is only useful when \( \beta \) is small, that is, \( \beta < 0.01 \) as its bias is related to \( \beta/\hat{\beta} \). This is particularly problematic for PR curves, as we will show later.

3. The relationship between the rank CDF of positives and contingency tables

We now prove some relationships between the CDF of ranks of sets of positives within an overall ranking \( \mathcal{P}_L \subset \mathcal{R} \) at a given rank \( r \) and the corresponding contingency table. We then use these relationships to prove bounds on the FPR at a given rank \( r \) when the ranking includes unlabeled examples, some of which are latent positives.

3.1. Rank distributions and contingency tables based on given subsets of positives in an overall ranking

We begin by considering given sets of positives within an overall ranking. The implications of each lemma are visualized where appropriate.

**Lemma 1.** Given a rank \( r \) and two disjoint subsets of positives \( \mathcal{P}_1 \) and \( \mathcal{P}_2 \) within an overall ranking \( \mathcal{R} \). If \( |\mathcal{P}_1| = |\mathcal{P}_2| \), and \( \text{TPR}(\mathcal{P}_1, r) > \text{TPR}(\mathcal{P}_2, r) \), then \( \text{FPR}(\mathcal{P}_1, r) < \text{FPR}(\mathcal{P}_2, r) \).

**Proof:** The numerator of FPR is the number of false positives, this is the number of positive predictions minus the number of true positives. Via Equations (3) and (4), this is \( r \) and \( \text{TPR}(\mathcal{P}, r) \cdot |\mathcal{P}| \), respectively:

\[
\text{FPR}(\mathcal{P}, r) = \frac{r - \text{TPR}(\mathcal{P}, r) \cdot |\mathcal{P}|}{|\mathcal{R}| - |\mathcal{P}|}. \tag{10}
\]
Since |P_1| = |P_2|, the denominators of FPR(P_1, r) and FPR(P_2, r) are equal, so TPR(P_1, r) > TPR(P_2, r) ⇔ FPR(P_1, r) < FPR(P_2, r).

**Lemma 2.** Given a rank r and two disjoint sets of positives P_1 and P_2 in a ranking R and P_Ω = P_1 ∪ P_2. If TPR(P_1, r) = t_1 < TPR(P_2, r) = t_2 then:

\[
TPR(P_1, r) < TPR(P_Ω, r) < TPR(P_2, r)
\]

**Proof:** write TPR(P_Ω, r) in terms of t_1 and t_2:

\[
TPR(P_Ω, r) = \frac{t_1 |P_1| + t_2 |P_2|}{|P_1| + |P_2|},
\]

since \(t_1 < t_2\), we get \(t_1 < TPR(P_Ω, r) < t_2\).

**Corollary 1.** Given a rank r and three sets of positives P_A, P_B and P_C within a ranking R such that P_A ∩ P_B = ∅ and P_A ∩ P_C = ∅ and |P_B| = |P_C|.

\[
TPR(P_B, r) = t_B < TPR(P_C, r) = t_C \leftrightarrow TPR(P_A ∪ P_B, r) < TPR(P_A ∪ P_C, r).
\]

**Proof:** via Equation (11): all terms are equal for TPR(P_A ∪ P_B, r) and TPR(P_A ∪ P_C, r) except \(t_B < t_C\).

### 3.2. Contingency tables based on partially labeled data

Lemmas 1 and 2 describe relationships between rank distributions and contingency tables of different (but known) sets of positives within an overall ranking. However, computing contingency tables from partially labeled data requires accounting for the unknown set of latent positive examples \(P_U\) in the unlabeled data, that is, \(P_Ω = P_L ∪ P_U\), such that |P_U| > 0 and |P_U| = β · |U|. Assuming a known proportion of latent positive examples in the unlabeled data, we now show how to construct two contingency tables from partially labeled data that correspond to the greatest-lower and least-upper bound of the FPR at a specified rank.

**Theorem 1.** Given an overall ranking \(R = P_L ∪ N_L ∪ U\), consisting of disjoint sets of known positives \(P_L\), known negatives \(N_L\) and unlabeled instances \(U\), where \(U\) contains an unknown set of latent positives \(P_U\) ∋ U of known size |P_U| = β · |U|. Given a rank r and an upper bound \(T_{ub}(r) ≥ TPR(P_U, r)\), a tight lower bound on FPR(P_Ω, r) with \(P_Ω = P_L ∪ P_U\) can be found without explicitly identifying \(P_U\).

**Proof:** Step 1: assign a set of surrogate positives \(P^*_U\):

\[
P^*_U = \arg \min_{P_U ⊂ U} TPR(P^*_U, r),
\]

subject to TPR(P^*_U, r) ≥ T_{ub}(r) and |P^*_U| = β · |U|, then TPR(P^*_U, r) ≥ TPR(P_U, r) by construction. If |head(U, r)| < βT_{ub}(r) · |U|, then no \(P^*_U\) exists that satisfies the constraint TPR(P^*_U, r) ≥ T_{ub}(r) in Equation (12). In this case, define:

\[
P^*_U = \arg \max_{P_U ⊂ U} TPR(P^*_U, r),
\]

that is assign positive labels to all instances in head(U, r). This trivially implies that TPR(P^*_U, r) ≥ TPR(P_U, r).

Step 2: define \(P^*_Ω = P_L ∪ P^*_U\). Using Corollary 1 yields TPR(P^*_Ω, r) ≥ TPR(P_U, r). Since |P_Ω| = |P_Ω|, using Lemma 1 yields the lower bound on FPR:

\[
FPR(P^*_Ω, r) ≤ FPR(P_Ω, r).
\]

Applying Theorem 1 yields a nontrivial lower bound on FPR(P_Ω, r). In Lemma 3 we prove that FPR(P_Ω, r) is the greatest achievable lower bound based on a given \(U ⊂ R\).

**Lemma 3.** Minimizing TPR(P^*_Ω, r) in Equation (12) of Theorem 1 ensures FPR(P^*_Ω, r) is the greatest achievable lower bound on FPR(P_Ω, r) given β, T_{ub}(r), R and U.

**Proof (by contradiction):** suppose another set of surrogate positives \(P^*_U \subset U\) exists with |P^*_U| = β · |U|, such that P^*_U ≠ P^*_U, and TPR(P^*_U, r) ≥ T_{ub}(r) and for P^*_U = P_L ∪ P^*_U:

\[
FPR(P^*_U, r) < FPR(P^*_Ω, r) ≤ FPR(P_Ω, r).
\]

Via Corollary 1 this implies TPR(P^*_U, r) < TPR(P^*_U, r), which contradicts the definition of FPR (Eq. (12)).

Due to its symmetry, Theorem 1 can also be used to obtain the least achievable upper bound of FPR(P_Ω, r) given the overall ranking R by assigning surrogates P^*_U such that:

\[
P^*_U = \arg \max_{P_U ⊂ U} TPR(P^*_U, r),
\]

subject to TPR(P^*_U, r) ≤ T_{ub}(r) and |P^*_U| = β · |U|, where T_{ub}(r) ≤ TPR(P^*_U, r).

\(^2\)A surrogate positive is an example that we treat as if its ground truth label is positive (even though in reality its ground truth label is unknown) when constructing a contingency table.

\(^3\)An infeasibility implies that T_{ub}(r) and/or β are too high.
4. Efficiently computing the bounds

We now describe the practical details for using Theorem 1 and Lemma 3 to compute the contingency tables corresponding to the greatest lower bound and least upper bound on FPR(\(P_{\Omega}, r\)) from a finite sample. First, we propose an efficient method to compute contingency tables via Theorem 1. Second, we describe how to obtain bounds on \(T_{\text{lb}}(r)\) and \(T_{\text{ub}}(r)\).

4.1. Computing the contingency table corresponding to greatest-lower bound on FPR at given rank \(r\)

First, we discuss how to compute contingency tables for a given cutoff rank \(r\), by using Theorem 1. We focus on building the contingency table that corresponds to the lower bound on the FPR. Estimating an upper bound on the FPR is analogous.

The proof of Theorem 1 requires creating sets of surrogate positives \(P_U^r\) from the unlabeled data. However, it is not necessary to explicitly construct sets of surrogate positives because the full rank distribution of \(P_U^r\) is not needed to compute the contingency table at rank \(r\). The only requirement is that the ranks in \(P_U^r\) are distributed such that \(\text{TPR}(P_U^r, r) \geq T_{\text{ub}}(r)\) where \(T_{\text{ub}}(r)\) is an upper bound on \(\text{TPR}(P_U, r)\) (Section 4.2 describes computing \(T_{\text{ub}}(r)\)).

To construct a contingency table at a given cutoff rank \(r\), we decompose the computation to consider the labeled and unlabeled instances separately:

\[
\begin{bmatrix}
    \text{TP}'_U & \text{FP}'_U \\
    \text{FN}'_U & \text{TN}'_U
\end{bmatrix} = \begin{bmatrix}
    \text{TP}'_L & \text{FP}'_L \\
    \text{FN}'_L & \text{TN}'_L
\end{bmatrix} + \begin{bmatrix}
    \text{TP}'_U - \text{TP}'_L & \text{FP}'_U - \text{FP}'_L \\
    \text{FN}'_U - \text{FN}'_L & \text{TN}'_U - \text{TN}'_L
\end{bmatrix}.
\]

Partial contingency table based on labeled instances

For a given rank \(r\), we can easily compute partial contingency tables of the global ranking based on \(P_L\) and \(N_L\):

\[
\begin{align*}
    P_L & \rightarrow \text{TP}'_L = |\text{head}(P_L, r)|, \\
    & \quad \text{FN}'_L = |\text{tail}(P_L, r)|, \\
    N_L & \rightarrow \text{FP}'_L = |\text{head}(N_L, r)|, \\
    & \quad \text{TN}'_L = |\text{tail}(N_L, r)|.
\end{align*}
\]

Partial contingency table based on unlabeled instances

The partial contingency table from \(U\) for a given rank \(r\) can be determined based on \(T_{\text{lb}}(r), T_{\text{ub}}(r), P_L, U, \beta\). To obtain a lower bound on \(\text{FPR}(P_{\Omega}, r)\) via Theorem 1, we must compute a contingency table corresponding to surrogate positives \(P_U^r \subset U\), such that \(|P_U^r| = \beta \cdot |U|\) and \(\text{TPR}(P_U^r, r) \geq T_{\text{ub}}(r) \geq \text{TPR}(P_U, r)\).

To obtain the greatest lower bound on FPR via Theorem 1 we must compute a contingency table based on surrogate positives \(P_U^r\), where \(P_U^r\) is obtained via Equation (12). This translates into requiring \(\theta\) surrogate positives in \(\text{head}(P_U^r, r)\) and the rest in \(\text{tail}(P_U^r, r)\), where:

\[
\theta = \left[ T_{\text{lb}}(r) \cdot |P_U^r| \right] = \left[ \beta \cdot T_{\text{ub}}(r) \cdot |U| \right].
\]

By rounding up in Equation (15), we ensure that \(\text{TPR}(P_U^r, r) \geq T_{\text{ub}}(r)\) as mandated by Theorem 1.

It may occur that \(|\text{head}(U, r)| < \theta\), in which case it is impossible to assign \(\theta\) surrogates below rank \(r\) in \(U\). In this case, all of \(\text{head}(U, r)\) is assigned as a surrogate positive and the remaining surrogates are somewhere in \(\text{tail}(U, r)\) (this approach is equivalent to fulfilling Equation (13)).

The corresponding partial contingency table is:

\[
\begin{align*}
    \text{TP}'_U & = \min(|\text{head}(U, r)|, \theta), \\
    \text{FN}'_U & = |P_U^r| - \text{TP}'_U, \\
    \text{FP}'_U & = |\text{head}(U, r)| - \text{TP}'_U, \\
    \text{TN}'_U & = |U| - |P_U^r| - \text{FP}'_U.
\end{align*}
\]

To obtain the contingency table corresponding to the least upper bound on FPR(\(P_U, r\)), replace Equation (15) by:

\[
\theta = \left[ T_{\text{lb}}(r) \cdot |P_U^r| \right].
\]

4.2. Bounds on the rank distribution of \(P_U\)

In Theorem 1 we assumed a bound \(T_{\text{ub}}(r) \geq \text{TPR}(P_U, r)\) at rank \(r\) is available. To estimate a lower bound on FPR at rank \(r\), an upper bound on TPR is needed and vice versa, i.e. we need bounds \(T_{\text{lb}}(r) \leq \text{TPR}(P_U, r) \leq T_{\text{ub}}(r), \forall r\).

To obtain these bounds we assume known and latent positives have similar rank distributions (Eqs. (7) and (1)):

\[
\text{TPR}(P_L, r) = \text{TPR}(P_U, r) \pm \epsilon(r), \forall r,
\]

with small deviations \(\epsilon(r)\). This holds if the probability of the label being known is independent of the data vector, \(x \sim X^+ : x \in P_L\) and \(x \sim X^- : x \in P_U\), with \(X^+ \sim X^+\), that is when known positives are selected completely at random from \(P_O\). The transformation of input space distribution to rank distribution is fixed for a given classifier \(C\), so if the underlying input distributions are equal \((X^+_L = X^+_U)\), then \(P_L \sim C(X^+_L)\) and \(P_U \sim C(X^+_U)\), satisfying Equation (18). The assumption in Equation (18) is violated if the labeling process is biased (Chawla & Karakoulas, 2005), which may occur in certain applications (Doyle et al., 2010; Sifrim et al., 2013).

\[
\text{TPR}(P_L, r) \text{ is estimated via the empirical CDF of } P_L.
\]

Since this estimate is not a perfect estimate of the true CDF, we use confidence intervals (CIs). The assumption in Equation (18) implies that a CI of the CDF based on \(P_L\) is also a CI of the CDF of \(P_U\). A CI boundary is treated as a function mapping rank \(r\) to the estimated bound on the CDF: \(\tau(r) : \mathbb{N} \mapsto [0, 1] \). \(T_{\text{lb}}\) and \(T_{\text{ub}}\) denote the CI bounds:

\[
T_{\text{lb}}(r) \leq \text{TPR}(P_U, r) \leq T_{\text{ub}}(r) \leq 1, \forall r.
\]
We formalize the bounds of the CI of the CDF as functions of rank because an underlying set with that rank distribution does not necessarily exist in the overall ranking \( \mathcal{R} \).

We compute a standard bootstrap estimate of the confidence interval on the empirical CDF of \( \mathcal{P}_L \) with satisfactory results (narrow confidence interval, no distributional assumptions) (Efron & Tibshirani, 1994). A downside to the bootstrap is its computational demand when the number of known positives becomes large. An alternative is to construct a confidence band based on the empirical CDF via the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality (Dvoretzky et al., 1956; Massart, 1990).

5. Constructing ROC and PR curve estimates

Next, we describe how to estimate bounds on the true ROC and PR curves. While we focus our discussion on these two criteria, our approach can be used to estimate other metrics based on contingency tables such as accuracy or F1.

**ROC curves** Given a ranking, instead of constructing a single ROC curve, our approach computes two curves: one corresponding to the least upper bound and one corresponding to the greatest lower bound on rank CDF of all positives \( \mathcal{P}_\Omega \). To construct these curves, we use the methodology outlined in Section 4 to compute two contingency tables for each possible rank \( r \), corresponding to the greatest lower bound and least upper bound on FPR. The set of contingency tables corresponding to greatest lower bounds on FPR at each rank is associated to an upper bound on the ROC curve of all positives \( \mathcal{P}_\Omega \), whereas the set of contingency tables corresponding to the least upper bound on FPR corresponds to a lower bound on the ROC curve of \( \mathcal{P}_\Omega \).

It is important to understand how these estimates correspond to bounds in ROC space. By computing \( \theta \) as in Equation (15) to obtain the greatest lower bound on FPR(\( \mathcal{P}_U, r \)), the corresponding TPR is higher than TPR(\( \mathcal{P}_U, r \)). As such, the upper bound on the ROC curve is shifted upwards and to the left. Conversely, the lower bound on the ROC curve (based on the least upper bound on FPR at each rank, i.e. \( \theta \) as in Equation (17)) is shifted downward and to the right. This implies that the upper bound on the ROC curve completely dominates the curve of \( \mathcal{P}_\Omega \) and the lower bound is completely dominated by the curve of \( \mathcal{P}_\Omega \), provided that \( T_{lb}(r) \leq \text{TPR}(\mathcal{P}_U, r) \leq T_{ub}(r), \forall r \in \{1, \ldots, |\mathcal{R}|\} \).

**PR curves** Given the contingency tables used to generate the least upper bound and greatest lower bound ROC curves, it is straightforward to construct the corresponding bounds in PR space. Each contingency table contains all the required information for generating a point in PR space.

A key result relating ROC and PR curves is that one curve dominates another in ROC space if and only if it also dominates in PR space (Davis & Goadrich, 2006). Given this result, mapping the bounds we obtain for ROC curves to PR space directly yields (tight) bounds on the corresponding true PR curve. Since the upper bound in ROC space completely dominates the true curve, and the lower bound in ROC space is completely dominated by it, the same holds for the bounds on PR curves.

PR curves are very sensitive to class skew. However, the empirical class balance is contingent on the estimated fraction \( \beta \) of positives in the unlabeled part of the test set. If the uncertainty on the estimate of \( \beta \) is large, the bounds in PR space are inevitably loose (cfr. Figure 9).

6. Properties of the estimates

Next we demonstrate the convergence properties and the effect of over/underestimating \( \beta \) via ROC curves based on proposed method for estimating contingency tables.

6.1. Convergence properties

The convergence properties of our ROC estimates are contingent on those of the confidence interval of the empirical CDF, which are well studied in statistics: via the strong law of large numbers the empirical CDF \( \hat{F}_n(x) \) is a consistent pointwise estimator of the true CDF \( F(x) \). By the Glivenko-Cantelli theorem \( \hat{F}_n(x) \) converges uniformly to \( F(x) \) for increasing \( n \) (Van der Vaart, 2000).

Figure 4 shows the convergence of the bounds on the area under the curve for the estimated lower and upper bound of the ROC curve for increasing amounts of known positives in simulated rankings. The range of estimates depends on the width of the CI on rank CDF, which in turn depends on the number of known positives \( |\mathcal{P}_L| \) (higher is better) and the size of the total data set \( |\mathcal{R}| \) (lower is better).

Figure 4. The effect of \( |\mathcal{P}_L| \) on estimated AUC, with \( |U| = 100,000, N_L = 0 \) and \( \beta = \beta = 0.2 \). Lower and upper bounds on rank CDF were obtained via the bootstrap. The depicted confidence intervals are based on 200 repeated experiments.
6.2. The impact of estimating $\hat{\beta}$

Until now we assumed the exact number of latent positives $|\mathcal{P}_L|$ is known. In practice, we are often limited to an estimate of the fraction of latent positives in the unlabeled set $\hat{\beta}$, such that $\hat{\beta} \approx \beta$. If $\hat{\beta}$ is not exact, the conditions to apply Lemma 1 are violated where it is used within Theorem 1. The effects of set size on FPR is characterized in Lemma 4, which will help us understand the effect of over or under estimating $\hat{\beta}$.

**Lemma 4.** Given two sets of positive labels $\mathcal{P}_1$ and $\mathcal{P}_2$ within an overall ranking $\mathcal{R}$ and a rank $r$, such that $\text{TPR}(\mathcal{P}_1, r) = \text{TPR}(\mathcal{P}_2, r) = t$ and $|\mathcal{P}_1| > |\mathcal{P}_2|$, then:

(a) $\text{FPR}(\mathcal{P}_2, r) < t \rightarrow \text{FPR}(\mathcal{P}_1, r) < \text{FPR}(\mathcal{P}_2, r)$,

(b) $\text{FPR}(\mathcal{P}_2, r) > t \rightarrow \text{FPR}(\mathcal{P}_1, r) > \text{FPR}(\mathcal{P}_2, r)$.

(a) corresponds to a ranking and cutoff that is better than random (i.e. TPR($\mathcal{P}, r$) > FPR($\mathcal{P}, r$)), whereas (b) corresponds to a ranking and cutoff that is worse than random.

![Illustration of Lemma 4](image)

**Proof:** take the derivative of FPR to $|\mathcal{P}|$ while fixing $r$, based on Equation (10):

$$
\frac{d\text{FPR}(\mathcal{P}, r)}{d|\mathcal{P}|} = \frac{r - t \cdot |\mathcal{R}|}{(|\mathcal{R}| - |\mathcal{P}|)^2},
$$

$$
= \frac{r - t \cdot |\mathcal{P}| - t \cdot |\mathcal{R} - \mathcal{P}|}{(|\mathcal{R}| - |\mathcal{P}|)^2}.
$$

$r - t \cdot |\mathcal{P}|$ is the number of negatives in the top ranking (false positives) and $t \cdot |\mathcal{R} - \mathcal{P}|$ is the number of false positives at FPR = $t$. The derivative is negative if the FPR is below $t$ and vice versa, therefore if the ranking is better than random (TPR = $t > \text{FPR}$), increasing $|\mathcal{P}|$ leads to a lower FPR at rank $r$ and vice versa.

The impact of Lemma 4 on the proof of Theorem 1 is of high practical value. If the ranking of $\mathcal{P}_U$ is better than random, then in terms of ROC curves overestimating $\hat{\beta}$ is useful to obtain a (loose) upper bound on the curve and underestimating $\hat{\beta}$ can be used to obtain a (loose) lower bound. This also implies that treating the full unlabeled set as negative results in an overly pessimistic estimate of the ROC curve. The effect of varying $\hat{\beta}$ is shown in Figure 6.

![Figure 6](image)

**Figure 6.** The effect of $\hat{\beta}$ on estimated ROC curves, based on 2,000 known positives, 100,000 unlabeled instances and $\beta = 0.3$.

Given the relationship between ROC and PR curves as explained in Section 5, the pattern in Figure 6 translates to PR curves as well: overestimating $\hat{\beta}$ results in bounds that are too optimistic (can be used as a loose upper bound) and underestimating $\hat{\beta}$ yields bounds that are too pessimistic (can be used as a loose lower bound).

7. Results and discussion

To demonstrate the applicability and robustness of our approach, we show the resulting estimates of ROC and PR curves on two public data sets in a PU learning setting and compare our estimates to the ground truth.

We used rankings and corresponding true labels of simulations done by Claesen et al. (2015), in which classifiers were tuned and trained in a PU learning context and tested on an independent, fully labeled test set. Since all test labels are available, this enables us to compute the true curves as a reference. From this study, we used rankings on the covtype and sensit data sets (Blackard & Dean, 1999; Duarte & Hen Hu, 2004). The covtype test set consists of 20,000/20,000 positives/negatives and the sensit test set consists of 5, 250/14, 455 positives/negatives.

To produce our estimations, we discarded all negative labels and then look at removing 80%, 90%, 95%, and 99% of positive labels (at random). We focus on discussing the 95% case; the remaining results can be found in the supplementary material. This setup yields $|\mathcal{P}_L| = 1,000$ and $\beta \approx 49\%$ for covtype and $|\mathcal{P}_L| = 262$ and $\beta \approx 26\%$.

4We can assess this via the lower bound on rank CDF $T_{lb}(r)$.

5Python code to reproduce the results is available at [https://github.com/claesenm/semisup-metrics](https://github.com/claesenm/semisup-metrics).
for sensit. We always used \( \hat{\beta} = \beta \). We estimated 95% confidence intervals on the rank CDF of \( P_L \) via a standard bootstrap approach, using 2,000 resamples. Figures 7 to 9 illustrate different aspects of the estimates for both data sets: rank CDFs and ROC/PR curves.

Figure 7 shows the estimate point-wise confidence intervals on the rank distribution, based on the empirical CDF of \( P_L \). Our approach assumes these confidence intervals contain the rank CDF of \( P_U \) (cfr. Equations (18) and (19)). For reference, we show the rank distribution of the latent positives \( P_U \), which is unknown in practice. If the rank CDF of \( P_U \) is close to or outside of the estimated confidence intervals based on \( P_L \), the true ROC/PR curve will also be close to or outside the estimated bounds, respectively.

![Figure 7. Rank cumulative distribution functions.](image)

Figure 8 shows the estimated bounds on the ROC curve of \( P_{\Omega} \) along with the true curve. The estimated bounds have the same shape as the true curve and are always strict. We additionally plotted the curve that is obtained by treating the unlabeled set as negative (\( \hat{\beta} = 0 \)), which clearly yields a more pessimistic approximation than our approach.

![Figure 8. Receiver Operating Characteristic curves.](image)

Figure 9 shows the resulting PR curves for both data sets. The estimated bounds are clearly much wider in PR space than in ROC space (Figure 8). Moreover, assuming \( \hat{\beta} = 0 \) yields a very poor estimate. Finally it must be noted that estimated PR curves are very sensitive to the estimation error in \( \hat{\beta} \), because precision is directly affected by class balance. As such, we recommend using ROC curves over PR curves when only a rough estimate of \( \beta \) is available.

![Figure 9. Precision-Recall curves.](image)

### 8. Conclusion

We presented an approach to compute contingency tables corresponding to a lower and upper bound on FPR at a given rank based on a confidence interval on the rank CDF of (known) positives. Via these estimated contingency tables many commonly used performance metrics become available in a semi-supervised setting.

To demonstrate the efficacy of our approach, we computed bounds on ROC and PR curves of all positives \( P_{\Omega} = P_L \cup P_U \) in two realistic experiments. We have shown that the quality of these bounds is limited only by the quality of estimated confidence intervals on the CDF of ranks of positives and the estimated fraction of latent positives \( \hat{\beta} \). Due to the large sensitivity of the estimated bounds in PR space to \( \beta \), we recommend using ROC curves in a partial labeling context unless \( \beta \) is known to be a good estimate (\( \beta \approx \hat{\beta} \)). The proposed bounds on ROC curves enable thorough assessment of models in a semi-supervised setting and hence address an important practical problem.

### Acknowledgments

STADIUS members are supported by Flemish Government: FWO: projects: G.0871.12N (Neural circuits), IWT: TBM-Logic Insulin(100793), TBM Rectal Cancer(100783), TBM IETA(130256); PhD grants, Industrial Research fund (IOF): IOF Fellowship 13-0260; iMinds Medical Information Technologies SBO 2015, ICON projects (MSIPad, MyHealthData) VLK Stichting E. van der Schueren: rectal cancer; Federal Government: FOD: Cancer Plan 2012-2015 KPC-29-023 (prostate); COST: Action: BM1104: Mass Spectrometry Imaging. Marc Claesen is funded by IWT (grant no. 111065). Jesse Davis is partially supported by the Research Fund KU Leuven (OT/11/051), EU FP7 Marie Curie Career Integration Grant.
Assessing binary classifiers using only positive and unlabeled data

(#294068) and FWO-Vlaanderen (G.0356.12).

References

Blackard, Jock A. and Dean, Denis J. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and Electronics in Agriculture*, 24 (3):131–151, December 1999.

Blum, Avrim and Mitchell, Tom. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th annual conference on Computational Learning Theory (COLT)*, pp. 92–100. ACM, 1998.

Bradley, Andrew P. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159, July 1997. ISSN 0031-3203. doi: 10.1016/S0031-3203(96)00142-2.

Bunescu, Razvan, Ge, Ruifang, Kate, Rohit J, Marcotte, Edward M, Mooney, Raymond J, Ramani, Arun K, and Wong, Yuk Wah. Comparative experiments on learning information extractors for proteins and their interactions. *Artificial Intelligence in Medicine*, 33(2):139–155, 2005.

Calvo, Borja, Larrañaga, Pedro, and Lozano, José A. Learning Bayesian classifiers from positive and unlabeled examples. *Pattern Recognition Letters*, 28(16):2375–2384, 2007.

Chapelle, Olivier, Schölkopf, Bernhard, Zien, Alexander, et al. *Semi-supervised learning*, volume 2. MIT press Cambridge, 2006.

Chapelle, Olivier, Sindhwani, Vikas, and Keerthi, Sathiya S. Optimization techniques for semi-supervised support vector machines. *Journal of Machine Learning Research*, 9:203–233, 2008.

Chawla, Nitesh V and Karakoulas, Grigoris I. Learning from labeled and unlabeled data: An empirical study across techniques and domains. *Journal of Artificial Intelligence Research*, 23:331–366, 2005.

Claesen, Marc, Simm, Jaak, Popovic, Dusan, Moreau, Yves, and De Moor, Bart. Easy hyperparameter search using Optunity. *CoRR*, abs/1412.1114, 2014. URL http://arxiv.org/abs/1412.1114. Software available at http://www.optunity.net.

Claesen, Marc, De Smet, Frank, Suykens, Johan A.K., and De Moor, Bart. A robust ensemble approach to learn from positive and unlabeled data using SVM base models. *Neurocomputing*, 160(0):73 – 84, 2015. doi: http://dx.doi.org/10.1016/j.neucom.2014.10.081.

Davis, Jesse and Domingos, Pedro. Deep transfer via second-order Markov logic. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pp. 217–224, 2009.

Davis, Jesse and Goadrich, Mark. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine learning (ICML)*, pp. 233–240. ACM, 2006.

Denis, François, Gilleron, Rémi, and Letouzey, Fabien. Learning from positive and unlabeled examples. *Theoretical Computer Science*, 348(1):70–83, 2005.

Duarte, Marco F and Hen Hu, Yu. Vehicle classification in distributed sensor networks. *Journal of Parallel and Distributed Computing*, 64(7):826–838, 2004.

Dvoretzky, Aryeh, Kiefer, Jack, and Wolfowitz, Jacob. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, pp. 642–669, 1956.

Efron, Bradley and Tibshirani, Robert J. *An introduction to the bootstrap*. CRC press, 1994.

Elkan, Charles and Noto, Keith. Learning classifiers from only positive and unlabeled data. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD ’08, pp. 213–220, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-193-4. doi: 10.1145/1401890.1401920.

Goadrich, Mark, Oliphant, Louis, and Shavlik, Jude. Learning ensembles of first-order clauses for recall-precision curves: A case study in biomedical information extraction. In *Proceedings of the 14th International Conference on Inductive Logic Programming (ILP)*, pp. 98–115. Springer, Porto, Portugal, 2004.

Goldman, Sally A. and Zhou, Yan. Enhancing supervised learning with unlabeled data. In *Proceedings of the 17th International Conference on Machine Learning (ICML)*, pp. 327–334, 2000.

Hanley, James A and McNeil, Barbara J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 1982.

Holmberg, Scott D. The estimated prevalence and incidence of HIV in 96 large US metropolitan areas. *American Journal of Public Health*, 86(5):642–654, 1996.
Assessing binary classifiers using only positive and unlabeled data

Khorana, Alok A, Kuderer, Nicole M, Culakova, Eva, Lyman, Gary H, and Francis, Charles W. Development and validation of a predictive model for chemotherapy-associated thrombosis. *Blood*, 111(10):4902–4907, 2008.

Kok, Stanley and Domingos, Pedro. Learning Markov logic networks using structural motifs. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, 2010.

Lessmann, Stefan, Baesens, Bart, Mues, Christophe, and Pietsch, Swantje. Benchmarking classification models for software defect prediction: A proposed framework and novel findings. *IEEE Transactions on Software Engineering*, 34(4):485–496, 2008.

Liu, Bing, Dai, Yang, Li, Xiaoli, Lee, Wee Sun, and Yu, Philip S. Building text classifiers using positive and unlabeled examples. In *3rd IEEE International Conference on Data Mining, 2003. ICDM 2003.*, pp. 179–186. IEEE, 2003.

Massart, Pascal. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *The Annals of Probability*, pp. 1269–1283, 1990.

Mordelet, Fantine and Vert, J-P. A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognition Letters*, 37:201–209, 2014.

Mordelet, Fantine and Vert, Jean-Philippe. ProDiGe: Prioritization of disease genes with multitask machine learning from positive and unlabeled examples. *BMC Bioinformatics*, 12(1):389, 2011.

Nigam, Kamal, McCallum, Andrew Kachites, Thrun, Sebastian, and Mitchell, Tom. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3):103–134, 2000.

Pencina, Michael J, D’Agostino, Ralph B, and Vasan, Ramachandran S. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, 27(2):157–172, 2008.

Raghavan, Vijay, Bollmann, Peter, and Jung, Gwang S. A critical investigation of recall and precision as measures of retrieval system performance. *ACM Transactions on Information Systems (TOIS)*, 7(3):205–229, 1989.

Raina, Rajat, Battle, Alexis, Lee, Honglak, Packer, Benjamin, and Ng, Andrew Y. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th International Conference on Machine Learning (ICML)*, pp. 759–766. ACM, 2007.