# Soft Receiver Operating Characteristics Curves

**William Klement** University of Ottawa, K1N 6N5 Canada, klement@site.uottawa.ca
**Peter Flach** University of Bristol, BS8 1UB United Kingdom, Peter.Flach@bristol.ac.uk
**Keywords**: Classification, Ranking, Probability Estimation, Performance Analysis, Machine Learning, ROC

## Abstract

This paper argues that evaluating a learning model should measure the combined performance of several learning tasks, in this case, classification, ranking, and probability estimation. We propose a method to incorporate the scores, which may be probability estimates and are assigned by the model to test data, into the ROC Curve to construct a Soft ROC curve. Our method to construct the Soft ROC curve preserves all characteristics of an ROC curve with the added incorporation of the magnitudes of the scores.

## 1. Introduction

Classification, ranking, and probability estimation are three methods commonly used in machine learning to build learning models. In binary class domains, a classification task is concerned with making decisions to categorize instances into the positive or the negative class. Ranking produces an order of instances from most positives (with high ranks) to most negatives (of low ranks). Probability estimation involves modeling the posterior probability distribution of instances being positive given the training data. The performance of classification is naturally assessed by measuring the accuracy of predictions. The performance of ranking is commonly measured by the Receiver Operating Characteristics (ROC) curve which can be reduced to a scalar metric by taking the area under the ROC curve (AUC) (Provost & Fawcett, 2001). It has been shown that the AUC is a better metric than accuracy to measure the performance of classification (Ling et al., 2003; Provost et al., 1998). The AUC is defined to measure the performance of ranking by assessing how well the model separates the two classes. The quality of probability estimation is measured by the mean squared error (MSE) or Brier score (Brier, 1950). The ROC analysis is a better method than accuracy due to its robustness to changing class or cost distributions (Provost & Fawcett, 2001; Wu et al., 2007). However, the scores assigned to instances by the learning model are excluded from the ROC analysis. It is clear that these performance methods measure different things. Given that a transformations is possible between any two of classification, ranking, or probability estimation tasks, we propose the question of what and how many learning tasks should one evaluate in a given domain? In many domains, it is not always obvious which learning task should be the subject of evaluation. In fact, obtaining good classification performance does not guarantee obtaining good rankings, and vise versa. Similarly, obtaining good probability estimates does not imply a better ranking. Therefore, we argue that measuring the performance of a learning model is not always specific to measuring the performance a single learning task. Instead, we propose that the performance be measured by a combination of these tasks together. A similar approach is used to develop the sAUC metric which combines the margins of ranks of a pair of instances into the AUC metric (Wu et al., 2007).

In this paper, we take the above position and present a generic method to incorporate the scores, assigned to instances by the learning model, into the ROC curve. The result is a generalized ROC curve that combines the performance of all three learning tasks. We call this curve a "*Soft ROC Curve*". We begin with a discussion of the ROC analysis (section 2) followed by a brief review of the three tasks; classification, ranking, and probability estimation tasks (section 3). Section 4 proposes a method to incorporate the scores into the ROC space and section 5 concludes with a summary.

## 2. The ROC analysis

Building a predictive model (a classifier, a ranker, or a probability estimator) involves training a learning algorithm on a set of instances. Then, this model's

*Table 1.* Scores produced by learning models $C_1$ and $C_2$.

| Labels | Scores by $C_1$ | Scores by $C_2$ | ROC Threshold |
|---|---|---|---|
| + | 0.80 | 0.99 | 1 |
| + | 0.70 | 0.78 | 0.79 |
| - | 0.60 | 0.68 | 0.69 |
| + | 0.40 | 0.58 | 0.59 |
| - | 0.30 | 0.38 | 0.39 |
| - | 0.20 | 0.01 | 0.29 |
| | – | – | 0.00 |

predictions, on a set of unseen instances, are evaluated to measure its performance. This evaluation can be performed using the ROC analysis. An ROC curve is generated by imposing a threshold on scores produced by the model, the subject of evaluation, to obtain classification decisions for instances and form a confusion matrix. Varying the threshold from 0 to 1 produces several confusion matrices for which the true positive rates are plotted against the false positive rates. This builds an ROC curve based on classification decisions for all threshold values. In this process, however, the scores are excluded from the analysis once classification decisions have been made. We suggest that the magnitudes, distribution, and order of such scores, particularly when they are probability estimates, can potentially provide additional insights beneficial to assessing the performance of the learning model.

Since the scores are only used to compare against the threshold and are, thereafter, eliminated from the analysis, ROC curves are unable to distinguish between instances whose scores differ significantly in magnitude. For instance, if the model produces probabilities of class membership as scores and if the classification threshold is set to 0.5, then, two positive instances whose scores are 0.9 and 0.51, respectively, will be classified as positives. However, the margin in their probabilities is ignored and, as far as the ROC curve is concerned, both are classified correctly. Furthermore, excluding the scores from the ROC analysis can result in plotting identical ROC curves for different predictive models despite them assigning significantly different scores to the same set of instances. For example, consider scores produced by models $C_1$ and $C_2$ shown in table 1. Given threshold values in the right column of table 1, their ROC curves are identical (figure 1) despite their scores being different for every instance. This suggests that the ROC analysis can be insensitive to differences in the scores.

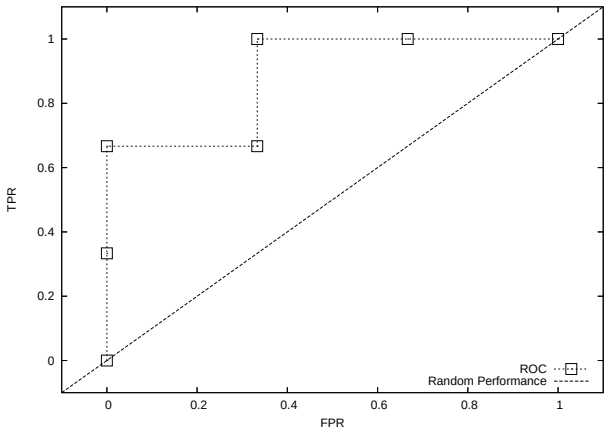


*Figure 1.* The ROC Curves are identical for $C_1$ and $C_2$

## 3. Classification, ranking, and probability estimation

Following the problem description presented by (Flach & Matsubara, 2007), we let $X$ be a set of $n$ instances where the $i^{th}$ instance is a vector $x_i$ of values for attributes $A_1, \cdots A_n$. In classification problems, a *classifier* is a mapping $\hat{c} : X \to C$ where $C$ is a set of labels with binary classification being a special case where $C = \{+, -\}$. A *scoring classifier* is a mapping $\hat{s} : X \to \mathbb{R}$ which assigns scores $\hat{s}(x_i)$ to each $x_i \in X$. For simplicity, we assume that the score $\hat{s}(x_i)$ relates to the expectation of instance $x_i$ being positive, i.e. instances with higher scores have higher expectation of being positive than those with lower scores. Therefore, ordering instances in $X$ by a decreasing value of their $\hat{s}$ scores produces a ranking of $X$. When a scoring classifier produces scores as probability estimates of instances being positive, then it is called a *probability estimator* which produces a mapping $\hat{p} : X \to [0, 1]$, where $\hat{p}(x)$ is taken as an estimate of the posterior $p(+|x)$. Ranking is a total order with potential ties represented by an equivalence relation over $X$ (Flach & Matsubara, 2007). A ranker is a function represented as $\hat{r} : X \times X \to \{>, =, <\}$ that decides, for any pairs of instances, whether the first is more likely ($>$), equally likely ($=$), or less likely ($<$) to be positive than the second. It is important to point out that a ranker may or may not assign scores to instances, i.e. a *scoring ranker* produces scores, which may well be probabilities, then ranks the instances based on their order of scores. Alternatively, a *non-scoring ranker* produces an order or ranks of instances without providing any scores or probabilities.

A classifier predicts labels for each instance. A scoring

classifier assigns scores to instances and requires (for binary classification tasks) a fixed threshold on these scores such that scores higher than (or equal to) the threshold are classified as positives and the remaining instances are classified as negatives. Similarly, a probability estimator assigns probabilities to instances and requires (for use as a classifier) a fixed threshold imposed on these probabilities. A ranker assigns ranks to instances and can also be used as a classifier by using the same technique of thresholding the ranks. However, the magnitudes of scores, probabilities, or ranks are ignored in classification tasks. The focus of classification is the class labels, whereas, for ranking, the focus is the order of instances. Similarly, for a probability estimator, the focus is computing these posterior probabilities of class membership.

## 4. Incorporating scores into ROC space

For ease of representation, we define some notations. For a data set of examples, let $n^+$ ($n^-$) be the number of positive (negative) examples, thus, $n = n^+ + n^-$ is the size of the data set. Let $s_i$ represent the score assigned by the learning model to the $i^{th}$ example. Also, let $m^+$ ($m^-$) be the average score of the positive (negative) examples. Finally, let $L_i$ be the label associated with the $i^{th}$ example.

Since their early introduction to the machine learning community, ROC Curves are produced by plotting the true positive rate against the false positive rate for all threshold values between 0 and 1 ($0 \leq threshold \leq 1$) (Provost & Fawcett, 2001). However, The work in (Fawcett & Niculescu-Mizil, 2007) presents an alternative method to plot an ROC curve. Their method plots the ROC curve by examining the set of examples one by one in a decreasing order of their $s_i$ scores. For each positive example, the true positive rate is incremented by one vertical-step-size and for each negative example, the false positive rate is incremented by one horizontal-step-size. These step sizes are computed by $\frac{1}{n^+}$ in the vertical direction and by $\frac{1}{n^-}$ in the horizontal direction. Plotting these true positive rates against their corresponding false positive rates produces the ROC curve.

Our idea of incorporating the scores into the ROC space is based on altering these step-sizes proportional to the magnitudes of the scores. However, we want the resulting curve to remain an ROC curve such that ROC analysis remain valid, particularly, the definition of the area under the ROC curve. The following sections examine possible methods of adjusting these step-sizes in two settings; first in a single direction only and second, in both directions simultaneously.

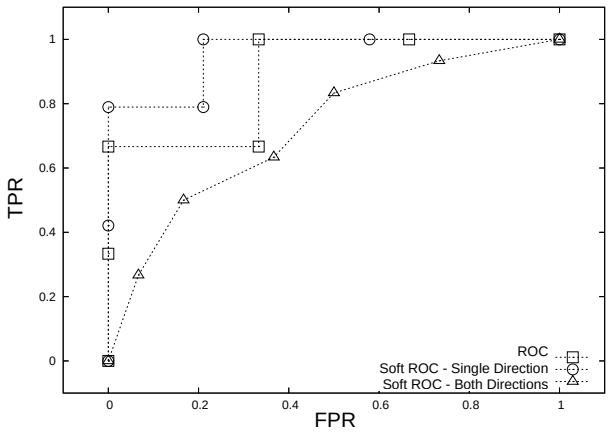


*Figure 2.* Soft ROC curves for $C_1$ in table 1

### 4.1. In a single direction only

First, we consider altering the step-sizes to move up proportional to $s_i$ for the positive examples and move right proportional to $(1-s_i)$ for the negative examples. This means that in each step, only a single change of a step-size can occur, either in the upward direction (for a positive example), or in the horizontal direction (for a negative example) but never both. Furthermore, we need to normalize the vertical step-sizes by $\frac{1}{m^+n^+}$ and by $\frac{1}{(1-m^-)n^-}$ in the horizontal direction to maintain the unit square in the ROC space. The effect of this approach can be viewed as a method of sampling the data based on the magnitudes of the scores. In a sense, the positive examples are sampled by the normalized magnitudes of their scores ($\frac{s_i^+}{m^+n^+}$) to boost the scale of the vertical climb of the ROC curve. Similarly, the negative examples are sampled by the normalized magnitudes of their negative scores ($\frac{1-s_i^-}{m^+n^+}$) on the horizontal run. This results in stretching the original ROC curve at the bottom left and the top right of the ROC space while shrinking the part of the ROC curve that is towards the top left corner of the space. This effect is illustrated in figure 2. We suggest that the Soft ROC curve (in a single direction) is an optimistic curve that mostly dominates the original ROC curve. Due to this bias in favor of the learning model, the usefulness of this approach remains in question.

### 4.2. In both directions simultaneously

An alternative way to incorporate the scores into the ROC space is to adjust the ROC step-sizes in both directions, vertically and horizontally simultaneously. Intuitively, the move upwards can be proportional to

$s_i$ whereas the move to the right can be scaled proportional to $(1-s_i)$. Such adjustment requires the normalization by $\frac{1}{mn}$ vertically and by $\frac{1}{(1-m)n}$ horizontally to maintain the unit square in the space. However, this approach fails to account for class information because the class labels are excluded from the process of plotting this ROC curve. With respect to performance analysis, this approach falls short, however, it may be useful to determining some form of an upper bound on performance. To remedy this short fall and account for class information, our algorithm should treat correctly classified examples favorably, whereas incorrectly classified examples should contribute a penalty into the performance assessment. In the ROC space, good performance is represented by a climb in the vertical direction while a run to the right represents a penalty. Therefore, if we impose a threshold value on the scores $s_i$ to make classification decisions, then, we can compare these decisions to the class labels. To compute the step-sizes in both directions simultaneously, our proposed method follows: for a correctly classified example $x_i$, it climbs up proportional to score $s_i$ while it runs horizontally proportional to $(1 - s_i)$. Also, for an incorrectly classified example $x_i$, our method climbs up proportional to $(1 - s_i)$ while it runs horizontally proportional to $s_i$.

In order to ensure that the rate of climb and the rate of run comply with the unit square of the ROC space, we need to normalize the step-sizes. This normalization becomes cumbersome and is omitted due to space constraints, however, it can be calculated incrementally. At this point, it suffices to state that we have successfully proposed a method to incorporate the scores into the ROC space. Our proposed method accomplishes this task while preserving all properties of the ROC curve in the ROC space. The idea can be viewed as sampling each example in the test set at a higher resolution proportional to the score assigned to it by the learning model. One remaining detail is the calculation of the threshold by which examples are classified. The answer is intuitive when the scores are calibrated probability estimates, we use 0.5. However, when the scores are probabilities and may not be calibrated, we calculate this threshold to be the weighted mean score value between the positive and the negative scores, weighted by the class distribution $c = \frac{n^+}{n^-}$. Such threshold value is computed on the training data and lies close to the middle point between most positive and negative examples.

Comparing the Soft ROC (in both directions method) curve to the ROC curve in figure 2 reveals that the former depicts a more pessimistic performance for model $C_1$ than that shown by the original ROC curve. This can easily be explained by inspecting the scores produced by $C_1$ in table 1. All the scores assigned to the positive instances are less than 1, therefore, they contribute with a penalty. Similarly, the scores assigned to the negative examples are all greater than zero, thus, they also contribute with a penalty to the Soft ROC curve (in both directions method).

## 5. Conclusions

In this paper, we have argued that evaluating a learning model should measure the combined performance of several learning tasks, in this case, these were; classification, ranking, and probability estimation. In addition, We have presented a preliminary investigation of methods aiming at incorporating the scores, assigned by the learning model to instances, into the ROC Curve for the purpose of performance analysis. We plan to further investigate the interpretations, analysis and potentials of the proposed Soft ROC curves to develop an evaluation framework that combines performance measures of these three learning tasks.

## References

Brier, G. (1950). Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review*, *78*, 1–3.

Fawcett, T., & Niculescu-Mizil, A. (2007). Pav and the roc convex hull. *Machine Learning*, *68*, 97–106.

Flach, P., & Matsubara, E. (2007). Obtaining calibrated probability estimates from simple lexicographic rankings. *In proceedings of ECML/PKDD 2007* (pp. 575–582).

Ling, C. X., Huang, J., & Zhang, H. (2003). Auc: A better measure than accuracy in comparing learning algorithms. *In proceedings of Canadian Conference on AI* (pp. 329–341).

Provost, F., & Fawcett, T. (2001). Robust classification systems for imprecise environments. *Machine Learning*, *42*, 203–231.

Provost, F. J., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *In proceeding of ICML'98* (pp. 445–453).

Wu, S., Flach, P., & Ferri, C. (2007). An improved model selection heuristic for auc. *In proceedings of ECML/PKDD 2007* (pp. 478–489).