Pattern Recognition 46 (2013) 3395–3411

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

ROC curves for regression

José Hernández-Orallo [*]

Departament de Sistemes Informàtics i Computació, Universitat Politècnica de València, Camí de Vera s/n, E-46022 València, Spain

ARTICLE INFO

Article history:
Received 27 December 2012
Received in revised form
1 May 2013
Accepted 11 June 2013
Available online 24 June 2013

Keywords:
ROC Curves
Cost-sensitive regression
Operating condition
Asymmetric loss
Error variance
MSE decomposition

ABSTRACT

Receiver Operating Characteristic (ROC) analysis is one of the most popular tools for the visual assessment and understanding of classifier performance. In this paper we present a new representation of *regression* models in the so-called regression ROC (RROC) space. The basic idea is to represent over-estimation against under-estimation. The curves are just drawn by adjusting a *shift*, a constant that is added (or subtracted) to the predictions, and plays a similar role as a threshold in classification. From here, we develop the notions of optimal operating condition, convexity, dominance, and explore several evaluation metrics that can be shown graphically, such as the area over the RROC curve (*AOC*). In particular, we show a novel and significant result: the *AOC* is equivalent to the error variance. We illustrate the application of RROC curves to resource estimation, namely the estimation of software project effort.

1. Introduction

Receiver Operating Characteristic (ROC) analysis [36,47,4,48,19,35,14,34] is a very popular technique for the graphical analysis of classification models. ROC analysis is profusely used in many areas [21,37,30,31]: radiology, medicine, statistics, bioinformatics, machine learning, pattern recognition, etc. Also, some metrics derived from the ROC curve, such as the Area Under the ROC Curve (AUC), are now key for the evaluation and construction of classifiers [15,38,50,42,32].

In classification, the traditional notion of operating condition is common and well understood. Classifiers may be trained for one cost proportion and class distribution (both making the operating condition) and then deployed on a different operating condition. ROC space decomposes the performance of a classifier in a dual way. On the x-axis we show the false positive rate (FPR) and on the y-axis we show the true positive rate (TPR). ROC curves neatly visualise how the TPR and the FPR change for different (crisp) classifiers or evolve for the same (soft) classifier (or ranker) for a range of thresholds. The notion of threshold is the fundamental idea to adapt a soft classifier to an operating condition. ROC analysis is the tool that illustrates how classifiers and threshold choices perform.

The adaptation of ROC analysis for regression has been attempted on many occasions. However, there is no such a thing as the 'canonical' adaptation of ROC analysis in regression, since regression and classification are different tasks, and the notion of operating condition may be completely different. In fact, the mere extension of ROC analysis to more than two classes has always been difficult because the degrees of freedom grow quadratically with the number of classes (see, e.g., [46,17,44]). Consequently it is even questionable whether a similar graphical representation of ROC curves in regression (or other tasks [26]) can even be figured out. Notable efforts towards ROC curves (or graphical tools) for regression are the Regression Error Characteristic (REC) Curves [3], the Regression Error Characteristic Surfaces (RECS) [51], the notion of utility-based regression [41] and the definition of ranking measures [43]. These approaches are based on gauging the tolerance, rejection rules or confidence levels. Some of these approaches actually convert the evaluation of a regression problem into a classification problem (tolerable estimation vs. intolerable estimation). However, none of these previous approaches started from a notion of 'operating condition', related to an *asymmetric loss function*. Also, the notion of threshold was not replaced by a similar concept playing its role for adjusting to the operating condition, and the dual positive-negative character in ROC analysis was blurred.

In this paper we present a graphical representation of regression performance based on a very usual view of operating condition. Many regression applications have deployment contexts where over-estimations are not equally costly as under-estimations (or vice versa).

[*] Tel.: +34 963877007; fax: +34 963877359.
E-mail address: jorallo@dsic.upv.es

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.06.014

This is called the *loss asymmetry*. Certainly, loss asymmetry is just one possible kind of operating condition (or one of its constituents), but it is a very common and important one in many applications.

The ROC space for regression (RROC space) is then defined by placing the total over-estimation on the *x*-axis and the total under-estimation on the *y*-axis. This duality leads to regions and isometrics in the ROC space where over-estimations have less cost than under-estimations and vice versa. There we can plot different regression models to see the notions of dominance. We also consider the construction of hybrid regression models by 'interpolating' between points in the RROC space. Moreover, the plot leads to *curves* (called RROC curves) when we use the notion of *shift*, which is just a constant that we can add (or subtract) to example predictions in order to adjust the model to an asymmetric operating condition. This notion is parallel to the notion of threshold in classification. Interestingly, while we can derive the best shift for a dataset given an existing model (which boils down to finding the shift that makes its average error equal to zero if the cost is symmetric), there are some effective methods to determine this shift for the deployment data given an operating condition, as has been recently explored by [1,55]. All this leads to a more meaningful interpretation of what the ROC curves for regression really mean, and what their areas represent.

The paper is organised as follows. Section 2 introduces some notation, the problem of cost-sensitive evaluation and the use of asymmetric costs in regression. The RROC space is defined in Section 3, where we represent several regression models as points, derive the isometrics of the space and develop the notions of hybrid models, dominance and convex hull. Section 4 introduces RROC curves, which are drawn by ranging a constant additive shift. We define an algorithm for plotting them and determine some of its properties in terms of segment slopes and convexity. Section 5 analyses the area over the RROC curve (*AOC*), proving its linear relation to error variance, and showing that the squared error decomposition can be shown in RROC space. A real example is included in Section 6, which illustrates how RROC curves are used from training to deployment. Finally, Section 7 closes the paper with an enumeration of issues for future investigation.

2. Background

In this section we introduce some notation and the basic concepts about cost-sensitive regression and the need of asymmetric loss functions.

2.1. Notation

Let us consider a multivariate input domain $\mathbb{X}$ and a univariate output domain $\mathbb{Y} \subset \mathbb{R}$. The domain space $\mathbb{D}$ is then $\mathbb{X} \times \mathbb{Y}$. Examples or instances are just pairs $\langle x, y \rangle \in \mathbb{D}$, and datasets are subsets (actually multi-sets) of $\mathbb{D}$. The length of a dataset will usually be denoted by n. A *crisp* regression model m is a function $m : \mathbb{X} \rightarrow \mathbb{Y}$. When the regression model is crisp, we just represent the true value by y and the estimated value by $\hat{y}$. Subindices will be used when referring to more than one example in a dataset. Vectors (unidimensional arrays) are denoted in boldface and its elements with subindices, e.g., $\mathbf{v} = (v_1, v_2, \ldots, v_n)$. Operations mixing arrays and scalar values will be allowed, specially in algorithms, as usual in the matrix arithmetic of many statistical computing languages. For instance, $\mathbf{v} + c$ means that the constant c is added to all the elements in the vector $\mathbf{v}$. The mean of a vector is denoted by $\mu(\mathbf{v})$ and its standard deviation as $\sigma(\mathbf{v})$—over the population, i.e., divided by n. Given a dataset with n instances $i = 1 \ldots n$, the error vector $\mathbf{e}$ is defined such that $e_i \triangleq \hat{y}_i - y_i$. The value $\mu(\mathbf{e}^2)$ is known as the mean squared error (*MSE*), $\mu(\mathbf{e})$ is known as the mean error (or mean error bias, *MEB*), $\mu(|\mathbf{e}|)$ is known as the mean absolute error (*MAE*) and $\sigma(\mathbf{e})^2$ as the error variance. Occasionally, we will drop the preceding *M*, especially when referring to total squared error (*SE*), total error bias (*EB*) and total absolute error (*AE*).

2.2. Cost-sensitive problems and loss functions

In cost-sensitive learning [13], there are several features which describe a context, such as the data distribution, the costs of using some input variables and the loss of the errors over the output variables [52]. In this paper, we focus on loss functions over the output variable, which is the kind of costs ROC analysis deals with (typically integrated, along with the class distribution, within the notion of *skew*). A loss function is defined as follows:

Definition 1. A loss function is any function $\ell : \mathbb{Y} \times \mathbb{Y} \rightarrow \mathbb{R}$ which compares elements in the output domain. For convenience, the first argument will be the estimated value, and the second argument the actual value.

Typical examples of loss functions are the absolute error (ℓ^A) and the squared error (ℓ^S), with $\ell^A(\hat{y}, y) \triangleq |\hat{y} - y|$ and $\ell^S(\hat{y}, y) \triangleq (\hat{y} - y)^2$. These two loss functions are *symmetric*, i.e., for every y and r we have that $\ell(y + r, y) = \ell(y - r, y)$. Two of the most common metrics for evaluating regression, the mean absolute error (*MAE*) and the mean squared error (*MSE*) are derived from these losses.

2.3. Asymmetric costs

Actually, although symmetric loss functions (and derived metrics) are common for the evaluation of regression models, it is rarely the case that a real problem has a symmetric cost. For instance, the prediction of sales, consumptions, calls, prices, demands, survival times, positions, reliabilities, etc., rarely has a symmetric loss. For instance, a retailing company may need to predict how many items will be sold next week for stock (inventory) management purposes, e.g., in order to calculate how many items must be ordered. Depending on the kind of product, it is usually not the same to over-estimate (increasing stocking costs) than to under-estimate (an item goes out of stock and it cannot be sold or sold with delays). In fact, it is also rare to find applications where even an asymmetric cost is invariable. In the above-mentioned example, depending on the warehouse saturation, the cost (and the asymmetry) may change in a weekly or daily fashion. Because of this, a specialised model for each fixed given asymmetry is not a practical solution in general. This motivates the adaptation (or reframing) of models, rather than their re-training for each new asymmetric loss. This variability of the operating condition is at the core of ROC analysis in classification.