Model Assessment with ROC Curves

Lutz Hamel
University of Rhode Island, USA

INTRODUCTION

Classification models and in particular binary classification models are ubiquitous in many branches of science and business. Consider, for example, classification models in bioinformatics that classify catalytic protein structures as being in an active or inactive conformation. As an example from the field of medical informatics we might consider a classification model that, given the parameters of a tumor, will classify it as malignant or benign. Finally, a classification model in a bank might be used to tell the difference between a legal and a fraudulent transaction.

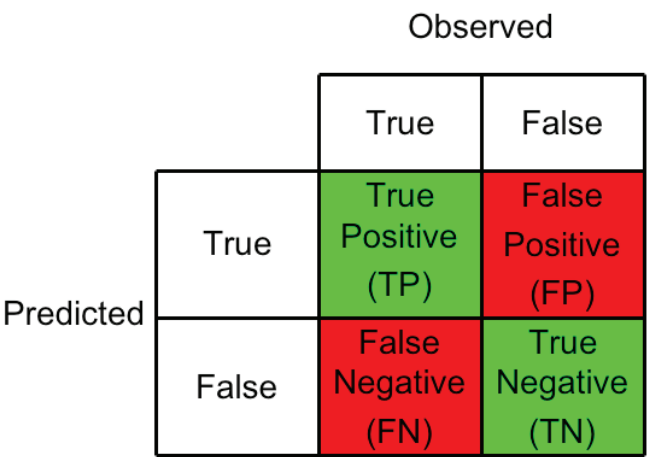
Central to constructing, deploying, and using classification models is the question of model performance assessment (Hastie, Tibshirani, & Friedman, 2001). Traditionally this is accomplished by using metrics derived from the confusion matrix or contingency table. However, it has been recognized that (a) a scalar is a poor summary for the performance of a model in particular when deploying non-parametric models such as artificial neural networks or decision trees (Provost, Fawcett, & Kohavi, 1998) and (b) some performance metrics derived from the confusion matrix are sensitive to data anomalies such as class skew (Fawcett & Flach, 2005). Recently it has been observed that Receiver Operating Characteristic (ROC) curves visually convey the same information as the confusion matrix in a much more intuitive and robust fashion (Swets, Dawes, & Monahan, 2000).

Here we take a look at model performance metrics derived from the confusion matrix. We highlight their shortcomings and illustrate how ROC curves can be deployed for model assessment in order to provide a much deeper and perhaps more intuitive analysis of the models. We also briefly address the problem of model selection.

BACKGROUND

A binary classification model classifies each instance into one of two classes; say a *true* and a *false* class. This gives rise to four possible classifications for each

Figure 1. Format of a confusion matrix



instance: a true positive, a true negative, a false positive, or a false negative. This situation can be depicted as a confusion matrix (also called contingency table) given in Figure 1. The confusion matrix juxtaposes the observed classifications for a phenomenon (columns) with the predicted classifications of a model (rows). In Figure 1, the classifications that lie along the major diagonal of the table are the correct classifications, that is, the true positives and the true negatives. The other fields signify model errors. For a perfect model we would only see the true positive and true negative fields filled out, the other fields would be set to zero. It is common to call true positives *hits*, true negatives *correct rejections*, false positive *false alarms*, and false negatives *misses*.

A number of model performance metrics can be derived from the confusion matrix. Perhaps, the most common metric is *accuracy* defined by the following formula:

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Other performance metrics include *precision* and *recall* defined as follows:

$$\text{precision} = \frac{TP}{TP + FP},$$

$$\text{recall} = \frac{TP}{TP + FN}$$

Note, that when we apply a model to a test dataset we obtain only one scalar value for each performance metric. Figure 2 shows two confusion matrices of one particular classification model built on the ringnorm data by Breiman (Breiman, 1996). Part (a) shows the classification model being applied to the original test data that consists of 7400 instances roughly split evenly between two classes. The model commits some significant errors and has an accuracy of 77%. In part (b) the model is applied to the same data but in this case the negative class was sampled down by a factor of ten introducing class skew in the data. We see that in this case the confusion matrix reports accuracy and precision values that are much higher than in the previous case. The recall did not change, since we did not change anything in the data with respect to the 'true' class. We can conclude that the perceived quality of a model highly depends on the choice of the test data. In the next section we show that ROC curves are not

Figure 2. Confusion matrices with performance metrics. (a) confusion matrix of a model applied to the original test dataset, (b) confusion matrix of the same model applied to the same test data where the negative class was sampled down by a factor of ten