An Overview of General Performance Metrics of Binary Classifier Systems

Sebastian Raschka
se.raschka@gmail.com

October 2, 2014

The purpose of this document is to provide a brief overview of different metrics and terminology that is used to measure the performance of binary classification systems. The equations are based on *An introduction to ROC analysis* by Tom Fawcett [1].

1 Confusion matrix

The *confusion matrix* (or *error matrix*) is one way to summarize the performance of a classifier for binary classification tasks. This square matrix consists of columns and rows that list the number of instances as absolute or relative "actual class" vs. "predicted class" ratios.
Let P be the label of class 1 and N be the label of a second class or the label of all classes that are *not class 1* in a multi-class setting.

		Predicted class	
		P	N
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

2 Prediction Error and Accuracy

Both the prediction *error (ERR)* and *accuracy (ACC)* provide general information about how many samples are misclassified. The *error* can be understood as the sum of all false predictions divided by the number of total predications, and the the accuracy is calculated as the sum of correct predictions divided by the total number of predictions, respectively.

$$ERR = \frac{FP + FN}{FP + FN + TP + TN} = 1 - ACC \tag{1}$$

$$ACC = \frac{TP + TN}{FP + FN + TP + TN} = 1 - ERR \tag{2}$$

3 False and True Positive Rates

The *True Positive Rate (TPR)* and *False Positive Rate (FPR)* are performance metrics that are especially useful for imbalanced class problems. In *Spam classification*, for example, we are of course primarily interested in the detection and filtering out of *spam*. However, it is also important to decrease the number of messages that were incorrectly classified as *spam* (*False Positives*): A situation where a person misses an important message is considered as "worse" than a situation where a person ends up with a few *spam* messages in his e-mail inbox. In contrast to the *FPR*, the *True Positive Rate* provides useful information about the fraction of *positive* (or *relevant*) samples that were correctly identified out of the total pool of *Positives*.

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \tag{3}$$

$$TPR = \frac{TP}{P} = \frac{TP}{FN + TP} \tag{4}$$

4 Precision, Recall, and the F_1-Score

Precision (PRE) and *Recall (REC)* are metrics that are more commonly used in *Information Technology* and related to the *False* and *True Prositive Rates*. In fact, *Recall* is synonymous to the *True Positive Rate* and is sometimes also called *Sensitivity*. The F_1-Score can be understood as a combination of both *Precision* and *Recall* [2].

$$PRE = \frac{TP}{TP + FP} \tag{5}$$

$$REC = TPR = \frac{TP}{P} = \frac{TP}{FN + TP} \tag{6}$$

$$F_1 = 2 \cdot \frac{PRE \cdot REC}{PRE + REC} \tag{7}$$

5 Receiver Operator Characteristic (ROC)

Receiver Operator Characteristics (ROC) graphs are useful tools to select classification models based on their performance with respect to the *False Positive* and *True Positive* rates.

The diagonal of a ROC graph can be interpreted as *random guessing* and classification models that fall below the diagonal are considered as worse than random guessing. A perfect classifier would fall into the top left corner of the graph with a *True Positive Rate* of 1 and a *False Positive Rate* of 0.

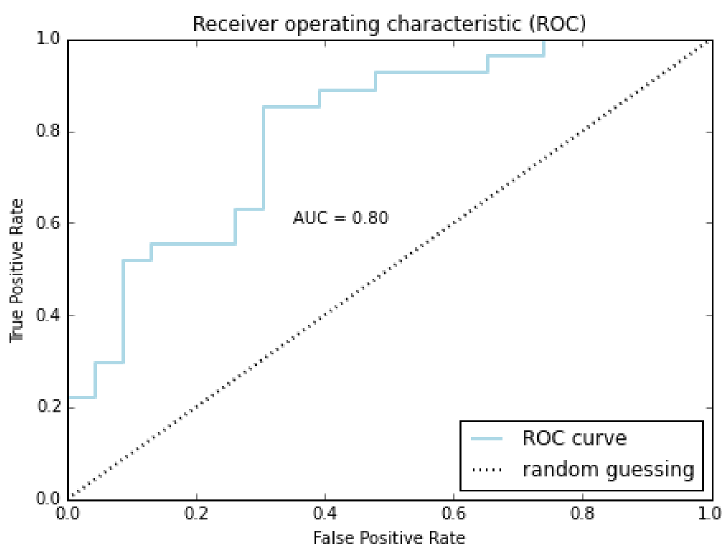

Figure 1: Example of a Receiver Operating Characteristic. This plot was created using the Python scikit-learn machine learning library.

The ROC *curve* can be computed by shifting the decision threshold of a classifier (e.g., the posterior probabilities of a naive Bayes classifier). Based on the ROC *curve*, the so-called *Area Under the Curve (AUC)* can be calculated to characterize the performance of a classification model.

References

[1] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

[2] Cyril Goutte and Eric Gaussier. A probabilistic interpretation of precision, recall and f-score, with implication for evaluation. In *Advances in Information Retrieval*, pages 345–359. Springer, 2005.