Model Assessment with ROC Curves

Lutz Hamel
Department of Computer Science and Statistics
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 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 Fig. 1. The confusion matrix juxtaposes the observed classifications for a phenomenon (columns) with the predicted classifications of a model (rows). In Fig. 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. Fig. 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 so dependent on the precise choice of test data, at least with respect to class skew.

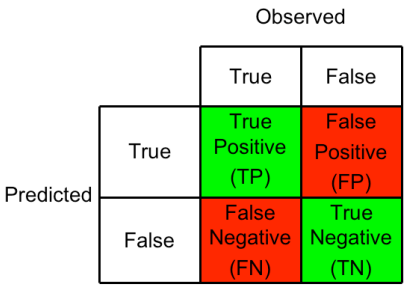


Figure 1: Format of a Confusion Matrix.

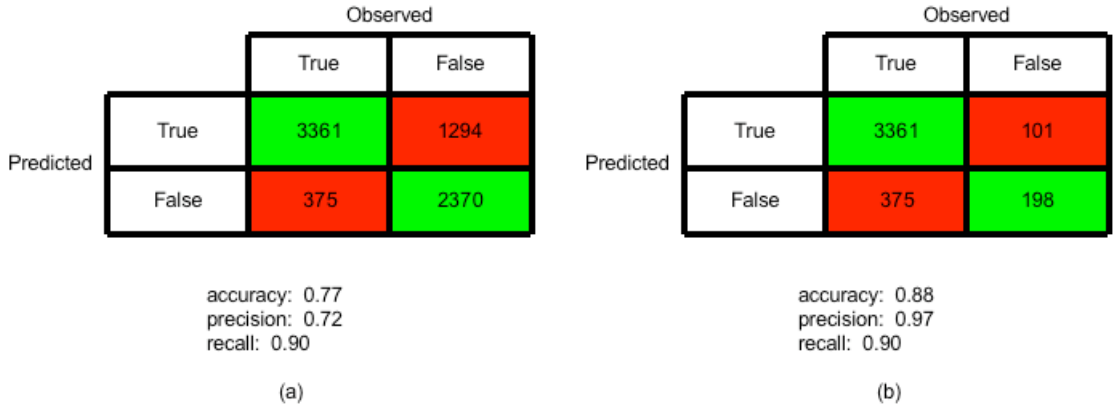


Figure 2: Confusion matrices with performance metrics. (a) Confusion matrix of a model applied to the original test dataset, (b) confusion matrix of the same model applied to the same test data where the negative class was sampled down by a factor of ten.

Main Focus of Chapter

ROC Curves – The Basics

ROC curves are two-dimensional graphs that visually depict the performance and performance trade-off of a classification model (Fawcett, 2004; P. Flach, Blockeel, Ferri, Hernandez-Orallo, & Struyf, 2003; P. Flach, 2004; P. A. Flach, 2003). ROC curves were originally designed as tools in communication theory to visually determine optimal operating points for signal discriminators (Egan, 1975).

We need to introduce two new performance metrics in order to construct ROC curves (we define them here in terms of the confusion matrix), the *true positive rate* (tpr) and the *false positive rate* (fpr):

$$\text{true positive rate} = \frac{TP}{TP + FN} = \text{recall},$$

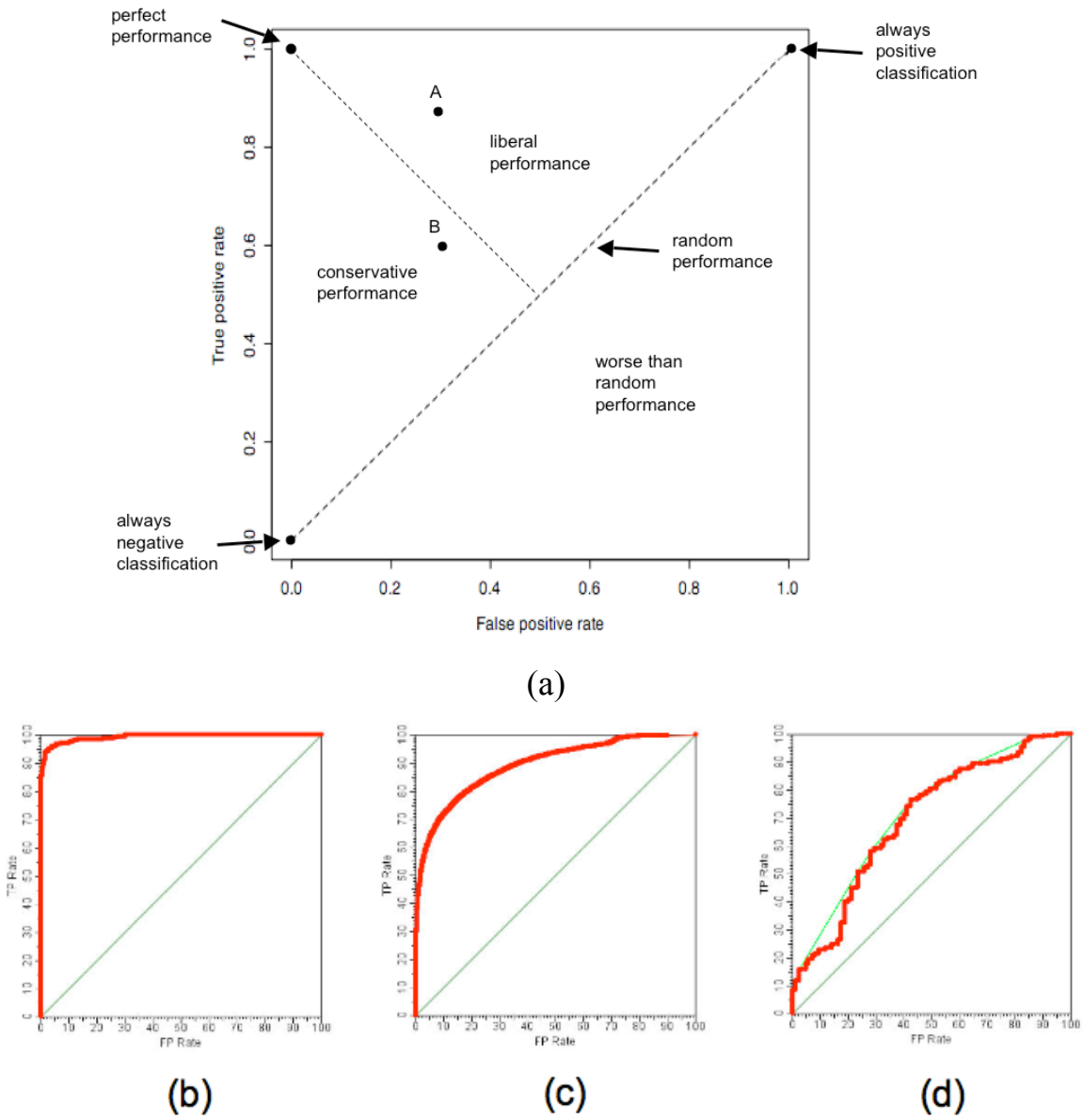
$$\text{false postive rate} = \frac{FP}{TN + FP}.$$

ROC graphs are constructed by plotting the true positive rate against the false positive rate (see Fig. 3(a)). We can identify a number of regions of interest in a ROC graph. The diagonal line from the bottom left corner to the top right corner denotes random classifier performance, that is, a classification model mapped onto this line produces as many false positive responses as it produces true positive responses. To the left bottom of the random performance line we have the conservative performance region. Classifiers in this region commit few false positive errors. In the extreme case, denoted by point in the bottom left corner, a conservative classification model will classify all instances as negative. In this way it will not commit any false positives but it will also not produce any true positives. The region of classifiers with liberal performance occupies the top of the graph. These classifiers have a good true positive rate but also commit substantial numbers of false positive errors. Again, in the extreme case denoted by the point in the top right corner, we have classification models that classify every instance as positive. In that way, the classifier will not miss any true positives but it will also commit a very large number of false positives. Classifiers that fall in the region to the right of the random performance line have a performance worse than random performance, that is, they consistently produce more false positive responses than true positive responses. However, because ROC graphs are symmetric along the random performance line, inverting the responses of a classifier in the "worse than random performance" region will turn it into a well performing classifier in one of the regions above the random performance line. Finally, the point in the top left corner denotes perfect classification: 100% true positive rate and 0% false positive rate.

The point marked with A is the classifier from the previous section with a tpr = 0.90 and a fpr = 0.35. Note, that the classifier is mapped to the same point in the ROC graph regardless whether we use the original test set or the test set with the sampled down negative class illustrating the fact that ROC graphs are not sensitive to class skew.

Classifiers mapped onto a ROC graph can be ranked according to their distance to the 'perfect performance' point. In Fig. 3(a) we would consider classifier A to be superior to a hypothetical classifier B because A is closer to the top left corner.

The true power of ROC curves, however, comes from the fact that they characterize the performance of a classification model as a curve rather than a single point on the ROC graph. In addition, Fig. 3 shows some typical examples of ROC curves. Part (b) depicts the ROC curve of an almost perfect classifier where the performance curve almost touches the 'perfect performance' point in the top left corner. Part (c) and part (d) depict ROC curves of inferior classifiers. At this level the curves provide a convenient visual representation of the performance of various models where it is easy to spot optimal versus sub-optimal models.



(a)

(b) (c) (d)

Figure 3: ROC curves: (a) regions of a ROC graph (a) an almost perfect classifier (b) a reasonable classifier (c) a poor classifier.[1]

ROC Curve Construction

In order to interpret ROC curves in more detail we need to understand how they are constructed. Fundamental to the construction of ROC curves is the notion of instance ranking or prediction confidence value. ROC curves can be directly computed for any

[1] Figures (b), (c), and (d) due to Peter Flach, *ICML'04 tutorial on ROC analysis*, International Conference on Machine Learning, 2004 (P. Flach, 2004).

classification model that attaches a probability, confidence value, or ranking to each prediction. Many models produce such rankings as part of their algorithm (e.g. Naïve Bayes (Mitchell, 1997), Artificial Neural Networks (Bishop, 1995), Support Vector Machines (Cristianini & Shawe-Taylor, 2000)). Techniques exist that compute an instance ranking for classification models that typically do not produce such rankings, i.e., decision trees (Breiman, Friedman, Olshen, & Stone, 1984). The instance ranking is used by the ROC algorithm to sweep through different decision thresholds from the maximum to the minimum ranking value in predetermined increments. The ranking values are typically normalized to values between 0 and 1 (as an aside, the default decision threshold for most classifiers is set to .5 if the ranking value expresses the actual probability value of the instance being classified as true). At each threshold increment, the performance of the model is computed in terms of the true positive and false positive rates and plotted. This traces a curve from left to right (maximum ranking to minimum ranking) in the ROC graph. That means that the left part of the curve represents the behavior of the model under high decision thresholds (conservative) and the right part of the curve represents the behavior of the model under lower decision thresholds (liberal).

The following algorithm[2] makes this construction a little bit more concrete,

Function Draw-ROC

Inputs:	
D	test set
p(i)	ranking of instance i in D, indicates the probability or confidence that the instance i is positive, normalized to [0,1]
P	set of *observed* positive instances in D, where $P \subseteq D$
N	set of *observed* negative instances in D, where $N \subseteq D$

for threshold = 1 **to** 0 **by** -.01 **do**
    FP $\Leftarrow$ 0

[2] Based on the algorithm published by Tom Fawcett (Fawcett, 2004).

```
        TP ⇐ 0
        for i ∈ D do
                if p(i) ≥ threshold then
                        if i ∈ P then
                                TP ⇐ TP + 1
                        else
                                FP ⇐ FP + 1
                        endif
                endif
        endfor
        tpr ⇐ TP/#P
        fpr ⇐ FP/#N
        Add point (tpr, fpr) to ROC curve
endfor
```

Notice how this algorithm sweeps through the range of thresholds from high to low and measures the number of mistakes the classifier makes at each threshold level. This gives rise to the tpr and fpr at each threshold level. This in turn can be interpreted as a point on the ROC curve.

Fig. 4(a) shows the ROC curve of the classifier from Fig. 2 with the decision thresholds annotated in color. From this we can see that the optimal decision threshold for this model (maximum tpr, minimum fpr, also called optimal operating point) occurs at a threshold of .35 in the green region representing a tpr = .95 and an fpr = .45. As we would expect from our confusion matrix analysis, we can observe that it is a reasonable classifier. We can also observe that it is a liberal classifier in that the optimal decision threshold of the curve lies in the liberal region of the ROC graph. It is also interesting to observe that the performance given by the confusion matrix maps to a suboptimal point on the curve (given as 'A' on the curve). This is due to the fact that the classification reported in the confusion matrix is based on the default decision threshold value of .5 instead of the optimal threshold value of .35.

In Fig. 4(b) we can see two ROC curves for the same classifier as in part (a), one is based on the original test data and the other one is based on the skewed test data. Both curves are virtually identical illustrating that ROC curves are not sensitive to class skew.

Returning to Fig. 3 above, we can now interpret these curves a little bit more carefully. In part (b) we see that the model only begins to commit false positive errors after it has almost reached a true positive rate of 100%. This means that at this point the decision threshold has been lowered to a point that observed, negative instances are classified as positive. Thus, when the decision threshold is set too low, a model will commit false positive errors. However, in a near perfect classification model this will not happen until the curve has almost reached the 'perfect performance' point.

In Fig. 3(c) we see that the model also behaves very nicely for a large range of decision threshold values. However, compared to the model in part (b) it starts to commit false positive errors much earlier and therefore the slope of the curve in part (c) is flatter. Another way of stating this is, that there exists no decision threshold for which the model is able to separate the classes perfectly. The model in Fig. 4(d) is not only inferior because its curve is the farthest away from the 'perfect performance' point but we can observe that for large ranges of the ranking values the model commits more false positive errors than it provides true positive classifications. This shows up as concavities in the curve indicating that for certain ranges of the decision threshold the classification model performs worse than a random classifier.

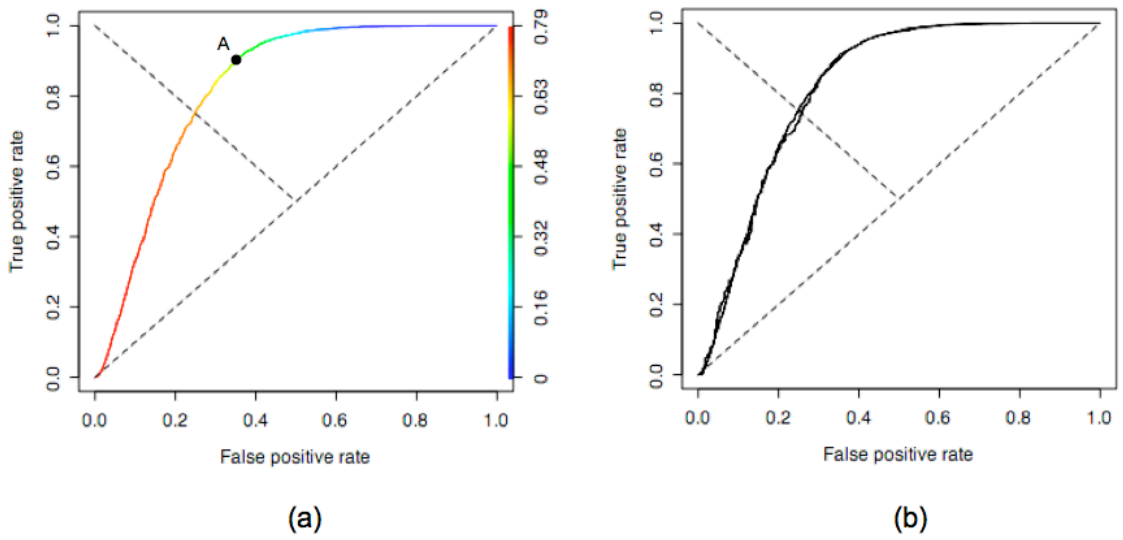


Figure 4: ROC curves of the classifier given in Fig. 2. (a) ROC curve with decision threshold values, (b) ROC curves of the classifier evaluated against original test data and the down sampled data.

Model Selection

A key notion in model assessment is model selection, that is, given two or more classification models, we need to pick one in order to be deployed. The criterion to pick one model over the other(s) has to answer two fundamental questions: (a) it needs to be general enough to describe model performance over a broad range of possible scenarios and (b) it needs to be able to discern whether the performance difference between models is statistically significant. It turns out that ROC curves answer both of these questions in a highly visual manner. Consider Fig. 5(a), here we have two classifiers plotted in a ROC graph together with their respective 95% confidence bands (vertical bars) (Macskassy & Provost, 2004). It is easy to see that the curve that stretches almost into the top left corner represents the performance of the superior model (for a tpr = 0.9 this model commits virtually no false positives). In addition, because the confidence bands of the two curves are clearly separated we can state that the performance difference between the two models is statistically significant. In Fig. 5(b) the situation is not so clear-cut.

We again have two classifiers and the curve that reaches closer to the top left corner of the graph denotes the better performing model. However, since the confidence bands overlap, the performance difference between these two models is not statistically significant. In addition, closer inspection reveals that the confidence band for the upper curve is slightly wider than for the lower curve suggesting greater variability in the performance of the better performing model.

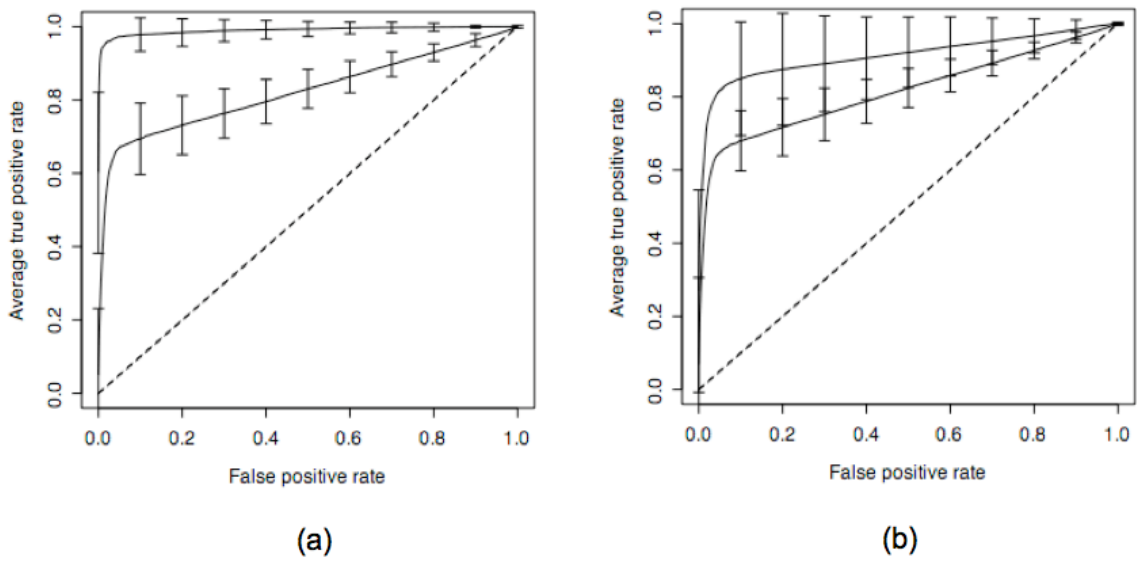


Figure 5: ROC curves with 95% confidence bands. (a) Two classifiers with a statistically significant difference in their performance. (b) Two classifiers whose difference in performance is not statistically significant.

Future Trends

ROC analysis enjoys a continued growth of interest. Since 2004 there have been regularly scheduled workshops, the *Workshops on ROC Analysis in Machine Learning* (ROCML), which bring together an international group of researchers. Robust tools such as the ROCR package[3] for the R environment (Sing, Sander, Beerenwinkel, & Lengauer, 2005) contribute to the rapid adoption of ROC analysis as the preferred model analysis

[3] We used the ROCR package for this work.

technique. At a technical level, the most important development is the extension of this analysis technique from binary classification problems to multi-class problems providing a much wider applicability of this technique (Everson & Fieldsend, 2006; Lane, 2000; Srinivasan, 1999).

Conclusions

Although brief, we hope that this overview provided an introduction to the fact that ROC analysis provides a powerful alternative to traditional model performance assessment using confusion matrices. We have shown that in contrast to traditional scalar performance metrics such as accuracy, recall, and precision derived from the confusion matrix, ROC analysis provides a highly visual account of a model's performance over a range of possible scenarios. We have also shown that ROC analysis is robust with respect to class skew, making it a reliable performance metric in many important application areas where highly skewed data sets are common (e.g. fraud detection).

References

Bishop, C. (1995). *Neural networks for pattern recognition.*Oxford University Press, USA.

Breiman, L. (1996). *Bias, variance and arcing classifiers* No. Tech. Report 460, Statistics Dept., University of California.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees.*Wadsworth.

Cristianini, N., & Shawe-Taylor, J. (2000). *An introduction to support vector machines and other kernel-based learning methods.*Cambridge University Press.

Egan, J. P. (1975). *Signal detection theory and ROC analysis.*Academic Press New York.

Everson, R. M., & Fieldsend, J. E. (2006). Multi-class ROC analysis from a multi-objective optimisation perspective. *Pattern Recognition Letters, Special Number on ROC Analysis in Pattern Recognition,*

Fawcett, T. (2004). ROC graphs: Notes and practical considerations for researchers. *Machine Learning, 31*

Fawcett, T., & Flach, P. A. (2005). A response to webb and Ting's on the application of ROC analysis to predict classification performance under varying class distributions. *Machine Learning, 58*(1), 33-38.

Flach, P. (2004). *Tutorial at ICML 2004: The many faces of ROC analysis in machine learning.* Unpublished manuscript.

Flach, P., Blockeel, H., Ferri, C., Hernandez-Orallo, J., & Struyf, J. (2003). Decision support for data mining: Introduction to ROC analysis and its applications. *Data mining and decision support: Aspects of integration and collaboration* (pp. 81-90)

Flach, P. A. (2003). The geometry of ROC space: Understanding machine learning metrics through ROC isometrics. *Proceedings of the Twentieth International Conference on Machine Learning,* 194–201.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction.*Springer.

Lane, T. (2000). Extensions of ROC analysis to multi-class domains. *ICML-2000 Workshop on Cost-Sensitive Learning,*

Macskassy, S., & Provost, F. (2004). Confidence bands for ROC curves: Methods and an empirical study. *Proceedings of the First Workshop on ROC Analysis in AI (ROCAI-2004) at ECAI-2004,* Spain.

Mitchell, T. M. (1997). *Machine learning*McGraw-Hill Higher Education.

Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. *Proceedings of the Fifteenth International Conference on Machine Learning,* , 445–453.

Sing, T., Sander, O., Beerenwinkel, N., & Lengauer, T. (2005). ROCR: Visualizing classifier performance in R. *Bioinformatics (Oxford, England), 21*(20), 3940-3941.

Srinivasan, A. (1999). *Note on the location of optimal classifiers in n-dimensional ROC space* (Technical Report No. PRG-TR-2-99). Oxford, England: Oxford University Computing Laboratory.

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Better decisions through science. *Scientific American; Scientific American, 283*(4), 82-87.

Key Terms and their Definitions

Receiver Operating Characteristic (ROC) Curve: A cost-benefit plot that describes the performance of a classification model.

Model Assessment/Evaluation: The process of evaluating the key performance characteristics of a classification model. This is usually done within the context of a problem domain with specific model performance requirements.

Confusion Matrix: A table that relates actual and predicted classifications by a model.

Class Skew: In probability theory and statistics skewness is a measure of asymmetry of a distribution. Class skew refers to the asymmetry of the class distribution.

Classification Model/Classifier: A mathematical construct such as a decision tree or neural network that models the relationship between independent and dependent variables of a classification problem. Once such a model has been constructed it can be used to predict classifications on new data instances.

Model Selection: The process of selecting a model from a set of potential models given a specific classification problem. The selection is usually based on a specific set of performance metrics dictated by the problem domain.

Performance Metric: A performance-related measurement.

Optimal Operating Point: The point on the ROC curve where a model has the largest true positive rate while committing the smallest number of false positives.