The Relationship Between Precision-Recall and ROC Curves

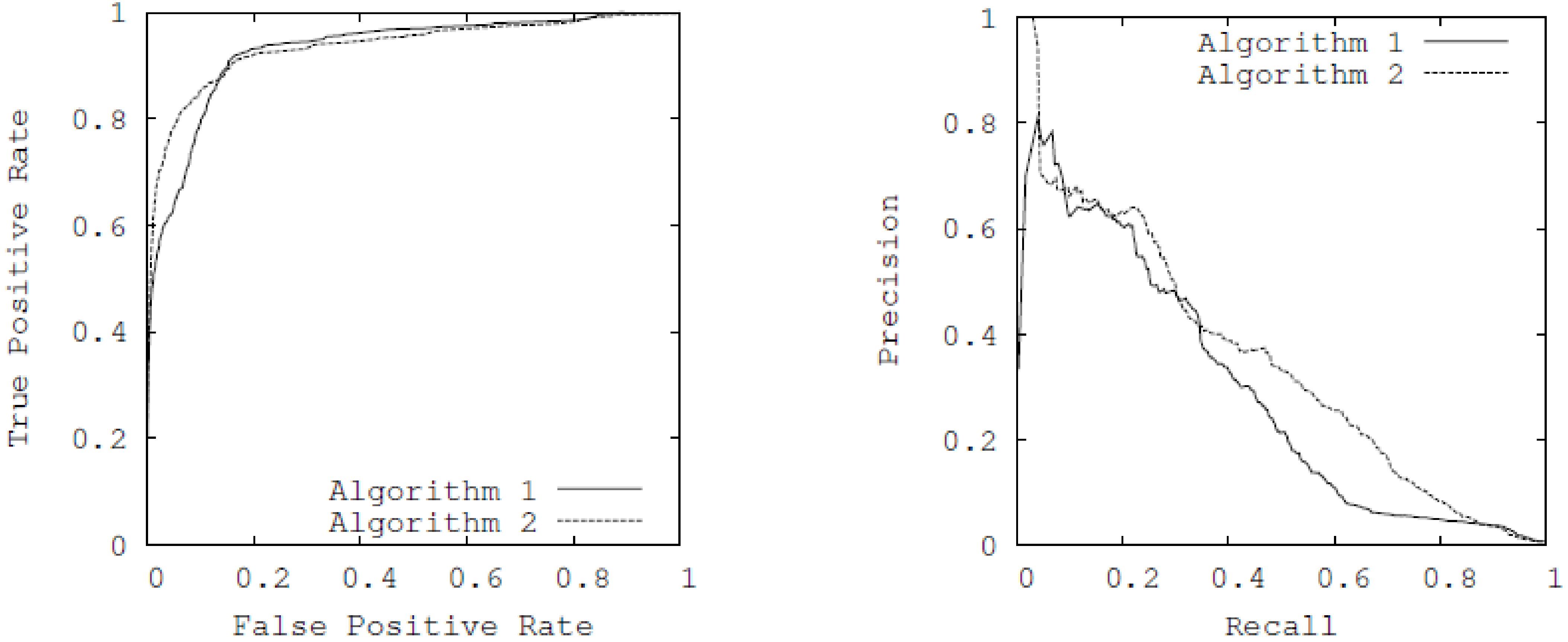
BY. F.NOORBEHBAHANI
FALL 2013

Abstract

- Receiver Operator Characteristic (ROC) curves are commonly used to present results for binary decision problems in machine learning.
- However, when dealing with highly skewed datasets, Precision-Recall (PR) curves give a more informative picture of an algorithm's performance.
- We show that a deep connection exists between ROC space and PR space, such that a curve dominates in ROC space if and only if it dominates in PR space.

- Provost et al. (1998) have argued that simply using accuracy results can be misleading.
- They recommended when evaluating binary decision problems to use Receiver Operator Characteristic (ROC) curves, which show how the number of correctly classified positive examples varies with the number of incorrectly classified negative examples.
- However, ROC curves can present an overly optimistic view of an algorithm's performance if there is a large skew.
- Drummond and Holte (2000; 2004) have recommended using cost curves to address this issue. Cost curves are an excellent alternative to ROC curves, but discussing them is beyond the scope of this paper.

- Precision-Recall (PR) curves, often used in Information Retrieval (Manning & Schutze, 1999; Raghavan et al., 1989), have been cited as an alternative to ROC curves for tasks with a large skew in the class distribution.
- An important difference between ROC space and PR space is the visual representation of the curves.
- Looking at PR curves can expose differences between algorithms that are not apparent in ROC space.



(a) Comparison in ROC space

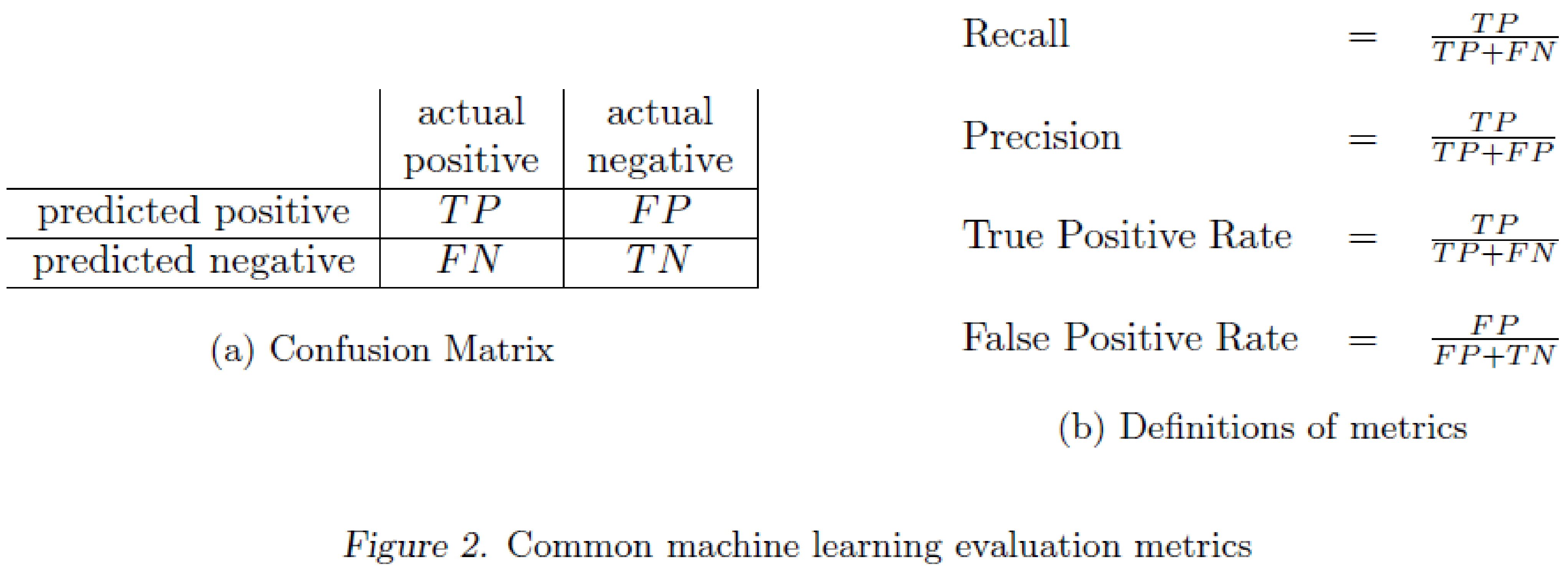
(b) Comparison in PR space

Figure 1. The difference between comparing algorithms in ROC vs PR space

- The performances of the algorithms appear to be comparable in ROC space, however, in PR space we can see that Algorithm 2 has a clear advantage over Algorithm 1.
- This difference exists because in this domain the number of negative examples greatly exceeds the number of positives examples. Consequently, a large change in the number of false positives can lead to a small change in the false positive rate used in ROC analysis.
- Precision, on the other hand, by comparing false positives to true positives rather than true negatives, captures the effect of the large number of negative examples on the algorithm's performance.

Review of ROC and Precision-Recall

- In a binary decision problem, a classifier labels examples as either positive or negative. The decision made by the classifier can be represented in a structure known as a confusion matrix or contingency table.
- The confusion matrix has four categories: True positives (TP) are examples correctly labeled as positives.
- False positives (FP) refer to negative examples incorrectly labeled as positive. True negatives (TN) correspond to negatives correctly labeled as negative.
- Finally, false negatives (FN) refer to positive examples incorrectly labeled as negative.



	actual positive	actual negative
predicted positive	TP	FP
predicted negative	FN	TN

(a) Confusion Matrix

$$\begin{aligned}
\text{Recall} &= \frac{TP}{TP+FN} \\
\text{Precision} &= \frac{TP}{TP+FP} \\
\text{True Positive Rate} &= \frac{TP}{TP+FN} \\
\text{False Positive Rate} &= \frac{FP}{FP+TN}
\end{aligned}$$

(b) Definitions of metrics

Figure 2. Common machine learning evaluation metrics

$$\text{fp rate} = \frac{FP}{N}$$

$$\text{tp rate} = \frac{TP}{P}$$

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

$$\text{recall} = \frac{TP}{P}$$

$$\text{accuracy} = \frac{TP+TN}{P+N}$$

$$\text{F-measure} = \frac{2}{1/\text{precision}+1/\text{recall}}$$

The Inaccuracy of Accuracy

- A tacit assumption in the use of classification accuracy :
  - class distribution among examples
    - is constant
    - and relatively balanced

- As the class distribution becomes more skewed evaluation based on accuracy breaks down

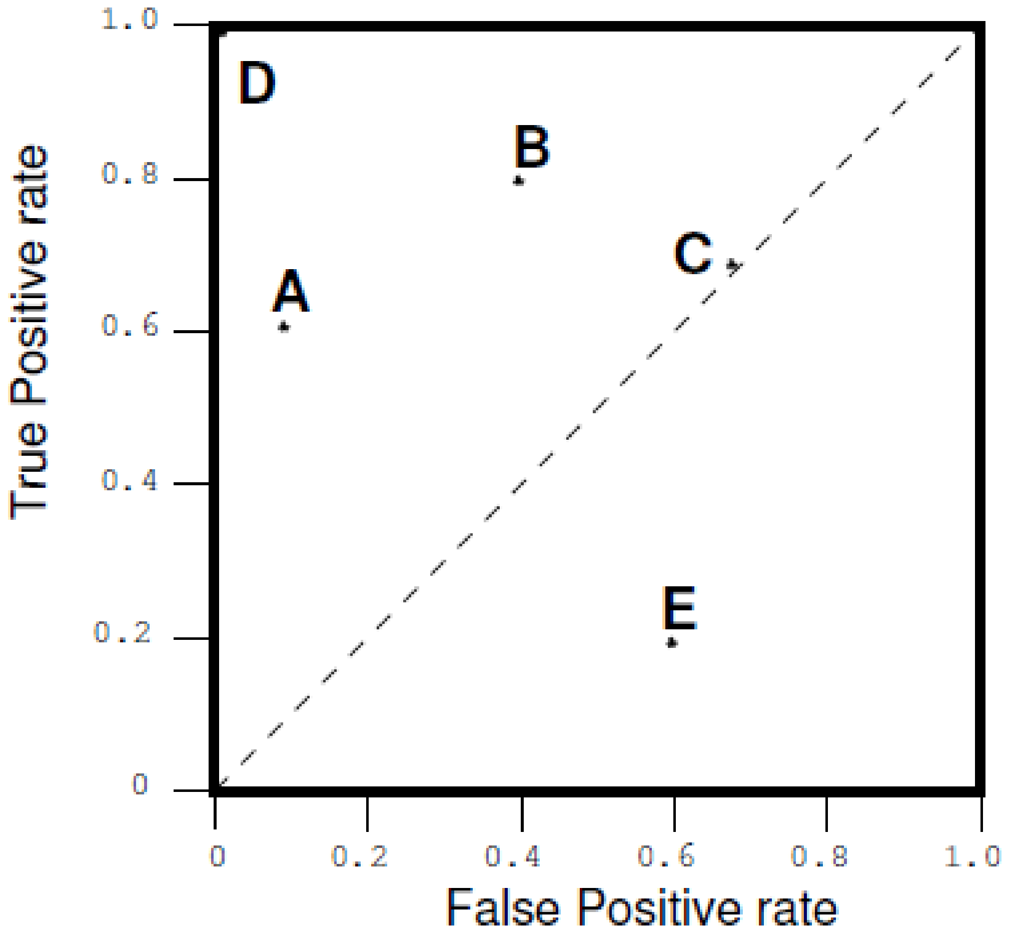
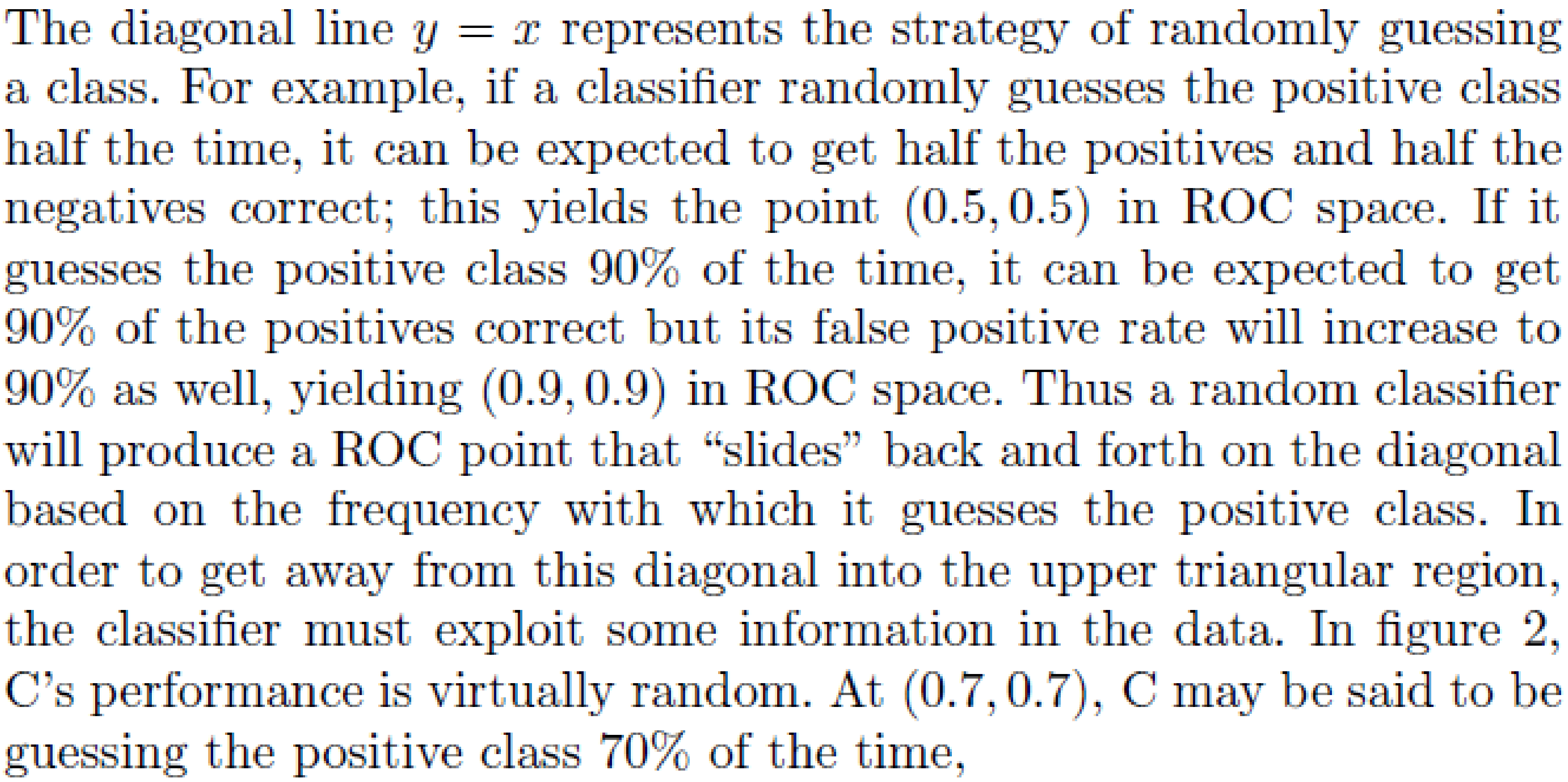


Figure 2. A basic ROC graph showing five discrete classifiers.

- Any classier that appears in the lower right triangle performs worse than random guessing. This triangle is therefore usually empty in ROC graphs.
- Given an ROC graph in which a classifier's performance appears to be slightly better than random, it is natural to ask “is this classifier's performance truly significant or is it only better than random by chance?"
- There is no conclusive test for this, but Forman (2002) has shown a methodology that addresses this question with ROC curves.

Random Performance



The diagonal line $y = x$ represents the strategy of randomly guessing a class. For example, if a classifier randomly guesses the positive class half the time, it can be expected to get half the positives and half the negatives correct; this yields the point $(0.5, 0.5)$ in ROC space. If it guesses the positive class 90% of the time, it can be expected to get 90% of the positives correct but its false positive rate will increase to 90% as well, yielding $(0.9, 0.9)$ in ROC space. Thus a random classifier will produce a ROC point that "slides" back and forth on the diagonal based on the frequency with which it guesses the positive class. In order to get away from this diagonal into the upper triangular region, the classifier must exploit some information in the data. In figure 2, C's performance is virtually random. At $(0.7, 0.7)$, C may be said to be guessing the positive class 70% of the time,

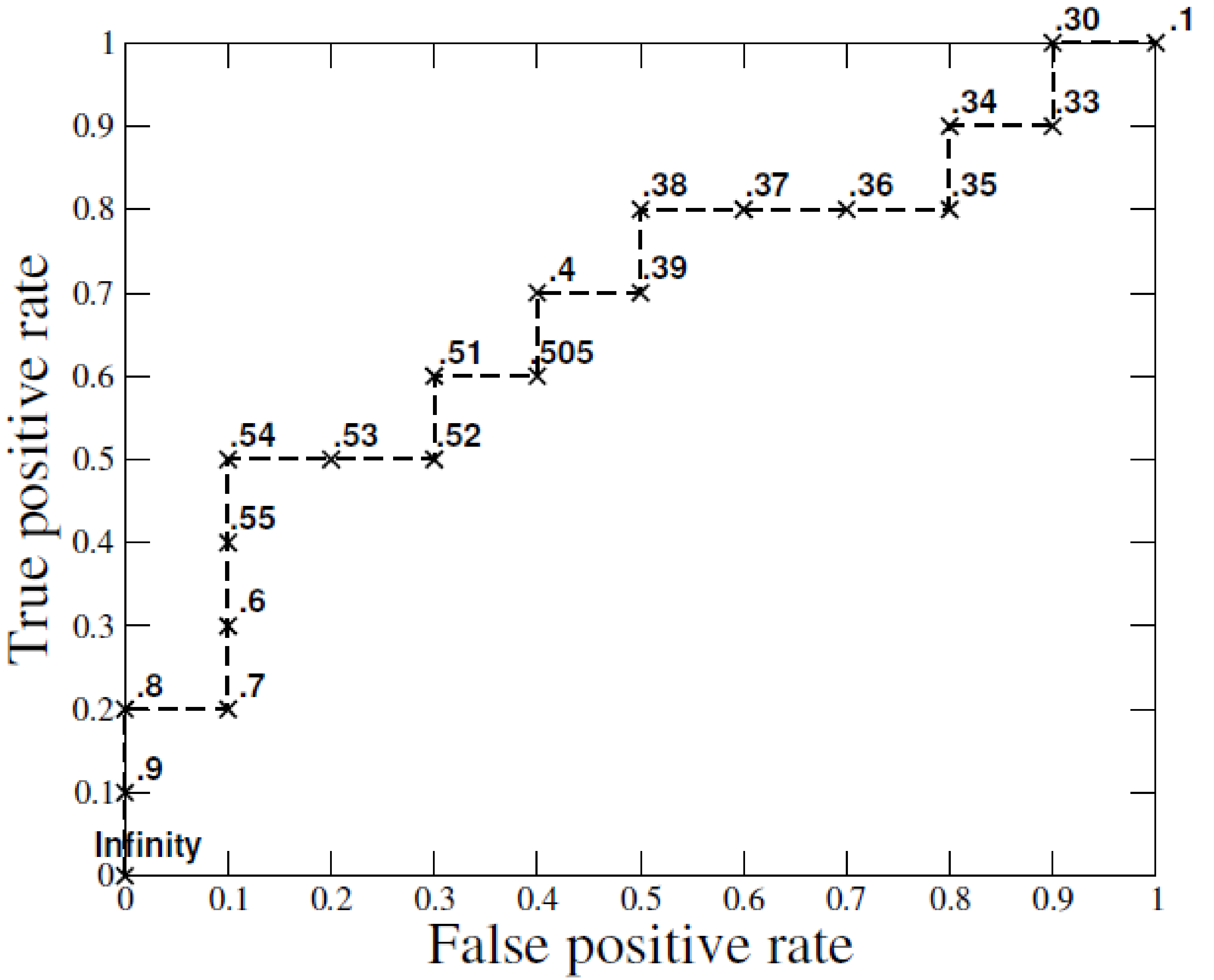
Curves in ROC space

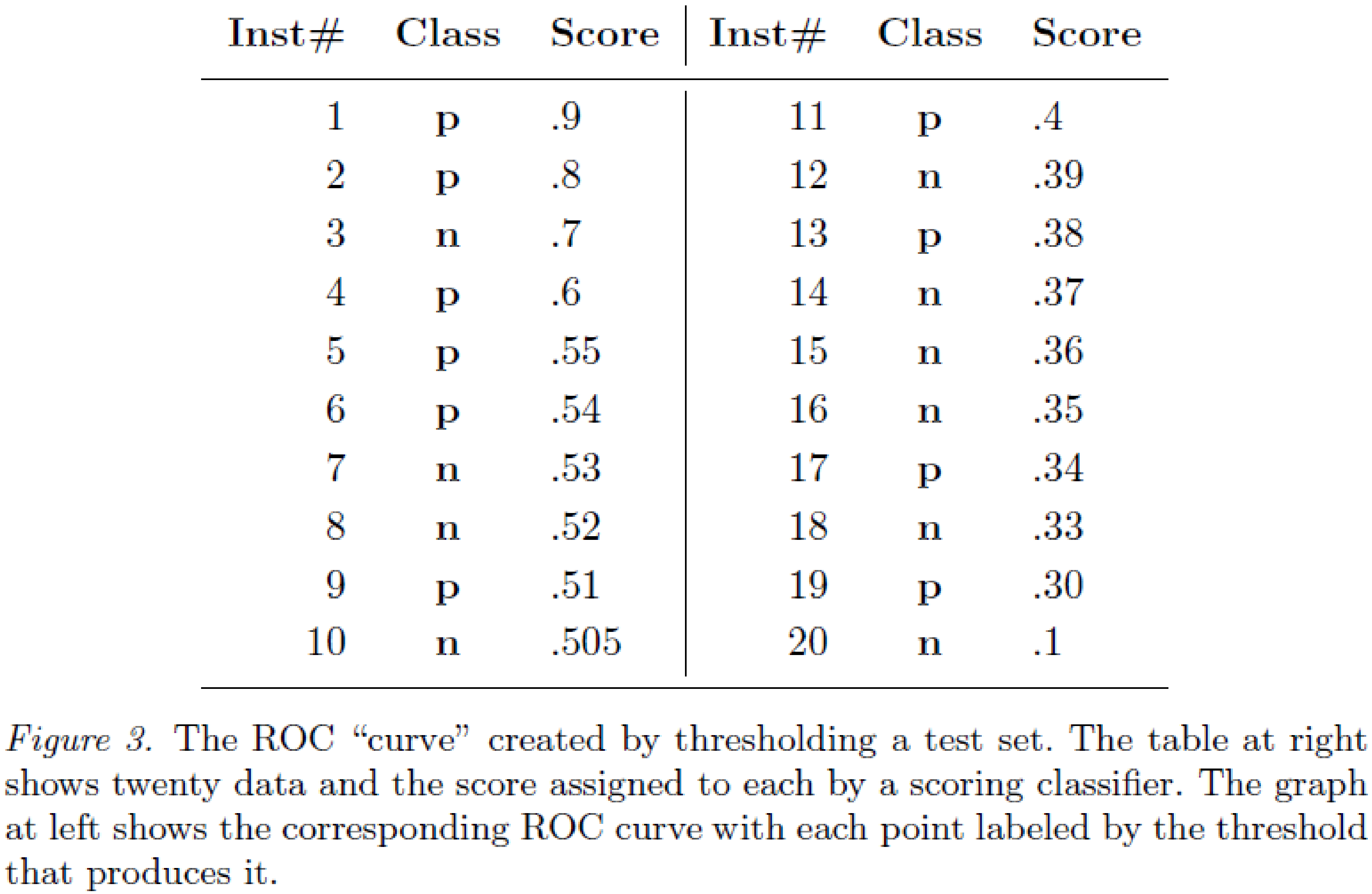
- Many classifiers, such as decision trees or rule sets, are designed to produce only a class decision, i.e., a Y or N on each instance.
- When such a discrete classier is applied to a test set, it yields a single confusion matrix, which in turn corresponds to one ROC point.
- Thus, a discrete classier produces only a single point in ROC space.

Curves in ROC space

- Some classifiers, such as a Naive Bayes classier or a neural network, naturally yield an instance probability or score, a numeric value that represents the degree to which an instance is a member of a class.
- These values can be strict probabilities, in which case they adhere to standard theorems of probability; or they can be general, uncalibrated scores, in which case the only property that holds is that a higher score indicates a higher probability.
- We shall call both a probabilistic classier, in spite of the fact that the output may not be a proper probability.

- Such a ranking or scoring classifier can be used with a threshold to produce a discrete (binary) classifier: if the classifier output is above the threshold, the classifier produces a Y, else a N.
- Each threshold value produces a different point in ROC space.
- Computationally, this is a poor way of generating an ROC curve, and the next section describes a more efficient and careful method.


True positive rate
False positive rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Infinity
.9
.8
.7
.6
.55
.54
.53
.52
.51
.505
.4
.39
.38
.37
.36
.35
.34
.33
.30
.1

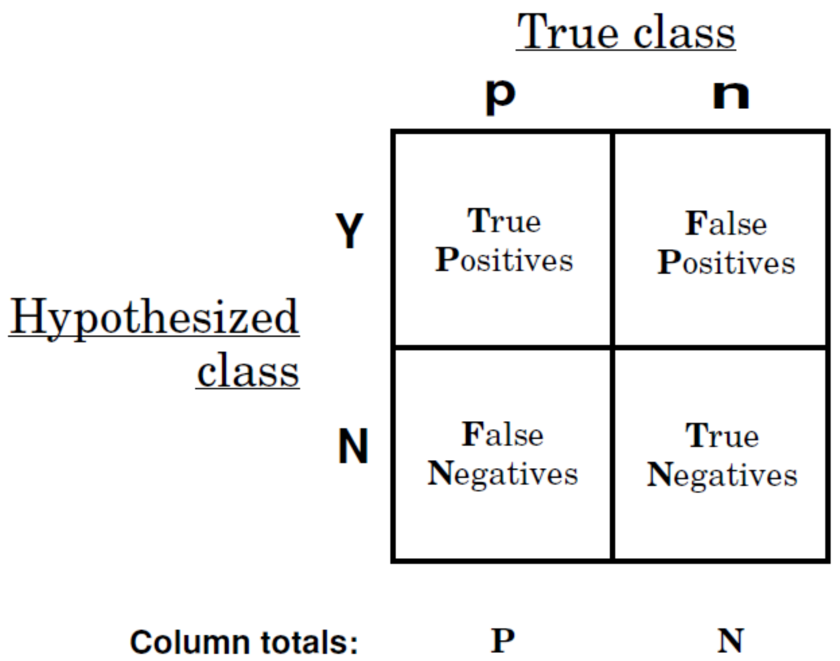


Inst#	Class	Score
1	p	.9
2	p	.8
3	n	.7
4	p	.6
5	p	.55
6	p	.54
7	n	.53
8	n	.52
9	p	.51
10	n	.505
11	p	.4
12	n	.39
13	p	.38
14	n	.37
15	n	.36
16	n	.35
17	p	.34
18	n	.33
19	p	.30
20	n	.1

Figure 3. The ROC "curve" created by thresholding a test set. The table at right shows twenty data and the score assigned to each by a scoring classifier. The graph at left shows the corresponding ROC curve with each point labeled by the threshold that produces it.

Class skew

- ROC curves have an attractive property: they are insensitive to changes in class distribution. If the proportion of positive to negative instances changes in a test set, the ROC curves will not change.

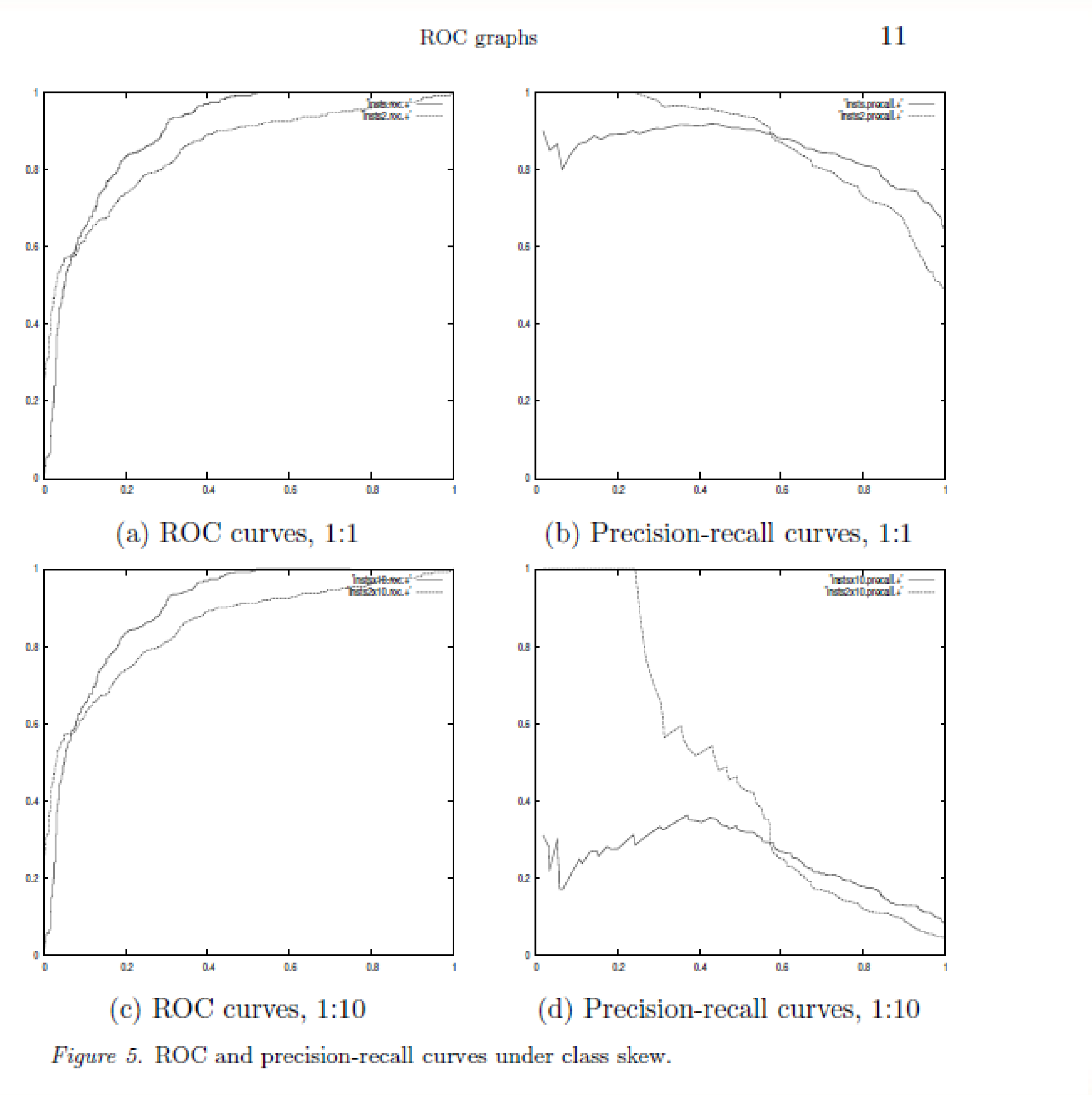


		True class p	 n
Hypothesized class	Y	True Positives	False Positives
	N	False Negatives	True Negatives
Column totals:		P	N

$$\text{fp rate} = \frac{FP}{N} \qquad \text{tp rate} = \frac{TP}{P}$$

$$\text{precision} = \frac{TP}{TP+FP} \qquad \text{recall} = \frac{TP}{P}$$

$$\text{accuracy} = \frac{TP+TN}{P+N} \qquad \text{F-measure} = \frac{2}{1/\text{precision}+1/\text{recall}}$$



ROC graphs 11

(a) ROC curves, 1:1

(b) Precision-recall curves, 1:1

(c) ROC curves, 1:10

(d) Precision-recall curves, 1:10

Figure 5. ROC and precision-recall curves under class skew.

Creating scoring classifiers

- Many classier models are discrete: they are designed to produce only a class label from each test instance.
- However, we often want to generate a full ROC curve from a classier instead of just a single point. To this end we want to generate scores from a classier rather than just a class label. There are several ways of producing such scores.
- Many discrete classier models may easily be converted to scoring classifiers by looking inside" them at the instance statistics they keep.
- For example, a decision tree determines a class label of a leaf node from the proportion of instances at the node; the class decision is simply the most prevalent class. These class proportions may serve as a score

Creating scoring classifiers

- A rule learner keeps similar statistics on rule confidence, and the confidence of a rule matching an instance can be used as a score
- Even if a classier only produces a class label, an aggregation of them may be used to generate a score. MetaCost (Domingos, 1999) employs bagging to generate an ensemble of discrete classifiers, each of which produces a vote. The set of votes could be used to generate a score
- Finally, some combination of scoring and voting can be employed. For example, rules can provide basic probability estimates, which may then be used in weighted voting.

Area under an ROC Curve (AUC)

- An ROC curve is a two-dimensional depiction of classier performance. To compare classifiers we may want to reduce ROC performance to a single scalar value representing expected performance.
- A common method is to calculate the area under the ROC curve, abbreviated AUC .

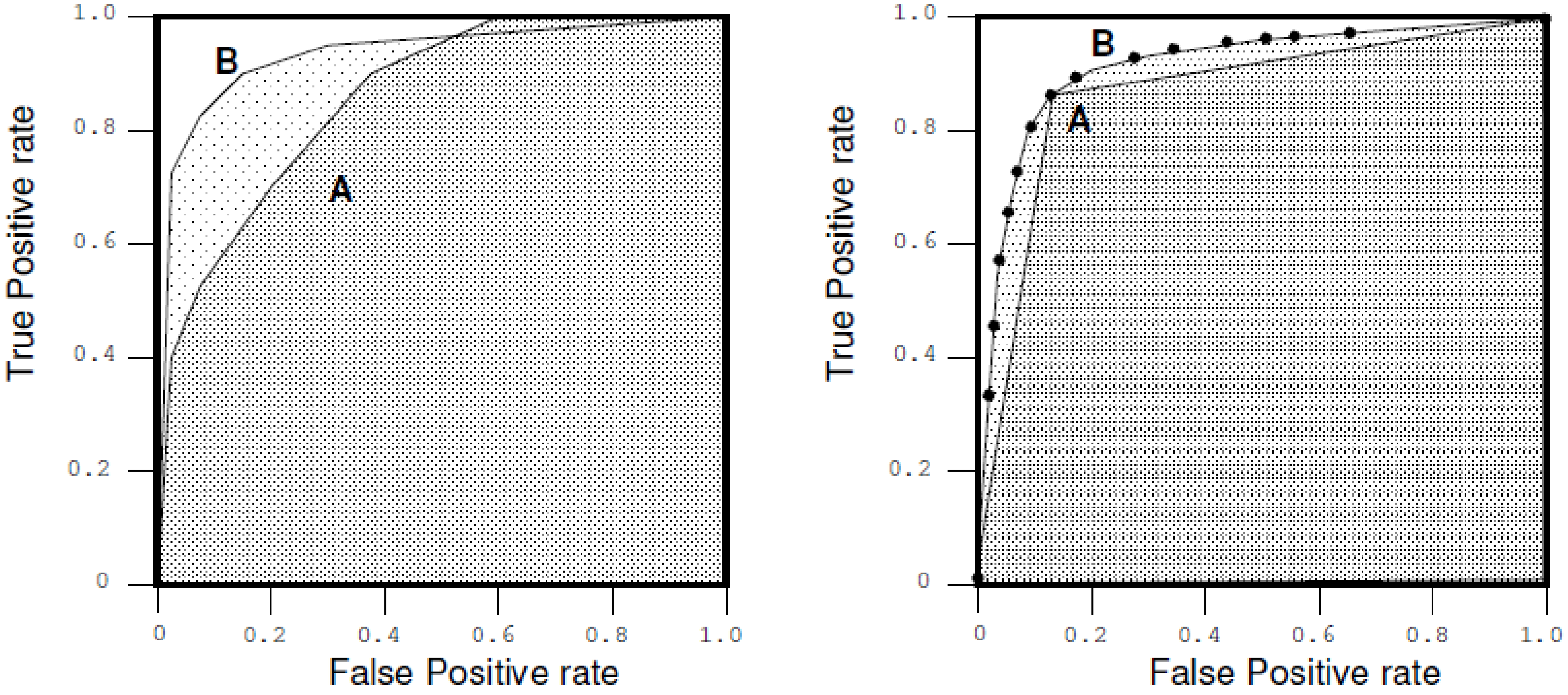
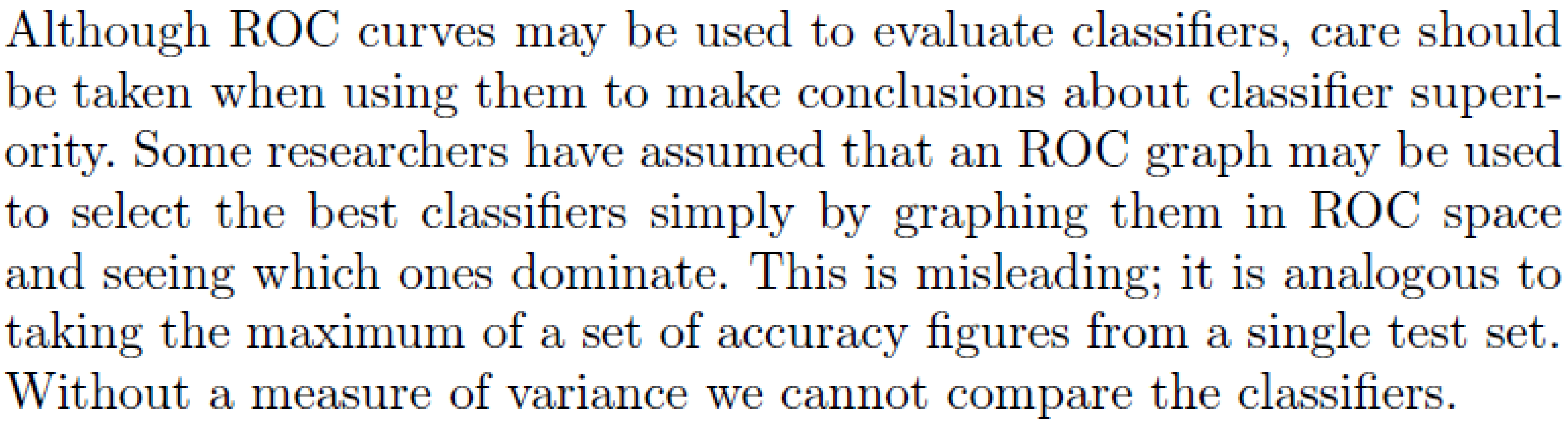


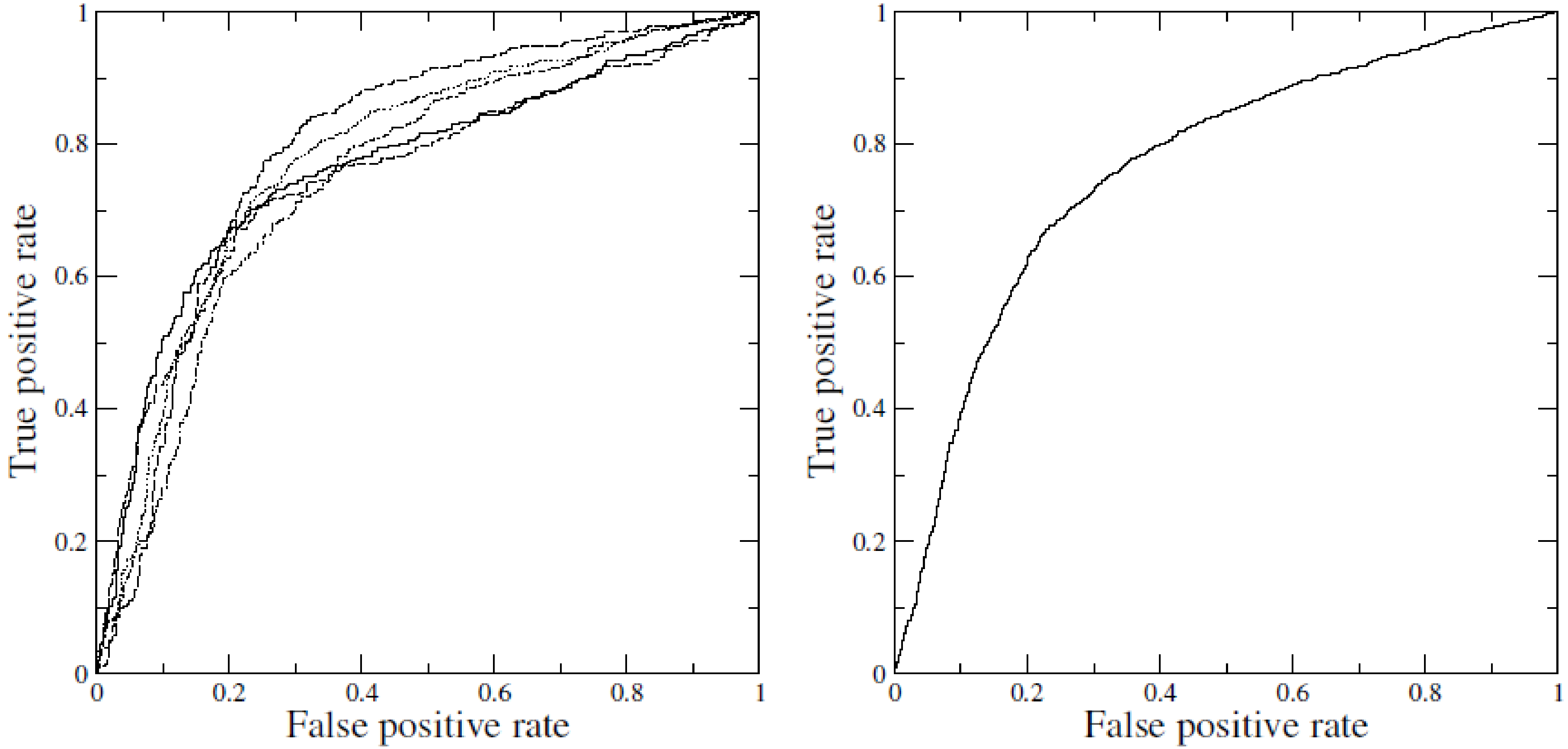
Figure 7. Two ROC graphs. The graph on the left shows the area under two ROC curves. The graph on the right shows the area under the curves of a discrete classifier (A) and a probabilistic classifier (B).

Averaging ROC curves



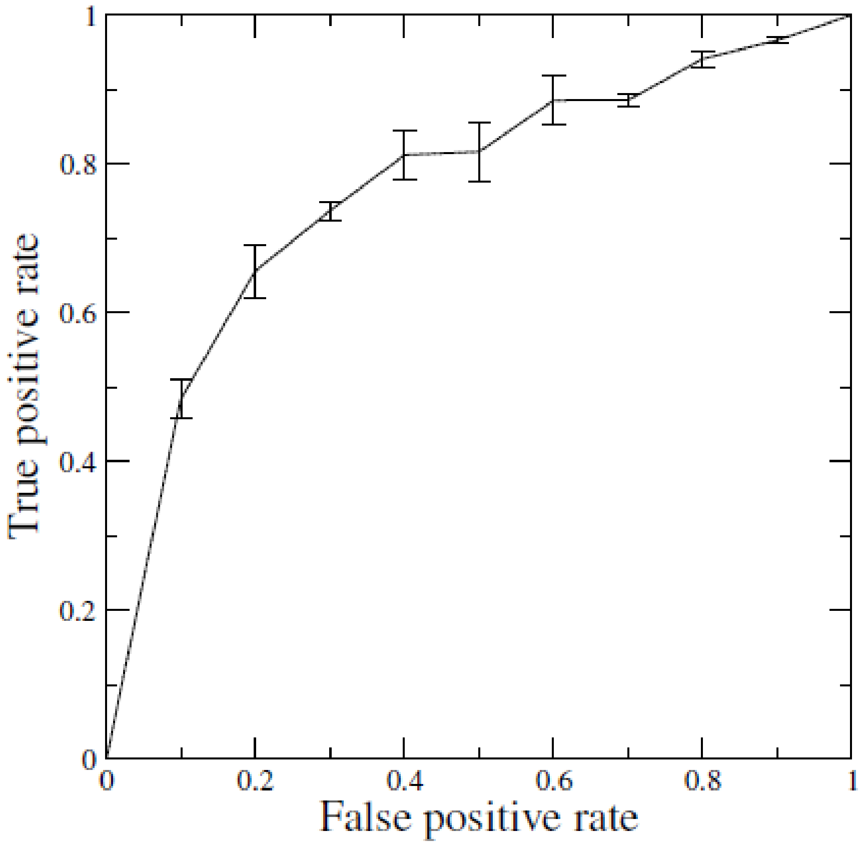
Although ROC curves may be used to evaluate classifiers, care should be taken when using them to make conclusions about classifier superiority. Some researchers have assumed that an ROC graph may be used to select the best classifiers simply by graphing them in ROC space and seeing which ones dominate. This is misleading; it is analogous to taking the maximum of a set of accuracy figures from a single test set. Without a measure of variance we cannot compare the classifiers.

- two methods for averaging ROC curves:
  - vertical
  - threshold averaging.

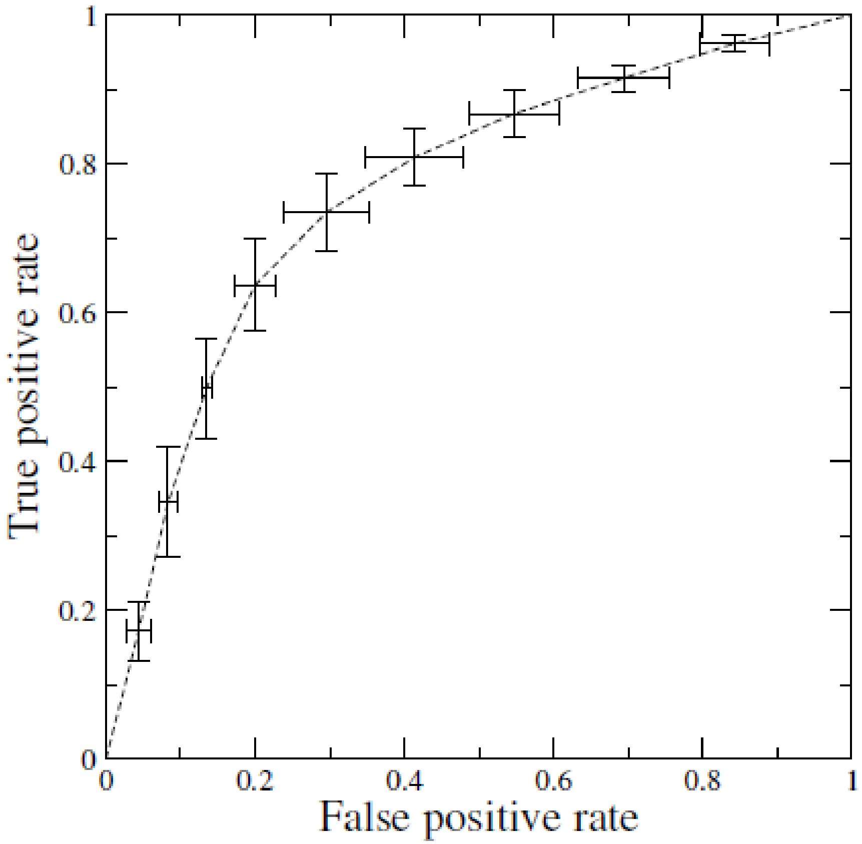


(a) ROC curves of five instance samples

(b) ROC curve formed by merging the five samples



(c) The curves of a averaged vertically



(d) The curves of a averaged by threshold

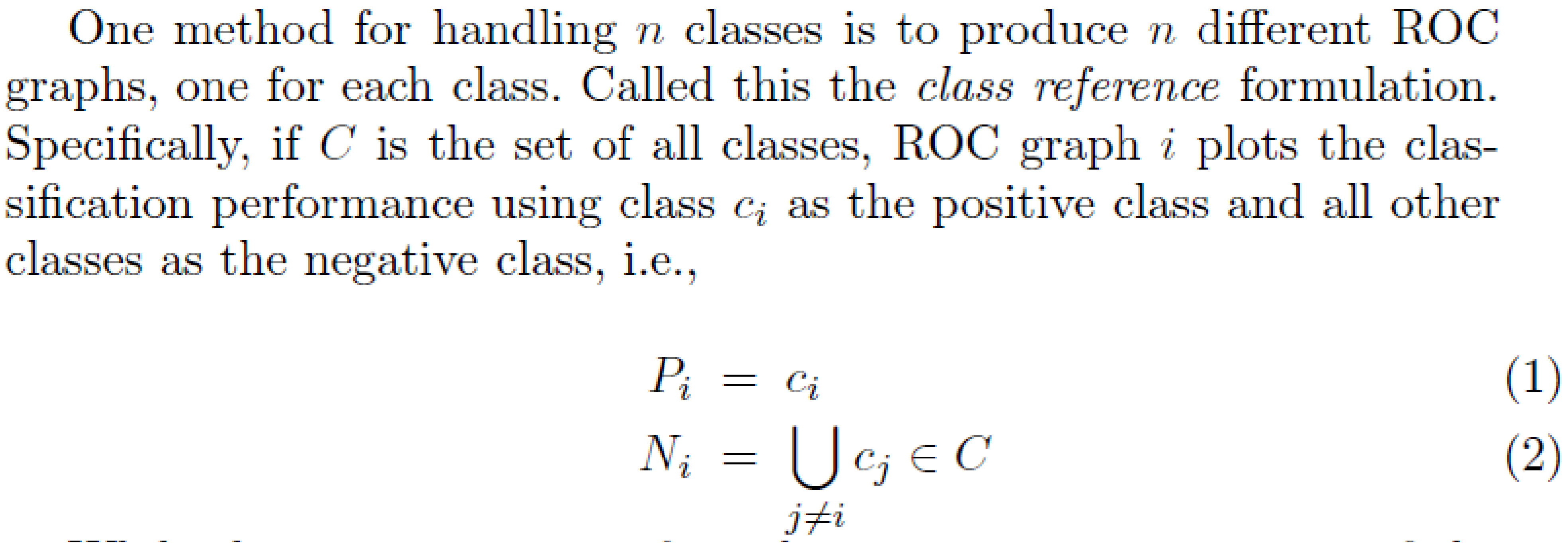
Vertical averaging

- Vertical averaging takes vertical samples of the ROC curves for fixed FP rates and averages the corresponding TP rates.
- Such averaging is appropriate when the FP rate can indeed be fixed by the researcher, or when a single-dimensional measure of variation is desired.

Threshold averaging

- Instead of sampling points based on their positions in ROC space, as vertical averaging does, it samples based on the thresholds that produced these points.
- The method must generate a set of thresholds to sample, then for each threshold it finds the corresponding point of each ROC curve and averages them.

Multi-class ROC graphs

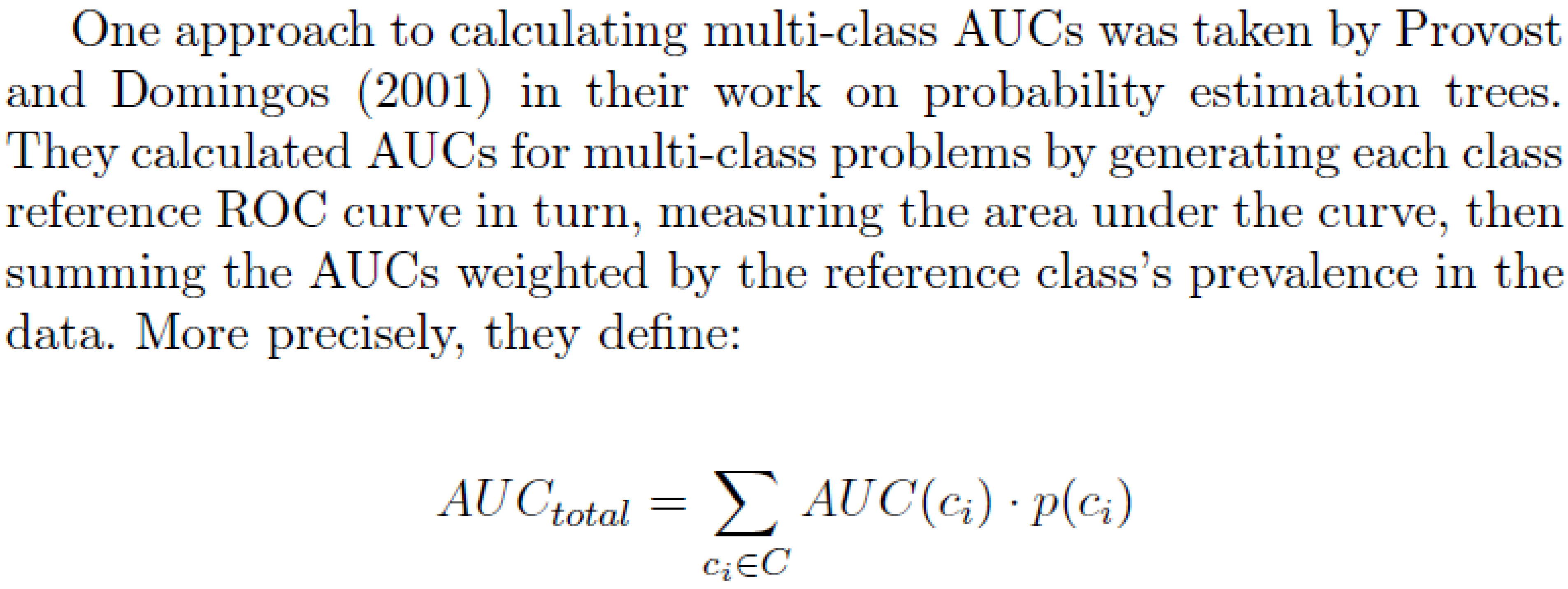


One method for handling n classes is to produce n different ROC graphs, one for each class. Called this the *class reference* formulation. Specifically, if C is the set of all classes, ROC graph i plots the classification performance using class c_i as the positive class and all other classes as the negative class, i.e.,

$$P_i = c_i \tag{1}$$

$$N_i = \bigcup_{j \neq i} c_j \in C \tag{2}$$

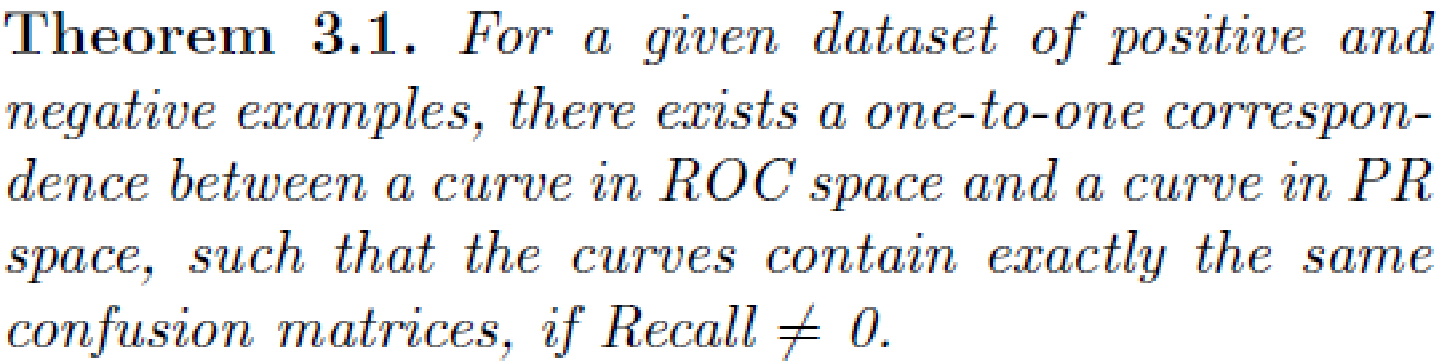
Multi-class AUC



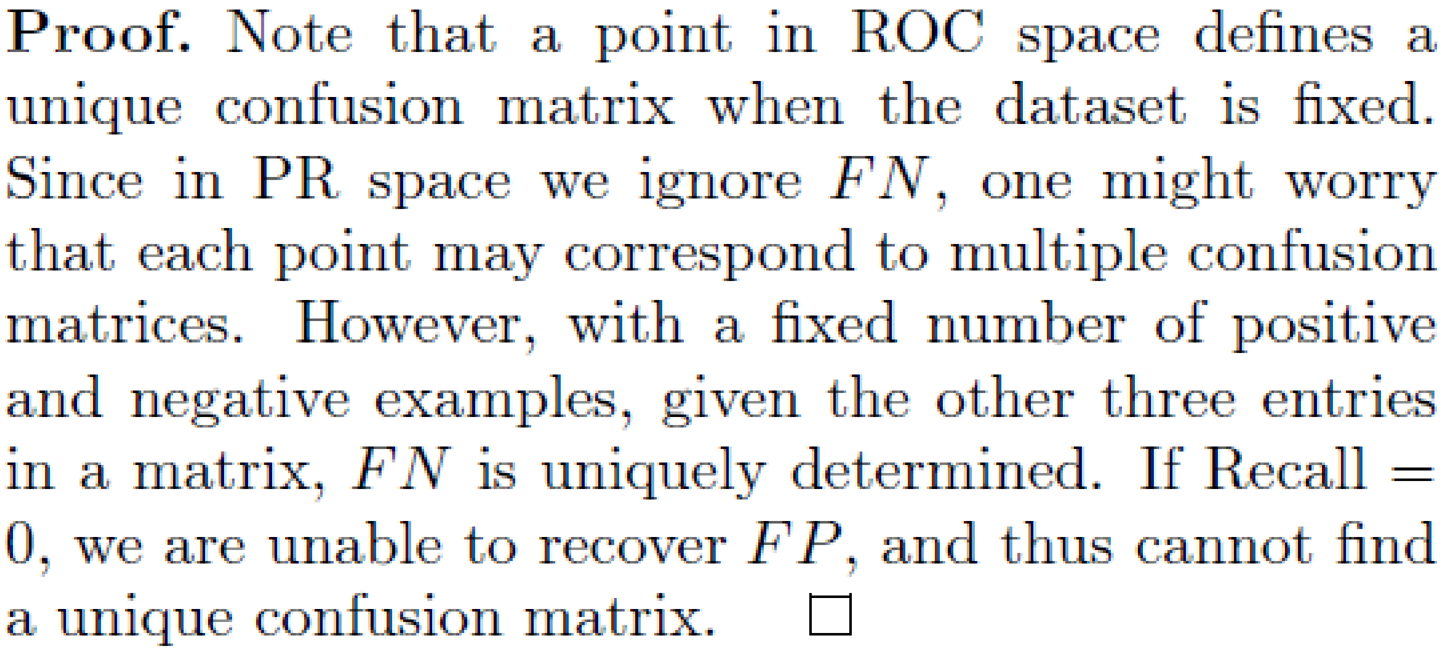
One approach to calculating multi-class AUCs was taken by Provost and Domingos (2001) in their work on probability estimation trees. They calculated AUCs for multi-class problems by generating each class reference ROC curve in turn, measuring the area under the curve, then summing the AUCs weighted by the reference class's prevalence in the data. More precisely, they define:

$$AUC_{total} = \sum_{c_i \in C} AUC(c_i) \cdot p(c_i)$$

Relationship between ROC Space and PR Space

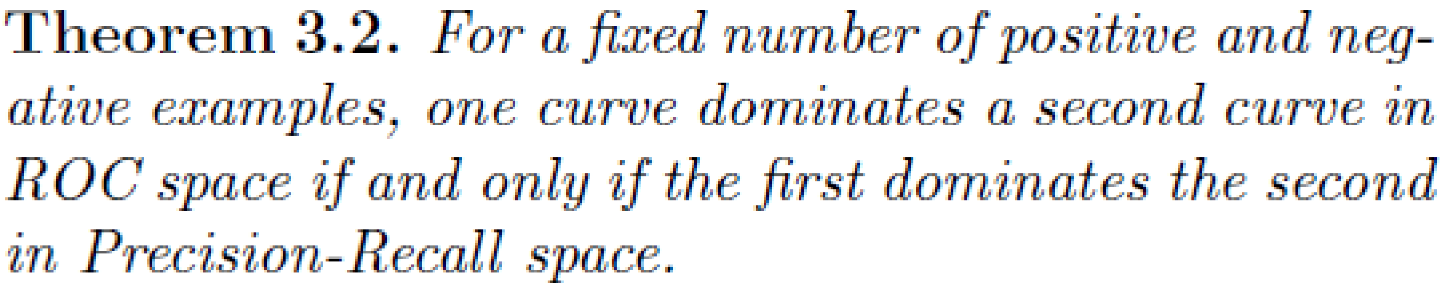


Theorem 3.1. *For a given dataset of positive and negative examples, there exists a one-to-one correspondence between a curve in ROC space and a curve in PR space, such that the curves contain exactly the same confusion matrices, if Recall* $\neq$ *0.*

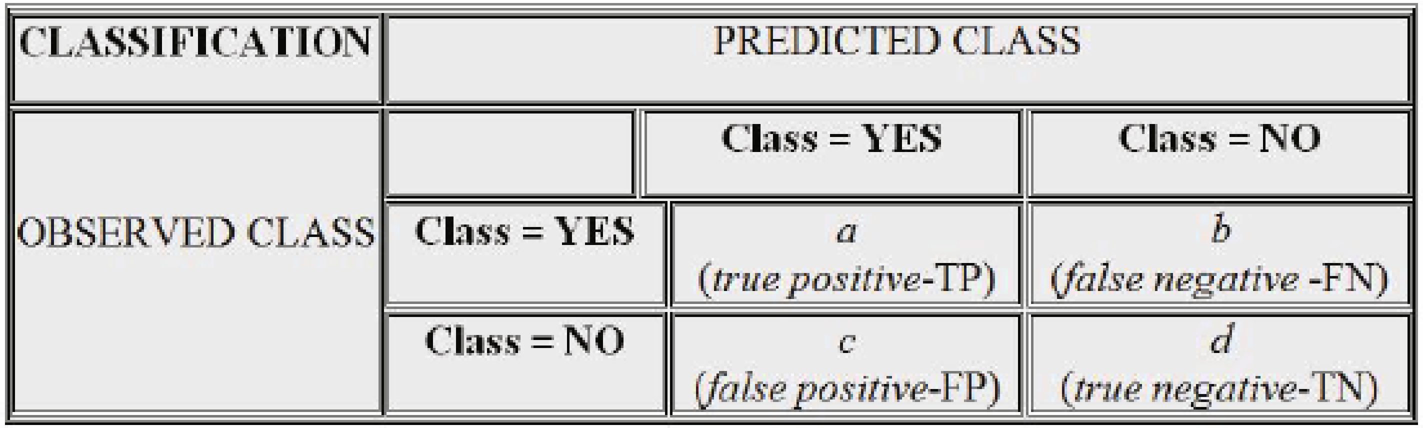


Proof. Note that a point in ROC space defines a unique confusion matrix when the dataset is fixed. Since in PR space we ignore FN, one might worry that each point may correspond to multiple confusion matrices. However, with a fixed number of positive and negative examples, given the other three entries in a matrix, FN is uniquely determined. If Recall = 0, we are unable to recover FP, and thus cannot find a unique confusion matrix. □

- One important definition we need for our next theorem is the notion that one curve dominates another curve, \meaning that all other...curves are beneath it or equal to it.

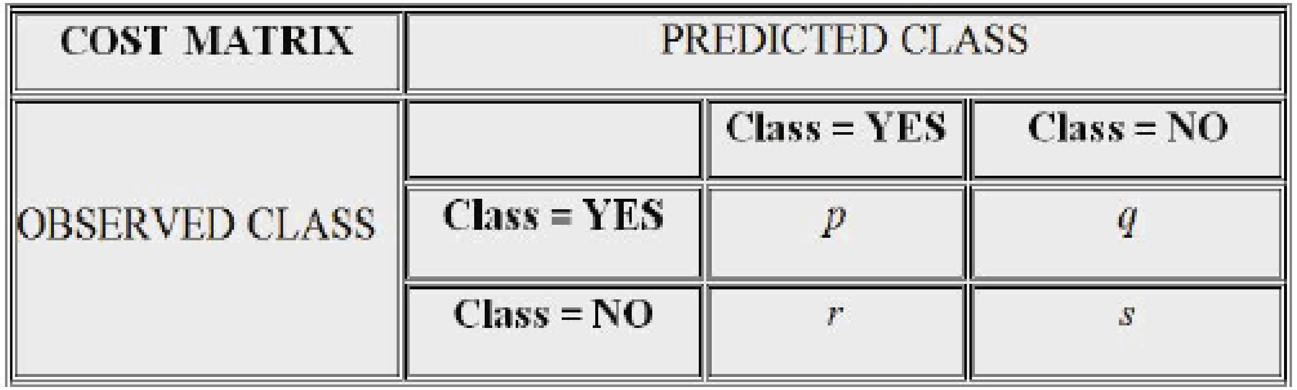


Theorem 3.2. *For a fixed number of positive and negative examples, one curve dominates a second curve in ROC space if and only if the first dominates the second in Precision-Recall space.*



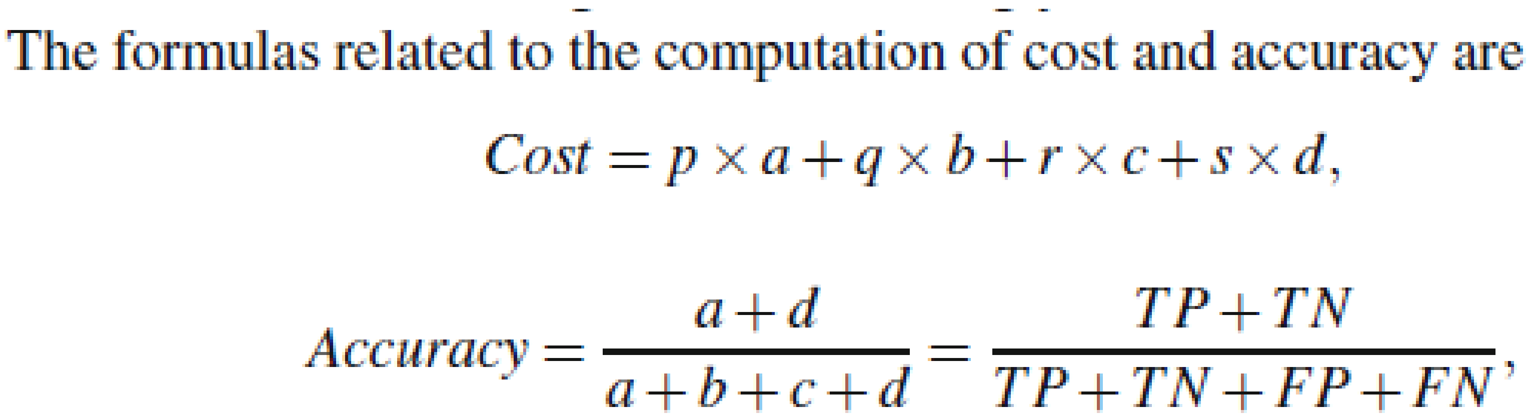
CLASSIFICATION	PREDICTED CLASS		
OBSERVED CLASS		Class = YES	Class = NO
	Class = YES	*a* (*true positive*-TP)	*b* (*false negative* -FN)
	Class = NO	*c* (*false positive*-FP)	*d* (*true negative*-TN)

Fig. 6.1 Confusion matrix for a 2-class model



COST MATRIX	PREDICTED CLASS		
OBSERVED CLASS		Class = YES	Class = NO
	Class = YES	*p*	*q*
	Class = NO	*r*	*s*

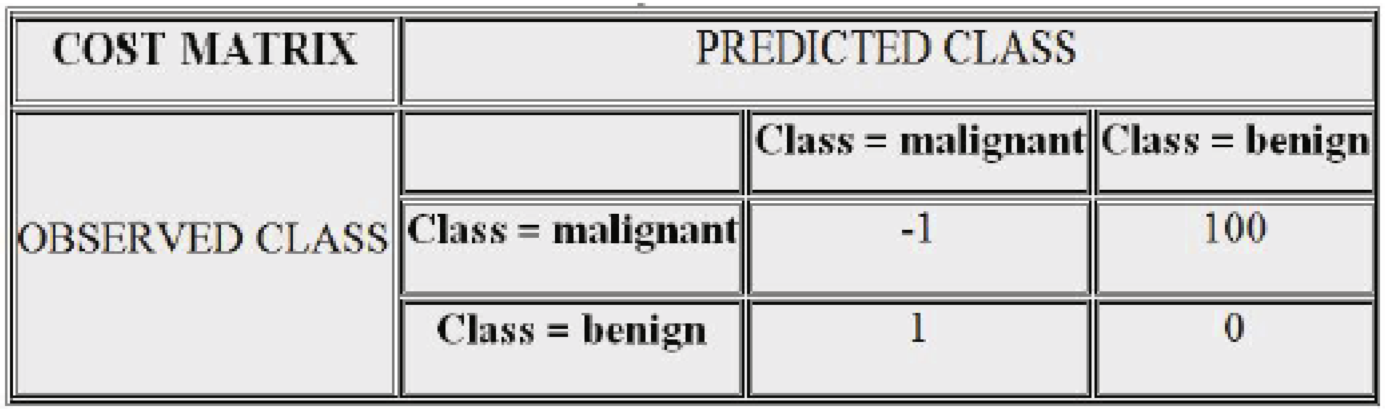
Fig. 6.2 Cost matrix for a 2-class model



The formulas related to the computation of cost and accuracy are the following:

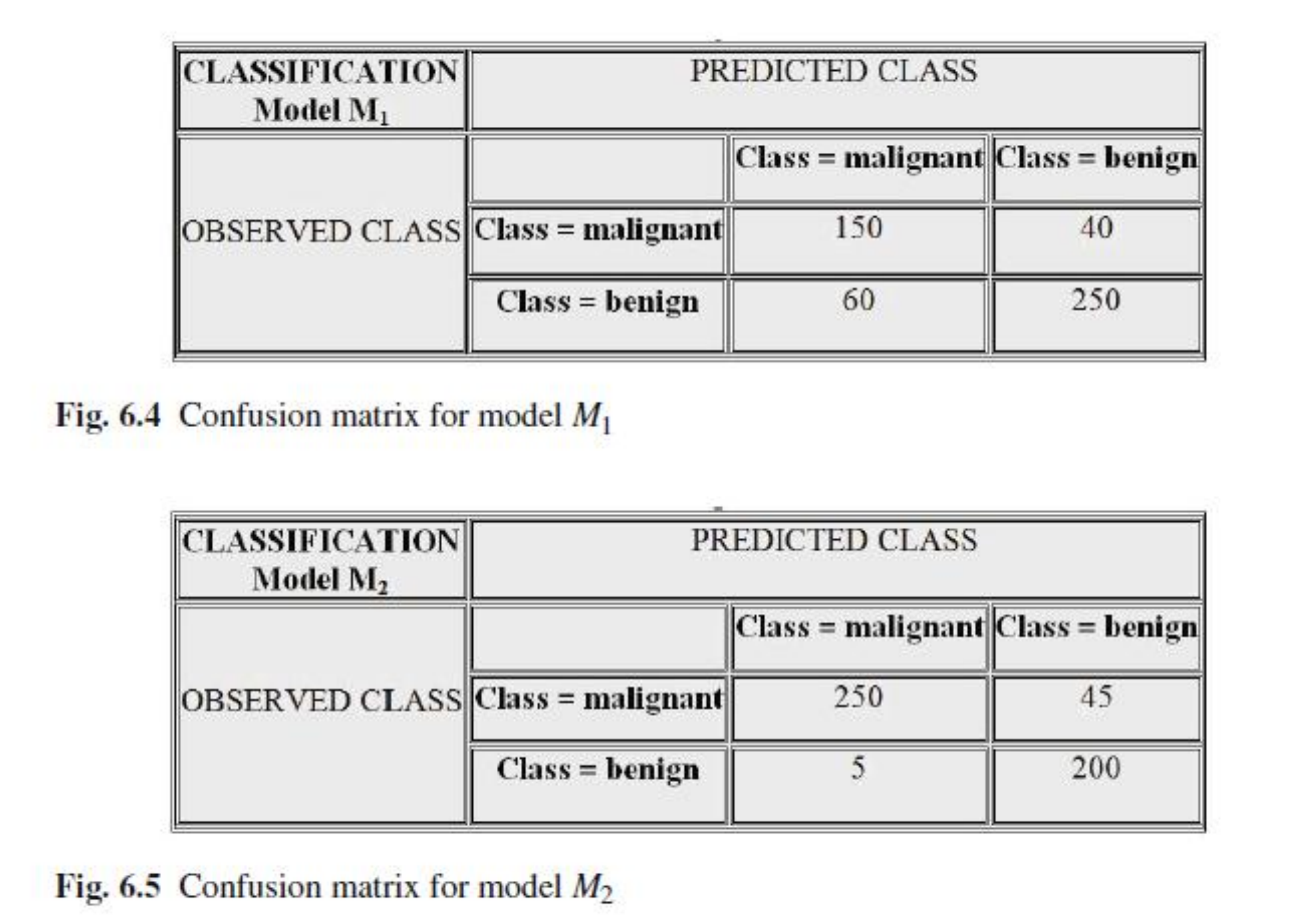
$$Cost = p \times a + q \times b + r \times c + s \times d,$$

$$Accuracy = \frac{a+d}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN},$$



COST MATRIX	PREDICTED CLASS		
OBSERVED CLASS		Class = malignant	Class = benign
	Class = malignant	-1	100
	Class = benign	1	0

Fig. 6.3 Cost matrix



CLASSIFICATION Model M_1	PREDICTED CLASS		
OBSERVED CLASS		Class = malignant	Class = benign
	Class = malignant	150	40
	Class = benign	60	250

Fig. 6.4 Confusion matrix for model M_1

CLASSIFICATION Model M_2	PREDICTED CLASS		
OBSERVED CLASS		Class = malignant	Class = benign
	Class = malignant	250	45
	Class = benign	5	200

Fig. 6.5 Confusion matrix for model M_2

$$Cost_{M_1} = -1 \times 150 + 100 \times 40 + 1 \times 60 + 0 = 3910,$$

$$Cost_{M_2} = -1 \times 250 + 100 \times 45 + 1 \times 5 + 0 = 4255,$$

$$Accuracy_{M_1} = \frac{400}{500} = 80\%,$$

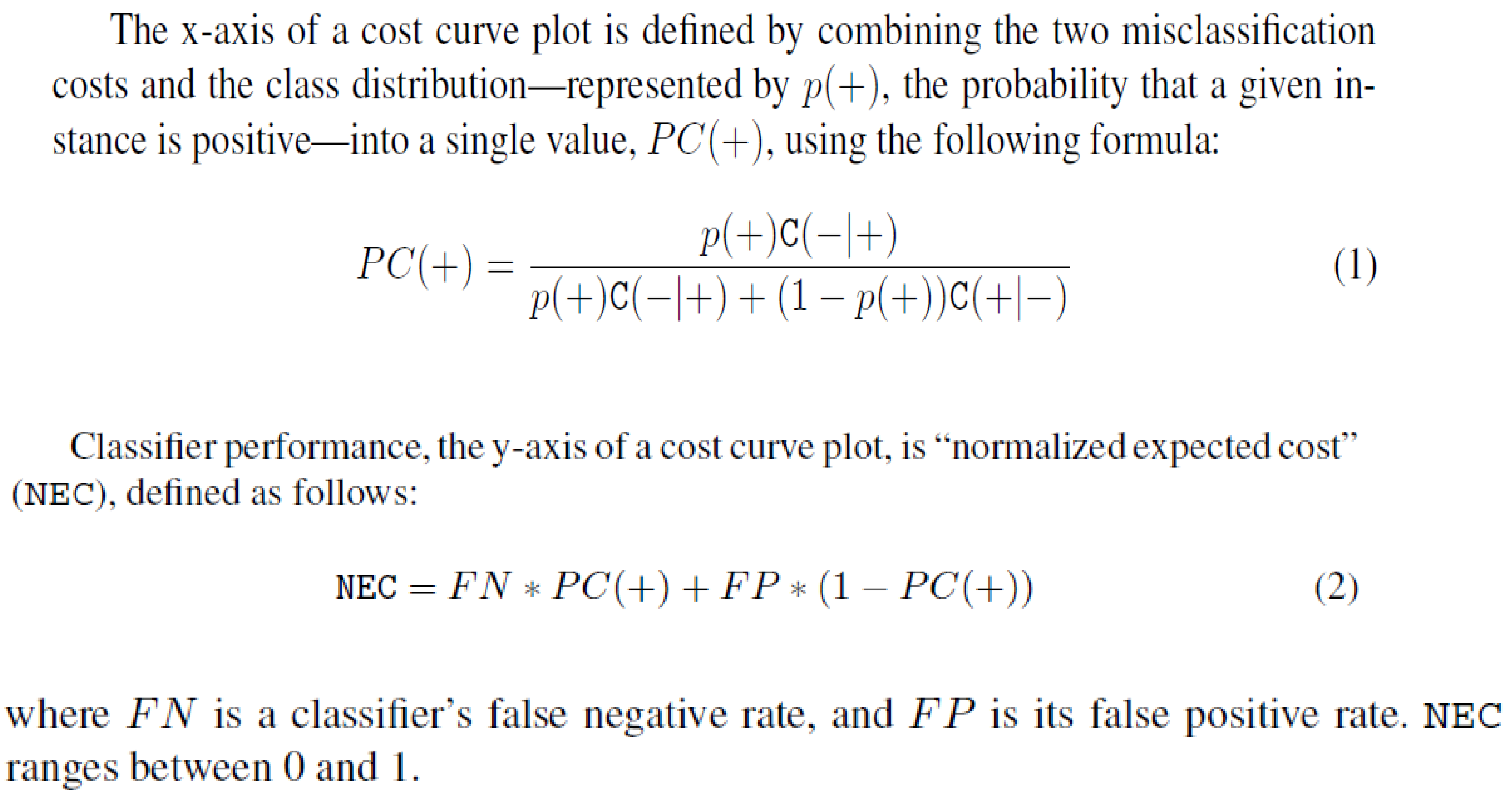
$$Accuracy_{M_2} = \frac{450}{500} = 90\%.$$

Cost-sensitive Classifier Evaluation using Cost Curves

- Methods for evaluating the performance of classifiers fall into two broad categories:
  - Numerical : produce a single number summarizing a classifier's performance
    - accuracy, expected cost, precision, recall, and area under a performance curve (AUC).
- recall, and area under a performance curve (AUC)
  - Graphical: depict performance in a plot that typically has just two or three dimensions so that it can be easily inspected by humans
    - ROC curves, precision-recall curves, DET curves , regret graphs , loss difference plots , skill plots , prevalence-value-accuracy plots, and the method presented in this talk, cost curves .
- Graphical methods are especially useful when there is uncertainty about the misclassification costs or the class distribution that will occur when the classier is deployed.

- In this setting, graphical measures can present a classifier's actual performance for a wide variety of different operating points (combinations of costs and class distributions),
- whereas the best a numerical measure can do is to represent the average performance across a set of operating points.

- Cost curves : perhaps the ideal graphical method in this setting because they directly show performance as a function of the misclassification costs and class distribution.



The x-axis of a cost curve plot is defined by combining the two misclassification costs and the class distribution—represented by $p(+)$, the probability that a given instance is positive—into a single value, $PC(+)$, using the following formula:

$$PC(+) = \frac{p(+)\mathtt{C}(-|+)}{p(+)\mathtt{C}(-|+) + (1 - p(+))\mathtt{C}(+|-)} \qquad (1)$$

Classifier performance, the y-axis of a cost curve plot, is "normalized expected cost" (`NEC`), defined as follows:

$$\mathtt{NEC} = FN * PC(+) + FP * (1 - PC(+)) \qquad (2)$$

where FN is a classifier's false negative rate, and FP is its false positive rate. `NEC` ranges between 0 and 1.

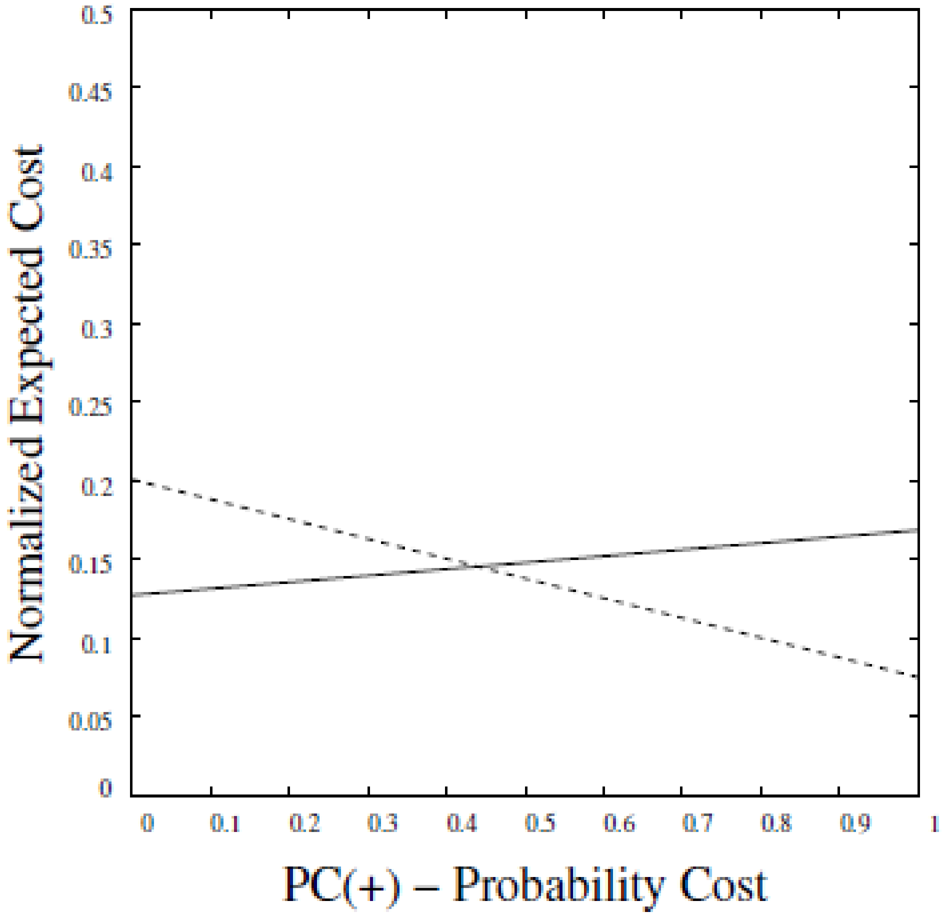


Fig. 1. Japanese credit - Cost curves for 1R (dashed line) and C4.5 (solid line)

cost curves have the following advantages over ROC curves

- Cost curves directly show performance on their y-axis, whereas ROC curves do not explicitly depict performance. This means performance and performance differences can be easily seen in cost curves but not in ROC curves.
- When applied to a set of cost curves the natural way of averaging two-dimensional curves produces a cost curve that represents the average of the performances represented by the given curves. By contrast, there is no agreed upon way to average ROC curves, and none of the proposed averaging methods produces an ROC curve representing average performance.

Thanks for your attention