School of Computing Science,
University of Newcastle upon Tyne

Proper Use of ROC Curves in Intrusion/Anomaly Detection

R. A. Maxion and R. R. Roberts

Technical Report Series

CS-TR-871

November 2004

Proper Use of ROC Curves in Intrusion/Anomaly Detection

Roy A. Maxion and Rachel R. Roberts

Abstract

ROC curves (receiver operating characteristic curves) are commonly used to portray the performance of detectors in signal-detection tasks, such as intrusion detection. This report introduces the origins of signal-detection-theory, and the underpinnings of ROC curves. It provides examples of how to construct these curves, as well as how to measure, interpret and compare them. Information about accommodating cost of error is included. Materials are suggested for further reading.

1 Introduction

This manual will explain what an ROC (receiver operating characteristic) curve is, and how to use one. It is a starting point for practitioners who need to evaluate a single signal detector, and for those who need to compare multiple detectors together. At the end, you should understand what an ROC curve is, how to build one from data, how to use the ROC curve to evaluate a single detector, and how to use it to compare multiple detectors together.

This material is appropriate for binary classification settings. That is, it applies whenever the task of a signal detector is to decide whether a given event is "signal" or "noise". Some signal detectors produce only a single final decision; these are binary classifiers. Other signal detectors may output a larger range of numbers (for example, an integer from 0 to 100 or a real number from 0 to 1) that represents the confidence in a particular judgment. ROC curve methodology is especially useful for continuous-output classifiers, but it can also be used for binary classifiers. This report will dwell on analysis of continuous-output classifiers.

Another key focus of this manual is on using ROC curves in a way that does not make special assumptions about the underlying distributions of signal and noise data that will be analyzed. Statisticians call this brand of analyses "nonparametric". This is in contrast to uses of the ROC curve requiring distributional assumptions, called "parametric". Since no special assumptions about the data are made in advance, the material in this manual is applicable to the widest number of settings. References are provided at the end for further information about the ROC curve, especially the parametric kinds of analyses which depend on prior assumptions about the distribution of the data.

In summary, the goal of this manual is to guide practitioners to use ROC-curve analyses successfully in the general case, in which nothing is known for sure about the probability distributions underlying the data. The only requirement is that the signal detector must distinguish between two events, called "signal" (the event of interest) and "noise" (the uninteresting event). The means of discrimination might involve a continuous-valued threshold (the decision cutoff value), but in the end all judgments terminate in a signal or noise classification.

1.1 What is an ROC curve?

An ROC curve is a two-dimensional depiction of the accuracy of a signal detector, as it arises on a given set of testing data. Two dimensions are required to show the whole story of how the true-positive rate of detection decreases as the false-positive rate of error increases.

An ROC curve plots the true-positive rate of detection, or *TP rate*, against the corresponding false-positive rate of error, or *FP rate*. These two numbers change in relation to each other (determined by theory or experiment) as the detection threshold, or decision cutoff, varies. Whenever the rate of true positives is the highest, the rate of false positives is the lowest, and the reverse is also true. Of course somewhere in the middle they will reach equal proportions. This fundamental tradeoff in the two components of accuracy will vary in a different way from one detector to the next, and from one data set to the next. This makes the shape of each ROC curve look different, meaning primarily more or less "bowed".

Figure 1 shows two views of the same ROC curve; the curve has a slight degree of "bowness". The view on the left shows the (*FP rate, TP rate*) points; each point corresponds to a different decision threshold, in the order of more strict to more lenient, as you travel up the curve toward the right. The view on the right shows the same points connected by straight line segments. A detector can operate at any point along the line-segmented "curve", but interpolation between tested points at run-time occurs in a special way (see Section 2.3).

It is important to realize that connecting the dots by straight lines does not imply that changing the detector threshold to some intermediate value will automatically result in the intermediate performances shown. Instead, interpolating performance between two points (following the straight line segment) requires alternating between the responses of the two endpoints in a proportional way.

Having now seen an ROC curve, it is easier to remember the primary purpose of the ROC curve, which is to illustrate the complete, i.e., two-dimensional accuracy of one or more signal detectors on a given data set. You can use an ROC curve to get a sense of how a single detector is behaving on a particular data set, or to compare the relative accuracies of two or more detectors on the same data.

Beyond visualizing detector accuracy, the ROC curve is also the proper starting point for more detailed analyses about the expected accuracy and cost of the detector in a well-characterized performance setting. For instance, if you know how prevalent signal events are in relation to noise events, you can estimate the final ratio of the two types of errors that will be made, for each threshold level. If

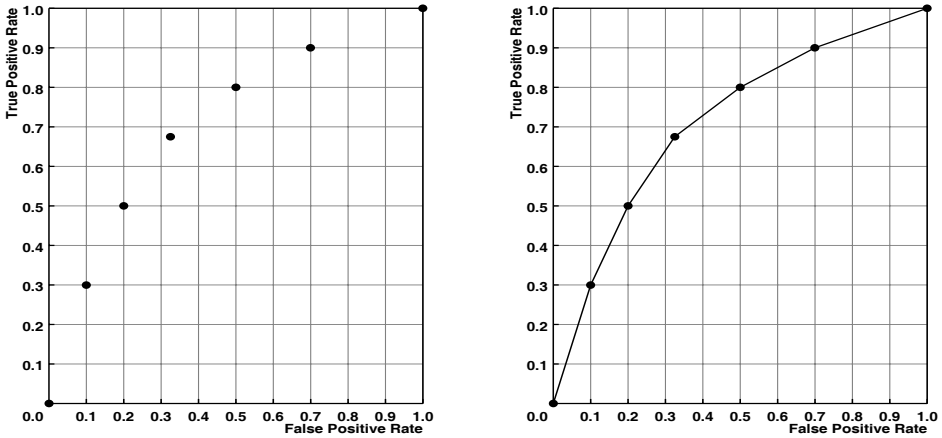

Figure 1: An ROC curve may be shown as unconnected points (a). or as points connected by straight lines (b). These views are appropriate when no underlying statistical models are assumed for the data, and when no attempt is made at curve-smoothing. Each point shown in (a) represents a decision threshold that was tested during evaluation. Each line segment in (b) represents possible future performance achievable by the detector, if proper alternation between endpoint responses occurs at run-time.

in addition you know the relationship between the costs of the two different error types, you can estimate the final cost-ratio of errors made, for each threshold. With this information, you can optimize the decision cutoff so that expected cost due to errors is minimized. Additionally, you can select an optimal signal detector out of many candidates, given class and cost distribution details about the anticipated performance setting (see Section 4).

Even when specific cost and class distribution details are unknown or imprecise, the ROC curve is invaluable in comparing detectors together. Specifically, you can easily tell whether there is a single detector that will *always* outperform others (on data similar to testing data), or whether more than one detector might be optimal, depending on specific error-cost and event-class distribution conditions at the time of detection (see Section 5). The knowledge of whether one ROC curve dominates all others is important because it is robust to imprecise event-class and error-cost distribution estimates. Additionally, when there is no single dominator, the various ROC curves which share domination can be used to construct a hybrid detector which equals or outperforms any single detector, under all event-class and error-cost distribution circumstances related to the testing data (Provost and Fawcett, 2001).

Used by itself or in conjunction with additional analyses, the ROC curve is a choice tool for evaluating the two-dimensional accuracy of one or more signal detectors.

1.2 Quick guide to ROC curve interpretation

The purpose of an ROC curve is to indicate the accuracy of the corresponding signal detector. This accuracy information, revealed in the shape of the curve, is two-dimensional because there are two kinds of events, and hence two kinds of accuracies possible. The first dimension is the success rate of detecting signal events, which is shown along the y-axis (the vertical axis). The second dimension is the error rate of falsely identifying noise events, which is shown along the x-axis (the horizontal axis). Since success is good and error is bad, an ideal ROC curve will have y-values which grow at a faster rate than its x-values, resulting in a curve shape which rises swiftly upward. Later on, as the decision threshold changes to become more and more lenient, the error values for noise (x-values) must also grow large, catching up with the success values for signals (y-values). This makes the curve bend over to the right, until it touches the point (0,1).

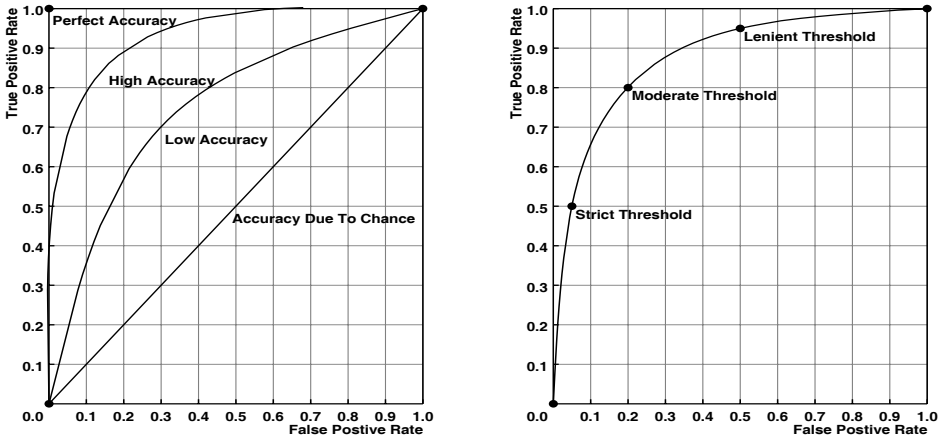

Figure 2: Detector accuracy is indicated by the rise or "bowness" of the ROC curve (a). A perfect ROC curve passes through the point (0,1) while an ROC curve no better than chance lies along the positive diagonal. Shown in (b), each point along the ROC curve corresponds to a different operating mode, or decision threshold. Points closer to the origin (0,0) indicate a more exclusive criterion; points closer to the upper-right (1,0) indicate a more inclusive one. Curves are idealized.

Figure 2 illustrates two important aspects of the ROC curve. In (a), different shapes of ROC curves indicate different levels of detector accuracy. The better (more accurate) curves will bulge further outward to the upper-left, nearing the point of perfection at (0,1). A perfect detector will have a success rate of 1.0 for signals while having an error rate of 0.0 for noises. Unfortunately, this result is difficult to achieve, so most ROC curves will tend to look less angular and more gently bowed, like the middle two curves shown. The worst ROC curve is the positive diagonal line stretching from (0,0) to (1,0); this accuracy could have resulted if one

merely guessed according to chance (assuming equal numbers of signal and noise events). It is not possible to do worse then chance, because then the detector would be perverse (and you could simply invert its output to achieve a better-than-chance result).

In Figure 2(b), three different threshold cutoffs are illustrated. Every ROC curve is based on the measurement of detector performance at various decision thresholds. On the ROC curve, the stricter thresholds appear closer to the point (0,0) and the more lenient thresholds appear closer to the point (1,0). The point (0,0) corresponds to wanting to say "No" all the time (leading to a low false-positive rate but also missing many true positives), and the point (1,0) corresponds to wanting to say "Yes" all the time (capturing nearly all the true positives, but at the expense of a high false-positive rate).

An easy way to start out reading ROC curves is to choose a fixed *FP rate* and then look at the corresponding coordinate for the *TP rate*. This action is analogous to assuming a maximum level of error for false positives, and then seeing how much true-positive success you have achieved. If you are willing to increase the level of false-positive error, then you may be able to gain (a little or a lot) more true-positive success.

In an ROC curve, the accuracy that results overall is a combined result of the individual accuracies achieved at every threshold point tested. Again, evaluation at each threshold results in two numbers, the true-positive rate (*TP rate*) and the false-positive rate (*FP rate*). When the *TP rate* is much higher than the *FP rate*, over more values of *TP rate*, then the curve will have a more extremely bowed shape and represent the performance of a superior detector. It is worth remembering that it is not always just the overall shape that matters, but sometimes just the shape of a curve in a particular region. For instance, if the task necessitates maintaining a false-positive rate below 5% (or < 0.05 error given a noise event), then the only part of the curve of interest is that region which lies to the left of the vertical line passing through $x = 0.05$ on the graph. Thus, looking at the shape (and "bowness") of the curve in a particular region may be more important than the overall curve shape and "bowness".

To summarize, the particular use of the ROC curve will depend upon the detection goal at hand. Two of the most common goals include minimizing expected cost (due to making errors) and maximizing the true-positive rate given a fixed false-positive rate (this is called the Neyman-Pearson criterion). One can imagine other goals, such as minimizing risk (for a given definition of risk), or of achieving equal error rates between false-positive and false-negative errors. However, regardless of the specific detection goal, the starting point is an ROC curve. We will discuss mainly the generation of full (as opposed to partial) ROC curves because this case is the most general.

Finally, before moving on we will summarize the concept of detection outcomes, which will make the ROC curve easier to understand. Table 1 explains the true-positive rate and the false-positive rate in the context of all four possible detection outcomes. Since there are two possible true classes: signal (a.k.a. signal+noise, because noise is always present) and noise (a.k.a. noise alone), and two

possible decision classes: "Yes it's a signal" and "No it's a noise", there are four different results possible. Two of these outcomes are successful, i.e., when the decision matches truth; two are erroneous, i.e., when there's a mismatch between the decision and truth. Note that the outcomes within each column are directly related, but that the two columns are not directly related. Thus when reporting accuracy rates for a signal detector, it is sufficient to name one figure from each column (for instance, the hit rate and the false-alarm rate; or alternatively both error rates, the miss rate and the false-alarm rate). The outcomes of "Miss" (false negative) and "Correct Rejection" (true negative) will not be discussed much in this report because the ROC curve focuses on the other two, the "Hit" (true positive) and the "False Alarm" (false positive). We'll primarily use the terms "true positive" and "false positive" because they are frequently used in formulas and to label the axes of ROC curves.

		Truth	
		Signal + Noise	Noise alone
Observer Decision	Yes	True-positive Rate a.k.a. **Hit Rate:** $\frac{\textbf{\# Hits}}{\text{Total \# Signal Events}}$	False-positive Rate a.k.a. **False-alarm Rate:** $\frac{\textbf{\# False Alarms}}{\text{Total \# Noise Events}}$
	No	False-negative Rate a.k.a. **Miss Rate:** $\frac{\textbf{1 - \# Hits}}{\text{Total \# Signal Events}}$	True-negative Rate a.k.a. **Correct-rej. Rate:** $\frac{\textbf{1 - \# False Alarms}}{\text{Total \# Noise Events}}$

Table 1: Contingency table showing formulas for calculating success and error rates. The hit rate and correct-rejection rate give rates for successes, and the false alarm rate and miss rate give rates for errors.

1.3 When to use ROC curves

ROC curves are beneficial when you need to:

- Visualize the differentiated (two-dimensional) accuracy of a signal detector, over all its decision thresholds
- Recognize the role of adjustable bias (or decision threshold) in the discrimination capacity of the detector

- Disregard additional factors such as the basic *signal:noise* ratio and the differential cost of detection errors (the false positive error and the false negative error)
- Facilitate the comparison of multiple detectors.

Basic needs (assumptions) of ROC curves:

- Source data must be classifiable into two categories, signals and noises[1]
- The "ground truth" label for each data event is available, i.e., knowledge about what is really a signal, and what is really a noise

Desirable conditions, but not absolutely necessary:

- The output of the detector is continuous (not discrete), and in practice it is not sparse, i.e., there is a healthy range of data values represented
- Decision threshold of the detector can be varied continuously

Limitations of an ROC curve:

- The ROC curve is valid and useful only when the testing data used in evaluation are representative of the wider context of interest; since a signal detector's accuracy on one set of testing data may vary dramatically from its accuracy on a different set of testing data, no direct generalizations can be guaranteed[2]
- The ROC curve assumes that the distribution of signals and noises (within each category) is constant, i.e., it assumes that the same kinds of signals are seen in the same kinds of proportions, and that the same kinds of noises are seen in the same kinds of proportions; it is not necessary that the *signal:noise* ratio remain constant, but the within-category assumptions are stringent

Any signal detector which is not yet a perfect discriminator, and which can be tested using a representative set of reference data, is an excellent candidate for ROC-curve analysis. Classifiers having a continuously-valued, adjustable threshold benefit especially from the two-dimensional visualization of the ROC curve.

2 Constructing an ROC curve from data

2.1 Collecting the right data

Before building an ROC curve, you must have a signal detector, and competent evaluation data upon which the detector will be tested. The evaluation data may come in two parts, training and testing data, if the signal detector requires training at the outset. The evaluation data should be as representative of the deployment

[1] It is possible for ROC curves to handle multiple classes, but this topic won't be covered.

[2] A greater amount of generalization may be possible if the testing data is assembled according to known statistical properties, and the response of the signal detector can be accurately predicted for all types of events tested.

setting as possible (in terms of representing the kinds and numbers of signals and noises within each category) in order to make valid any generalizations about detector performance on the data. It should be sufficiently numerous and varied so that the detector output represents all ranges of possible scores (assuming a continuous-output detector).

Preliminary points about data collection:

- You must define what an "event" is, i.e., an observation interval. This definition should be precise, and should cover all necessary circumstances. Each event or observation interval will be scored independently by the signal detector.
- Each event must be labelled truthfully as "signal" or as "noise". A "gold standard" should be used to determine the ground-truth label.
- It is important to have many different, representative examples of both signals and noises, to cover the conditions which emerge in practice. Within each category of signals and noises, there should be representative kinds and proportions of events. [3]
- When the evaluation data elicits a nearly continuous range of values output by the signal detector (so that it is not sparse, but rather the range of data values is filled out), the shape of the ROC curve will be defined in a more precise way. [4]

How to collect data scores, given that a detector is already trained and ready to go:

1. Decide upon what kind of threshold to use (if an adjustable threshold is available). These may correspond to natural modes in the detector, to rating categories used by a human observer, or at the finest grain for continuous data, to levels just beyond every unique data value represented in the testing data. As an alternative, you can choose thresholds that correspond to fixed levels of false-positive rates. The number of thresholds you use will be the number of ROC-curve points on your graph.

2. Run the signal detector on the evaluation data, recording its output for every event.

   (a) For a binary-output classifier, if only one mode is possible, run the detector on the test data and record the labels produced by the detector.

   (b) If a classifier produces only a binary response, but nevertheless has discrete internal modes which may be varied, run the detector on the test data for each possible operating mode (or a subset thereof to test). You will receive a different set of labels for each internal operating mode tested.

[3] Representativeness need only hold *within* each category of signals and noises, if you're building an ROC curve. However, estimating the true-to-life *signal:noise* ratio will prove useful later on.

[4] If you have categorical data with a fixed number of categories, even just two categories, you can still build an ROC curve.

(c) If a classifier outputs discrete labels according to a rating scale, such as a human observer who chooses a number between 1 and 6 inclusive, then run the detector on the test data (or, ask to human observer to rate the test data), and record the labels produced for the single run.

(d) If the detector produces continuous output (such as a confidence level from 1 to 100, or a probability from 0 to 1), run the signal detector once to collect confidence-level or probability labels for each event on that run.

3. Compare the signal detector's result to the ground-truth label, over all events (and over all runs, in the case of the binary signal detector in 2.b. above for which many operating thresholds were induced).

If an automatic classifier (one which learns from prior examples) requires training, and especially if there is a single large pool of data which may become either training or testing data, you must decide how to split the training and the testing data. Ten-fold cross-validation[5] is a popular option in many classification domains. Here's the basic idea behind it:

- In 10-fold cross-validation, you divide all of the data into 10 parts. Nine parts (concatenated together) will be used for training, and the tenth will be used for testing. You will perform 10 separate runs, in which a different part of the data is designated as the testing data for each run. Note that in each run, the testing-data part is *not* included with the training-data part.
- In the end, you will receive 10 different possible scores for each threshold level tested; each of the 10 comes from a run using a different test set. These can be averaged together at each threshold point to create an "averaged ROC curve" (Provost et al., 1998).
- The advantage of cross-validation is that it will improve the generalizability of the accuracy measurements.

Once you have the results of the signal detector for each event in the evaluation data (over multiple runs, if applicable), then you can draw the ROC curve.

2.2 Drawing the right curve

Depending on the nature of the detector tested and the form of scores produced, there are different procedures to follow for drawing the ROC curve (we will use the word "curve" without requiring that the curve is smooth; however, it is at least piecewise continuous). However, each of these methods requires calculating measures called the true-positive rate (*TP rate*) and the false-positive rate (*FP rate*), which will be described forthwith.

To calculate the *TP rate*: (at a fixed threshold)

[5] A generalization of this concept is N-fold cross validation; sometimes called leave-one-out.

1. Divide the evaluation run scores into two categories: events in which the signal actually occurred, and events in which no signal occurred. For now, look *only* at the events in which the signal actually occurred.
2. Count up the number of times the detector found a signal, among the events which were truly signal events (these are called true positives, or "hits").
3. Divide this count by the total number of events in which the signal truly occurred.
4. The result is the rate of true positives, a.k.a. the *hit rate*. It is a conditional probability, $Pr(\textit{Detector said "Yes"}|\textit{Signal occurred})$. The formula can be represented as:

$$\begin{aligned} \textit{TP rate} &= \frac{\textit{\# of times the detector labelled a signal event as a signal}}{\textit{\# of signal events in evaluation data}} \\ &= Pr(\textit{Detector said "Yes"} \mid \textit{Signal occurred}) \\ &= \frac{Pr(\textit{Detector said "Yes"} \cap \textit{Truth is "Yes"})}{Pr(\textit{Truth is "Yes"})} \end{aligned}$$

   The final line expresses the formula in terms of the conditional probability.

To calculate the *FP rate*: (at a fixed threshold)

1. Continuing with the category division above (all events are divided into two categories, events which contain a signal and events which do not): this time, look *only* at the events in which no signal occurred (the noise-only events).
2. Count up the number of times the detector found a signal, among the events which were truly noise-alone events (these are called false positives, or "false alarms").
3. Divide this count by the total number of events in which only noises occurred.
4. The result is the rate of false positives, a.k.a. the *false-alarm rate*. It is a conditional probability, $Pr(\textit{Detector said "Yes"}|\textit{Noise occurred})$. The formula can be represented as:

$$\begin{aligned} \textit{TP rate} &= \frac{\textit{\# of times the detector labelled a noise event as a signal}}{\textit{\# of noise events in evaluation data}} \\ &= Pr(\textit{Detector said "Yes"} \mid \textit{Noise occurred}) \\ &= \frac{Pr(\textit{Detector said "Yes"} \cap \textit{Truth is "No"})}{Pr(\textit{Truth is "No"})} \end{aligned}$$

   Notice the similarity between the formulas for *TP rate* and *FP rate*: in both cases, the numerator is the number of signals that were found, given the category of interest. However, this number is different in both cases because the category of interest is different between the two formulas (for the *TP rate*, it is signal events; for the *FP rate* it is noise events).

Now, given the means to calculate the *TP rate* and *FP rate* coordinates, follow the appropriate procedure to draw the data points which make up the interior of the ROC curve. On each ROC graph, we will assume that the range on each axis is 0.0 to 1.0; the x-axis is the false-positive rate and the y-axis is the true-positive rate.

a. For a binary-output classifier, if only one mode is possible, plot the single point obtained, i.e., (*FP rate*, *TP rate*) on the ROC graph.

b. If a classifier produces only a binary response, but the classifier has N discrete internal modes which may be selected among, then plot the N different (*FP rate*, *TP rate*) points obtained on the ROC graph.

c. If a classifier outputs N discrete labels (rankable from lowest to highest; highest value looks most like a signal), the labels may look like L_1,..., L_n, which can be mapped to the numeric categories $1, \ldots, N$. Here, you must vary the decision threshold (or criterion) c so that it equals each of the rating levels in turn. At each level of the threshold, a new (*FP rate*, *TP rate*) point will be created.

   First set $c = N$, and make the necessary event counts; the times when the detector said "Yes" will be whenever the event seen was given a label of N. Plot the resulting (*FP rate*, *FP rate*) point on the ROC graph. Lower the threshold one notch at a time, making the appropriate counts. For example, when $c = 5$, the events tallied as times when the detector said "Yes" will be those which were labelled by the detector as 5,..., N. After each set of calculations, plot the corresponding (*FP rate*, *FP rate*) point on the graph.

d. If the detector produces continuous output (such as a confidence level from 1 to 100, or a probability from 0 to 1), a procedure similar to 2.c. above will be followed, except that the number of threshold points will be larger. Setting a new threshold point corresponding to each possible data value gives the finest-grain ROC curve available from the raw (discrete) evaluation data.

   An algorithm to build the ROC curve in a single pass through the data appears in Provost and Fawcett (2001). It assumes that the data values are a set of continuously-varying numeric scores from the detector, each accompanied by a truth label (knowledge of whether the event was actually a signal or actually a noise). Here is the algorithm, paraphrased:

   - Order the data values (numeric scores from the detector) from largest, or most confidently signal, to the smallest, or most confidently noise.
   - Set the threshold c to just below the first data value, so that no other data values are $\leq c$. Increment the number of "Detector said 'Yes'" events, and also the number of total signal or total noise events (depending on whether the data value was a signal or a noise). Calculate the resulting *FP rate* and *TP rate* so far, and add the point to the curve.
   - Given each successive unique data value, repeat the above procedure to obtain a new (*FP rate*, *TP rate*) point, and graph it.

- Handling "ties": If there are many data items with the same numeric value, they will be processed in a batch; that is, the threshold will be lowered only once for the group, but counts of "total signal events" and "total noise events" should still be incremented appropriately for each truth label present in that batch. The "Detector said 'Yes'" labels will also be incremented according to the size of the batch.

In addition to the points interior to the ROC curve (graphed per the relevant instructions above), the points (0,0) and (1,0) should be added to each ROC graph. The point (0,0) represents the detector *always* saying "No", and the latter represents the detector *always* saying "Yes". While these may not be practical operating modes in a useful detector, they are required to make the ROC graph complete.

Next, connect all points on the ROC graph, including the additional points (0,0) and (1,0), with straight line segments. If the final shape which emerges is nondecreasing and convex (bulging upward) the entire way, and entirely above the positive diagonal $y = x$, then you can stop; you have an ROC curve which represents the highest accuracy attainable by your detector on this testing data.

If the resulting ROC curve seems to "sag" a bit, meaning that it is not entirely convex, that it decreases for a stretch along the way, or that it falls below the positive diagonal $y = x$, then you will need to make some adjustments to your curve. Given one bad point (or few in a row), simply "enlarge" the curve a bit in that spot by connecting with a straight line the two nearest points which do not violate the criteria. Remember you can also use the points (0,0) and (1,0) as endpoints. A simple example is shown in Figure 3.

The justification (and imperative) for "enlarging" the ROC curve in this way is that in the field, you could set your detector to operate at only the best threshold points available. To achieve some intermediate threshold between them, you could decide to alternate detector response between the two nearest acceptable thresholds, achieving an accuracy equal to the straight-line interpolation between the two points. As long as this line lies "outside", or to the "northwest" of a former threshold point deemed unacceptable, then it is best to operate the detector on the principle of such an interpolation, as opposed to actually operating the detector at the threshold level corresponding to the "bad" point.

Note that it is still informative to know whether or not if the ROC curve produced naturally by a detector yields a nondecreasing, convex curve above the diagonal $y = x$. Knowledge to the contrary might call for improvements to the detector. However, for the purposes of measuring and comparing achievable accuracy, it is best to make the corrections above, because this is what should be done in practice when such a detector must be used.

If desired, for visualization purposes you could plot only those straight-line interpolated points which lie at fixed intervals of the *FP rate*; however for all calculations it is best to use the entire "raw" ROC curve achieved.

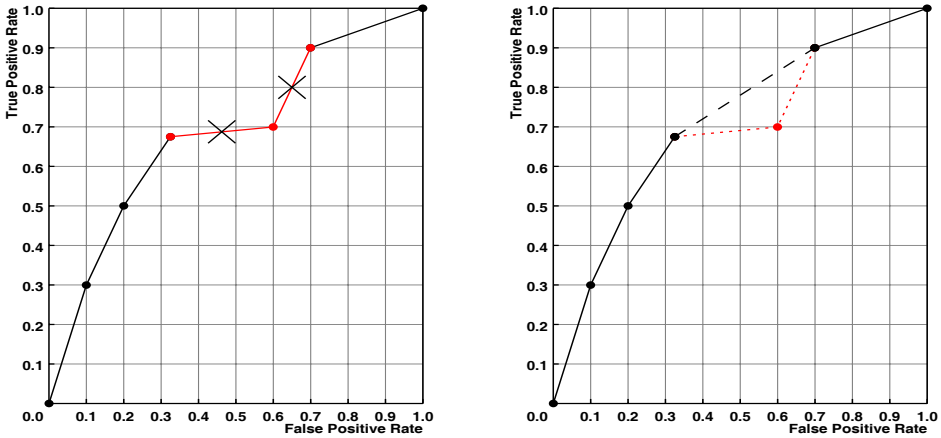

Figure 3: Correcting for an ROC data point in (a) which truly represents the data, but which corresponds to a detector operating at less than its potential accuracy. The corrections shown in (b) represent the detector's fullest potential accuracy, as achieved by alternating final response between the two nearest advantageous thresholds.

2.3 Straight-line interpolation versus smoothing

There are two ways to connect data points obtained from an empirical ROC testing procedure. One is using a straight-line interpolation (connecting successive data points by straight line segments), and the other is to attempt to smooth out the angular nature of the curve using a smoothing procedure. If a great deal of smoothing is used, then the curve may shift slightly from the original data points in places, and in the limit, a parametric model may be substituted for the original empirical curve. (Parametric models make assumptions about the underlying data distributions, yielding economy and generalizability, usually at the price of remaining strictly true to the data.)

In general, if your data values are continuous and represented throughout their entire range, then you should not need to smooth at all. The idea behind smoothing is to compensate for a potential underestimation of the true accuracy of the signal detector. This may occur because straight-line interpolations are conservative, though 100% true to the testing data observed (because this accuracy can be recreated faithfully on the same testing data set). Since a more "bowed" ROC curve is better, the jagged nature of the straight-line interpolations shrinks the total "bowness" from what it otherwise could be. However, when the data are continuous and not sparse, the discrepancy between the straight-line interpolation and the smoothed interpolation will be very slight.

In practice, to interpolate between two already-tested thresholds according to the straight-line (conservative) method, simply alternate the decision of the detector

between the two known endpoints in a randomized way, but weighted according to which point you'd most closely like to resemble. The long-run accuracy of the detector will then resemble the point found at the straight-line interpolation between the two thresholds operated at. Because such interpolations are possible in practice, using the representation of the straight-line "connected" ROC curve is valid. Remember that the straight-line interpolation must be achieved in this response-alternating way; it is *not* achieved by attempting to set a continuously-varying threshold on a signal detector to some intermediate value between the two thresholds already tested. (It is possible that the signal detector may behave either better, or worse at this untested point; there is no guarantee. Thus, the response-alternation method is the only way to achieve the expected accuracy.)

Returning to smoothing, if the testing data are somewhat limited, and it is strongly suspected (from previous experience in the domain) that the "true" ROC curve is more continuous in shape, then smoothing could be used to compensate for a possible underestimation by straight-line interpolation. You must now also decide how much to smooth the ROC curve; it's important not to over-smooth, else the accuracy of the curve will be *overestimated*.

Since smoothing is not strictly necessary, and because the straight-line interpolation gives the most conservative estimate of accuracy, treatment of smoothing is not covered in this manual. However, if your curves look extremely jagged and you wish to attempt smoothing, it is advised to consult an ROC expert or a statistician to help you.

In this section we have seen how to collect the right data and draw a reasonable ROC curve. Armed with an ROC curve, but keeping in mind the limitations of an empirical ROC curve (detailed in Section 1.3), we move on to the next step, which is calculating and using measures for ROC curves.

3 Measures provided by the ROC curve

It is important to remember that all measures made on the ROC curve are subject to the same limitations as the ROC curve itself; if the ROC curve is fallible, then the measures are likewise fallible. Some of the measures imply added ambiguity or uncertainty, as described below; they should always be used in conjunction with the larger picture of the ROC curve.

3.1 Accuracy measures

Simply inspecting an ROC curve visually provides clues to a detector's general accuracy. Wherever the curve is steeper, the relative confidence in a *signal* judgment is greater; wherever the curve is flatter, the relative confidence in a *noise* judgment is greater. You can also see how closely the curve approaches perfection at the "northwest" corner point (0,1), or on the other hand, how closely it lies to the line of accuracy due to random chance, the positive diagonal $y = x$. Looking at a fixed rate of false positives, e.g., 0.20 (or 20%) in Figure 3(b), you can see the corresponding rate of true positives, which turns out to be 0.60 (or 60%) in the same

figure.

However, often a need arises for a single-number summary of the detector's performance. While these numeric measures do not give as complete a picture as the whole ROC curve, they are often easier to digest and think about, so they help complement the ROC-curve picture with specific detail.

The two most common views of accuracy in nonparametric settings are the *TP rate at a fixed TP rate*, also known as the Neyman-Pearson criterion, and the nonparametric measure growing in renown, the *area under the curve* or *AUC*. The first is useful when you have a particular *FP rate* cutoff in mind, and the second is useful as a very general single-number summary of accuracy.

There are other measures for use in parametric settings, the most common of which is d′ (called "d-prime"). Roughly, d′ is a single measure of the departure of the bowed edge of the curve from the positive diagonal $y = x$. It works especially well when the underlying signal and noise distributions are both normal and have equal variance, because then the ROC curve is symmetric in shape, and a single number easily describes the extent of the "bowness" of the curve. It is less clear what such a single number might mean, whenever the curve is not symmetric. The *area under the curve* measure is then preferred, because it does not require the assumption of symmetry.

The Neyman-Pearson criterion and AUC will be described in further detail. They are both nonparametric, in the sense that they don't require special assumptions about the underlying statistical distributions of signal and noise data.

3.1.1 True-positive rate at a fixed false-positive rate

Maximizing the level of "hit" success attainable at a fixed level of "false alarms" is the essence of the Neyman-Pearson criterion of signal-detection optimization. Informally, this means looking at the TP rate, given a fixed FP rate. It requires that you first choose a false-positive rate beyond which detection performance is unacceptable.

For example, in writing up a report, a researcher may want to use a particular statistical test to determine the significance of a claim, in which there is no greater than a 5% chance (akin to 0.05 as an *FP rate*) of claiming an erroneous result. On the other hand, during exploratory work, it may be advantageous to look at the entire subset of features mostly likely to be important, accepting up to a 25% chance that a particular feature may be irrelevant. (In the hopes of catching almost every feature that is relevant, some unimportant ones may be entertained at first.) Thus, it is possible to fix an upper limit for the *FP rate*, and the value for the limit depends on the circumstances of its use.

After the *FP rate* is set, it remains to see what the best *TP rate* achievable for that level is. This number can then be reported, along with the fixed *FP* level.

Of course it is possible to first look at the entire picture of the ROC curve, and then decide upon a fixed *FP rate*. While this may not be acceptable in statistical hypothesis testing, there is nothing wrong in doing this in other stages of research, and wherever appropriate. The changing slope of the ROC curve provides an im-

portant clue: if the curve is steep in the region of interest, it may be worth a small loss in increasing the maximum acceptable *FP rate*, in order to win larger gains in an increased *TP rate*. Similarly, if the curve is already very flat in the region of interest, then a larger *FP rate* will not gain much, and perhaps decreasing it a bit will improve the *FP rate* itself while not detracting from the achievable *TP rate* in a significant way.

To compare together two or more ROC curves using this measure of accuracy, simply find the curve with the greater *TP rate* for a given fixed *FP rate*.

3.1.2 Dominance of a curve over a region of the graph

The topic of dominance will be fully explained in Section 5.1. However, it is worth mentioning here that a useful feature of ROC curves is whether or not any single curve lies more "outside" another (more "north" and more "west" of the other). This observation is slightly more general than the Neyman-Pearson criterion, because it considers curve "height" over an entire region rather than at a single fixed *FP rate*.

The fact that a single curve dominates all others shows that it is superior over that part of the range of the ROC graph; however, this superiority is more difficult to quantify in an unambiguous way (unlike the TP rate at a fixed FP rate measure). An initial effort toward this end is to compare together the partial areas under the curve, for different ROC curves (see Section 3.1.4). However, as will be mentioned, all area measures are slightly ambiguous and should therefore be used judiciously. However, given the additional knowledge of domination, one can be much more confident in the area measures, which are discussed next.

3.1.3 Total area under the curve

The most general measure of total detection accuracy is the area under the curve, or AUC. It summarizes the total accuracy of the detector in a way that accounts for both the gains in *TP rate* and the losses in *FP rate*. The number for AUC always ranges from 0.5 to 1.0, because the very worst ROC curve (due to chance alone) lies along the positive diagonal $y = x$ and has the corresponding area of 0.5. The very best ROC curve, passing through the "northwest" corner point (0,1), has an area of 1, or the unit square.

To find the AUC, simply calculate the trapezoidal area under each vertical slice of an empirical (unsmoothed) ROC curve having a straight-line segment as its top; then sum all the individual areas. For instance, if a binary detector produces only a single data point, then there will be two trapezoidal regions total for which to calculate area (one to the left of the point, and one to the right of the point).

Since every ROC curve is nondecreasing, following the trapezoidal will systematically underestimate the area under an ROC curve, if the true (completely continuous) ROC curve is actually smooth. However, if there are many threshold points in the ROC curve, and they are spaced rather closely, the discrepancy will not be large. If it is feared that the trapezoidal area is in fact too small, smoothing

can be applied to the ROC curve, before the area underneath it is calculated (Section 2.3 briefly discusses smoothing). However, the conservative and simple nature of the trapezoidal AUC may have greater advantages.

One drawback to the AUC measure is that given a single AUC number, you cannot reconstruct the shape of the curve which produced it. It is ambiguous precisely because it abstracts away the two-dimensional shape of the curve. If there is a specific use in mind for the signal detector, then this may not be the best measure to use, because it gives only the most general picture of total accuracy (similar to an average accuracy over all possible modes of the detector). One ROC curve having a greater AUC may actually perform less well under some conditions than a different ROC curve having a smaller AUC number.

Though, when comparing together two or more ROC curves using the AUC measure, the curve with the greater AUC is truly more accurate (when averaged over all detection thresholds). However, keep the point in mind that sometimes it is more important for a curve to be "above" another (to dominate) over certain entire regions of the graph, than for one curve to have a greater AUC than another.

In spite of its drawback of sometimes being ambiguous or even misleading (as mentioned previously), the AUC has a satisfying intuitive meaning.[6] The number for AUC is between 0.5 and 1.0; this lends naturally to an interpretation as a probability (between the accuracy due to chance at 0.5, and perfection at 1.0). In fact, the AUC is equivalent to the probability of correctly identifying the signal, given a forced-choice test where one must choose which event is more likely to contain a signal, out of one randomly chosen signal event, and one randomly chosen noise event. If the AUC is, say, 0.88, then there is an 88% chance that given two randomly selected alternatives (one signal and one noise), the correct one will be identified as signal. This measure reflects the inherent difficulty of the detection task while using a particular signal detector. Note, however, that the intuitive interpretation assumes that signal and noise events generally occur in equal amounts; whenever this is not true, the final number for accuracy will differ, being dependent upon the *signal:noise* ratio as well.

Because of its extremely general nature, the *area under the curve* measure is ideally suited for high-level detector comparisons, such as in evaluating core signal detector technology. It is also useful for summarizing the entire picture of a detector's performance. If you have more specific needs in a particular detection setting, it may be preferable to use the *partial area under the curve* (see next section) or even an iso-performance line in conjunction with an ROC convex hull (see Section 5) to provide more meaningful comparisons.

3.1.4 Partial area under the curve

Partial area under the curve (p-AUC) is just like the total area under the curve, except that only a subset of the ROC graph is considered. For instance, if it is known that a false-positive rate of 0.50 or greater is completely unacceptable, then only the left half of the ROC graph and curve need be considered. In this case,

[6] It also has ties to Wilcoxon-Mann-Whitney statistics, not mentioned here.

p-AUC will no longer range between 0.5 and 1.0, like the AUC measure did, but it will have a new minimum and maximum level (from 0.125 to 0.5, respectively) which always depends on how much of the graph is considered.

In order to narrow down the region of the curve of interest, it is necessary to have in mind a fixed maximum false-alarm rate, a fixed minimum true-positive rate, or the ideal slope or estimated slope range (see Section 3.2 and Section 4 for more about slope). In terms of focus it is somewhere between total area and the Neyman-Pearson criterion, because it considers accuracy over a range of the ROC graph but not over the entire graph. However, like the total AUC measure, it suffers from ambiguity because if the curves cross one another within the region of interest, it is not clear that one of the curves having a larger area will unambiguously be the better detector to use under deployment conditions. However, if a single curve dominates the region of interest, then the p-AUC measure becomes less problematic. (See section 5.1 for more on dominating ROC curves.)

3.2 Bias measure: Slope of the ROC curve

As we will learn later in Section 4, an important use of the slope of the ROC curve lies in selecting a detection threshold for operation. Before appreciating that role, it is important to understand how the slope of the ROC curve represents the bias of the detector at a given operating threshold (on the assumption that signals and noises occur with equal frequency).

On every ROC curve, the slope varies (either discretely or continuously) from a high value $m > 1$ to a low value $m = 1$. The worst ROC curve will have a constant slope $m = 1$, corresponding to the positive diagonal line $y = x$. The best ROC curve will have infinite slope ($m = \infty$) from (0,0) to (0,1), and then 0 slope ($m = 0$) from (0,1) to (1,1). For every other ROC curve, the slope will vary throughout the curve between $m = \infty$ (higher slopes occur near the "southwest" corner) and $m = 0$ (lower slopes occur near the "northeast" corner).

Because the slope m takes on a value between 0 and ∞, it may seem suspiciously similar to a likelihood ratio. In fact, it is one. The slope of ROC curve at a point is equivalent to the following likelihood ratio:

$$L(q) = \frac{Pr(q \mid q \in sn)}{Pr(q \mid q \in n)}$$

In the formula above, q represents one possible value of the signal detector's output (such as a numeric confidence level or discrete rating category). sn represents the category "truly signal", and n represents the category "truly noise". $q \in sn$ means that the observed value q is assumed to be a signal, and $q \in n$ means that q is assumed to be a noise. (The symbol $\in$ means "is an element of"; for example $A \in B$ means that A is a member of the set B.) The probabilities shown above are conditional probabilities. $Pr(q \mid q \in sn)$ answers: "Among true signals only, what is the probability that the level of evidence q occurs?" Similarly, $Pr(q \mid q \in n)$ answers: "Among true noises only, what is the probability that the level of evidence q occurs?" The likelihood ratio deals only with the observed level

of evidence; it ignores completely the fact whether signals or noises are more likely to occur, in general. (Or equivalently, it assumes that signals and noises occur with equal frequency.)

In effect, the likelihood ratio given above is a comparison (a ratio) between the probability of q having occurred, assuming it was a signal, and the probability of q having occurred, assuming it was a noise. If q is a very common type of signal, then $Pr(q \mid q \in sn)$ will be large (closer to 1 than to 0). If q is an uncommon type of signal, then the probability will be small (closer to 0 than to 1). Similar statements can be made for $Pr(q \mid q \in n)$ and whether q is a common, or an uncommon type of noise. Note that it doesn't matter necessarily whether the two conditional probabilities are large or small; the only thing that matters is what is their relationship to each other. When the two conditional probabilities are equal in size (no matter how large or small both are), then $L(q) = 1$. If the numerator (on top) is larger than the denominator (on bottom), then $L(q) > 1$; if the situation is reversed and the denominator (on bottom) is larger than the numerator (on top), then $L(q) < 1$. Just like the slope of the ROC curve, $L(q)$ ranges from 0 (when the numerator is zero and the denominator is non-zero) to ∞ (when the denominator is zero, the numerator "blows up" to ∞).

If the underlying distributions of signal and noise data are not known, then the likelihood ratio $L(q)$ cannot be calculated outright, without knowledge of the ROC curve. Once you have the ROC curve, you can calculate $L(c)$ at a given threshold c just by looking at the ROC curve. Given a data point on an empirical ROC curve, use the slope of the line segment that terminates just below (or to the left of) the point. For example in Figure 3(b), at the threshold point designated by (0.1, 0.3), $L(q) = 3$ because the slope of the line segment just below it between (0,0) and (0.1, 0.3) is exactly 3. Notice that this slope is > 1, because it's near the origin (0,0). Using similar reasoning, at the operating point threshold c represented by (1,1), $L(q) = 1/3$, which is < 1. Note that if the ROC curve happens to be smooth, then the slope at a point is equivalent to the slope of a line tangent to the curve at that point.

The practical interpretation of ROC-curve slope, or equivalently the likelihood ratio at a given threshold point, is the bias of the detector at that threshold. For instance, if $L(q) = 1$, then based on the value of the evidence q alone, the odds are equal that the evidence is signal or noise. While operating at this given threshold c, where $L(c) = 1$, the detector has no bias because it will always make a decision in keeping with the observed evidence. In this case, it pronounces "signal" whenever the observed evidence q is such that $L(q) > 1$ and it judges "noise" whenever the value of q is such that $L(q) < 1$. If the threshold c is altered from this point, then bias is introduced. When c is increased to a more stringent level (such that $L(c) > 1$), bias in favor of making noise judgments is introduced, because there must now be an overwhelming level of evidence q such that $Pr(q \mid q \in sn)$ exceeds $Pr(q \mid q \in n)$ by a factor greater than or equal to $L(c)$ in order to elicit a *signal* label. Similarly, if the threshold is decreased, the odds must be greatly against the determination of *signal* in order for a *noise* judgment to result. Thus the slope of the line corresponding to an operating point (having decision threshold c) indicates the

bias programmed into the detector, with respect to the level of observed evidence alone (that is, not considering the background *signal:noise* ratio).

A detector can be given as much or as little bias as desired; in the extreme case, a detector will always say "No, q is a noise" when $L(c) = \infty$, and a detector will always say "Yes, q is a signal" when $L(c) = 0$. Again, the value of the ROC-curve slope, or likelihood ratio $L(c)$ represents the ratio of probabilities that must be met or exceeded, in order to elicit the *signal* judgment.

At first glance, it may seem reasonable to want to avoid all bias (a good policy in many experimental situations). However, as will be seen shortly, bias is an essential characteristic of successful signal detectors. The important thing is to choose the bias carefully, so that the risk of errors is minimized (or a different appropriate detection goal is reached). If you know in advance the desired bias ratio, you can select an ideal operating threshold based on that knowledge. It is best to calculate the desired bias ratio based on as much information available as possible; Section 4 will discuss two invaluable sources of additional information, and show how to use them to calculate an ideal bias for the detector. (The likelihood ratio only tells part of the story, and if you have more information available then you should use that when calculating the ideal bias.)

4 Using extra information to select thresholds or detectors

As mentioned before, the ROC curve abstracts away important details of the problem context, which enhances its generality and broadens its applicability in empirical settings and theoretical scenarios.

However at some point (especially at deployment), if estimates are available, you may want to reintroduce at least the following two factors back into the realm of analysis: the base rate of signal prevalence (or *signal:noise* ratio) and the relative cost of making a detection error (or error-cost ratio). With knowledge of either or both, you can improve the choice of an operating threshold for a given signal detector, and improve the choice of which signal detector to use for the task, given a choice of many.

4.1 Class-distribution information predicts future error rates

Because it calculates separately the conditional probabilities for signal and noise conditions, the ROC curve abstracts away the base rate of signal prevalence, a.k.a. *signal:noise* ratio. If you take ROC-curve results literally, they assume the relative frequency of each kind of event (signal or noise) is the same. Since this is often not the case in reality, it is usually dangerous to use *only* the ROC curve in order to decide such questions as where to set the threshold, or which signal detector to use.

If it is possible to estimate the ratio of signal occurrence to noise occurrence (over a useful span of time), one can improve the prediction of total error rate implied by the ROC curve. All that is needed is to multiply each conditional probability (given a known threshold) by the base rate of event occurrence (for that category of events). This is equivalent to calculating the posteriori probability of an event,

given the prior probability (base rate) and the updating evidence (true-positive or false-positive rate, which are conditional probabilities).

Remember the formula to express the true-positive rate:

$$\begin{aligned} \textit{TP rate} &= Pr(\textit{Detector said ``Yes''}) \mid \textit{Signal occurred}) \\ &= Pr(\textit{``Yes''} \mid \textit{Signal}) \end{aligned}$$

Similarly, for the false-positive rate:

$$\begin{aligned} \textit{FP rate} &= Pr(\textit{Detector said ``Yes''}) \mid \textit{Noise occurred}) \\ &= Pr(\textit{``Yes''} \mid \textit{Noise}) \end{aligned}$$

Using the two conditional probabilities *TP rate* and *FP rate* as the "updating evidence", we can now calculate the posteriori probability of a true-positive event and of a false-positive event:

$$\begin{aligned} \textit{Updating Evidence} * \textit{Prior Probability} &= \textit{Posterior Probability} \\ Pr(\textit{``Yes''} \mid \textit{Signal}) * Pr(\textit{Signal}) &= Pr(\textit{True Positive}) \\ Pr(\textit{``Yes''} \mid \textit{Noise}) * Pr(\textit{Noise}) &= Pr(\textit{False Positive}) \end{aligned}$$

These probabilities express the relative frequency of each kind of error. Note that the new values $Pr(\textit{True Positive})$ and $Pr(\textit{False Positive})$ should not be confused with the *TP rate* and the *FP Rate*. The events "True Positive" and "False Positive" are two of the four possible detection outcomes; $Pr(\textit{True Positive})$ and $Pr(\textit{False Positive})$ together make up a portion of the total probability of 1 which is shared among the four detection outcomes.

As a reminder, Table 2 shows the four possible detection outcomes. When the frequencies of these outcomes are counted (in an empirical setting) and the relative frequency determined, the sum of all four probabilities will add to 1. This is different than the case of the *TP rate* and *FP rate*, which are independent of one another because the *signal:noise* ratio is not considered in their calculations.

As a final note on the posterior probabilities for "True Positive" events and "False Positive" events, respectively, it is assumed that the occurrence of each new event is independent of all the events just prior to it. Also, probabilities are just probabilities; they don't give actual numbers (per unit time). However, if you also knew the rate of arrival of events in general (in addition to the *signal:noise* ratio), you could additionally calculate the rate of arrival for each kind of error.

Beyond calculating the posterior probabilities outright, another useful concept is the *signal:noise* ratio. As the name implies, this ratio is simply the relative prevalence of signal events compared to the relative prevalence of noise events. The formula for the *signal:noise* ratio is:

$$\textit{Signal:Noise ratio} = \frac{Pr(Signal)}{Pr(Noise)}$$

		Truth	
		Signal + Noise	Noise alone
Observer Decision	Yes	True Positive **Hit**	False Positive **False Alarm**
	No	False Negative **Miss**	True Negative **Correct Rejection**

Table 2: Four detection outcomes are possible when an observer makes a decision about whether an event is a signal or a noise. Two sets of terms are shown; words in the lighter type often appear as axis labels and (in abbreviated form) in mathematical formulas. The words in darker type originated in military settings and are commonly used in parlance.

If you knew the *signal:noise* ratio, but not information about error cost (or, the cost of errors was equal), then you could calculate the ideal bias of the detector in order to minimize errors. The formula to calculate this slope is:

$$\textit{Ideal ROC slope}^{*} = \frac{Pr(Noise)}{Pr(Signal)}$$

* When the *signal:noise* ratio is known, but the error-cost ratio is either unknown or unity.

In effect, to find the ideal detector bias the *signal:noise* ratio is inverted, so that the bias given to the detector compensates for the skewed prevalence of one kind of event or the other. Given an ideal slope of the ROC curve, one can locate the threshold point which corresponds to the ideal slope. At that threshold point, the total number of errors will be minimized.

4.2 Error-cost information predicts future cost

The other major factor of importance in both setting a detection threshold and in selecting an appropriate detector is error cost. Since there are two kinds of error possible (false positives and false negatives), it follows that the average cost of an error may be differentiated primarily between these two categories. Depending on

which kind of error is more grievous, signal-detector operators will want to choose a signal detector and a decision threshold that are biased toward making the cheaper kind of error (knowing that it is impossible to avoid all error).

Similar to the *signal:noise* ratio, an *error-cost ratio* can be calculated which compares the sizes of the costs of error against each other. Here is the error-cost ratio:

$$\textit{Error-cost ratio} = \frac{Cost(\textit{False Negative})}{Cost(\textit{False Positive})}$$

On top in the numerator is the kind of error related to signals; on bottom in the denominator is the kind of error related to noises. If you wanted to select an appropriate detection threshold, based on error-cost information alone (assuming that the *signal:noise* is 1), this is the formula to find the ideal ROC slope:

$$\textit{Ideal ROC slope}^{**} = \frac{Cost(\textit{False Positive})}{Cost(\textit{False Negative})}$$

** When the error-cost ratio is known, but the *signal:noise* ratio is either unknown or unity.

Again, the crux of calculating the ideal slope is to bias the ROC curve in favor of making the cheaper type of error. This is done by inverting the error-cost ratio and making that the ideal slope. For example, if the error-cost ratio is *1:2*, meaning that false-positive errors are twice as expensive as false-negative (miss) errors, then the ideal slope of the ROC curve (ignoring *signal:noise* ratio for now) will be $\frac{2}{1} = 2$. Given a slope of 2, which is > 1, the detector is biased in favor of limiting the false positive error, since the criteria for awarding a *signal* judgment has now become more strict. Though the relative number of false-negative errors will now increase, the decrease in expensive false-positive errors will ensure that the final cost is minimized.

If you know the *signal:noise* ratio in addition to the error-cost ratio, you can combine the two to choose an ideal slope or detection threshold. The formula now becomes:

$$\textit{Ideal ROC slope}^{***} = \frac{Pr(Noise)}{Pr(Signal)} * \frac{Cost(\textit{False Positive})}{Cost(\textit{False Negative})}$$

*** Considering the *signal:noise* ratio and the error-cost ratio as additional factors.

Notice again that all factors pertaining to noises are on top (in the numerators), and all factors pertaining to signals are on bottom (in the denominators). However, the "ideal slope" has a signal-related figure (change in *TP rate*) in the numerator, and a noise-related figure (change in *FP rate*) in the denominator. The idea is to make the slope compensate (through inversion) for the relative importance of avoiding one kind of error over the other. Two more views of the equation will bring the relationship into perfect light:

$$\frac{\Delta \textit{TP rate}}{\Delta \textit{FP rate}} = \frac{Pr(Noise)}{Pr(Signal)} * \frac{Cost(\textit{False Positive})}{Cost(\textit{False Negative})}$$

And finally:

$$1 = \frac{\Delta \textit{FP rate}}{\Delta \textit{TP rate}} * \frac{Pr(Noise)}{Pr(Signal)} * \frac{Cost(\textit{False Positive})}{Cost(\textit{False Negative})}$$

Notice that the "ideal slope" term $(\Delta \textit{TP rate})/(\Delta \textit{FP rate})$ was inverted to make the last equation. Now, in the last equation, noise-related numbers are all in the numerators and signal-related numbers are all in the denominators. That the total product is equal to unity indicates an exact balance between the two kinds of numbers, noise and signal. (Imagine taking all the numbers out the denominator on the right-hand side of the equation, and moving them to the left side, through multiplication on both sides. Then, the product of the signal-related figures will be equivalent to the product of the noise-related figures.) The balance (or equality between signal-related numbers and noise-related numbers) means that a minimum cost has been found.

So far, we have mentioned only using the "ideal slope" calculation to choose the detection threshold of a single detector. This "ideal slope" can also be used to calculate the slope of an iso-performance line, used for selecting the best classifier out of a group. See Section 5.3 for more information on using ideal slope to select which detector is most appropriate for a given situation.

It is worth reviewing a basic assumption made so far, that the only cost incurred is due to the cost of a detection *error* made. In reality, there are many other costs: building a detector, testing a detector, running a detector, even handling "success" cases (true positives and true negatives). These costs are also likely to vary according to the detector. Thus, the choice of detector should ideally incorporate these many cost factors.

A simple formula incorporating many of these costs is:

$$\begin{aligned} C_{avg}(\textit{Detector D}) = C_0 &+ C_{TP} * P(TP) + C_{TN} * P(TN) \\ &+ C_{FP} * P(FP) + C_{FN} * P(FN) \end{aligned}$$

Above, C stands for cost; C_0 is the overhead cost of (developing, testing, running) a detector; the TP, TN, FP, FN are the four kinds of detection outcomes (true positive, true negative, false positive, false negative); $C_{avg}(\textit{Detector D})$ is the average cost of using a particular detector, D; the four probabilities were obtained using both the *signal:noise* ratio and the *TP rate* and *FP rate* given the threshold which produces the lowest cost for the detector. Using this kind of formula, the average cost for using different type of detectors can be compared together (assuming each detector minimizes its own total cost), and the detector having the minimum cost can be singled out. Given an additional number for the rate of event arrival, the total cost of using the detector (per unit time) can also be calculated.

As the last word on choosing an ideal slope, it is possible to entertain an estimated range for either the *signal:noise* ratio, the error-cost ratio, or both. It is quite

easy to calculate a range of ideal slopes for these cases. Just use the smallest set of figures to calculate one endpoint for the slope range, and use the largest set of figures to calculate the other endpoint. Given a range of possibly ideal slopes, one can still rule out certain detectors which do not perform better than others within this particular slope range. See also Section 5.3 for more discussion on utilizing an ideal-slope range.

Finally, as for any predictions about the future, the material in this section is subject to certain limitations. Most importantly, the above calculations assume that the class distribution and cost distribution ratios are static. In reality, the *signal:noise* ratio and the error-cost ratio may vary over time. During operation, it is wise to continually measure and re-calculate these ratios, so that predicted rates of errors and costs can be updated and the choice of threshold or detector can be flexible as well. These measures also contain all of the assumptions of a regular ROC curve, i.e., that testing data (or past data) is a perfect representation of the nature of future domain events, etc. Thus, they should be regarded as indicators, but not as perfect predictors; the final result is only as good as the quality of each estimate and the validity of each assumption made along the way.

5 Comparing many detectors and using the ROCCH

When multiple signal detectors are tested on the same data set, you can plot their ROC curves on the same graph. This facilitates visual comparison and directly leads to conclusions about dominance. (If if the curves are bunched up and hard to see, you can try to spread them out by showing only part of the graph, emphasizing the region of interest, or by changing the scale of one or both axes.)

5.1 Dominating ROC curve

If on a graph containing multiple ROC curves, a single curve lies outside of (above and to the left of) every other curve, then it "dominates" the others completely. A dominating ROC curve will outperform every other detector present, on data sets similar in composition to the testing data set (regardless of the *signal:noise* ratio or the error-cost ratio). The concept of dominance is illustrated in Figure 4(a), where curve A dominates curves B and C completely because it always lies "north" and "west" of them.

It is possible for an ROC curve to dominate partially, meaning over a certain region of the ROC space but not over the entire graph. In this case, that ROC curve will be optimal (better than the others) under a certain range of conditions related to *signal:noise* ratio and error-cost ratio. Partial dominance is illustrated in Figure 4(b), where each of the curves A, B, and C dominate over a select region of the ROC graph.

When two or more ROC curves coincide together outside the others, sharing domination over the same region, then they are equivalent in performance in the region over which they dominate. If different ROC curves dominate over different portions of the ROC graph, then you can run the corresponding detectors in parallel

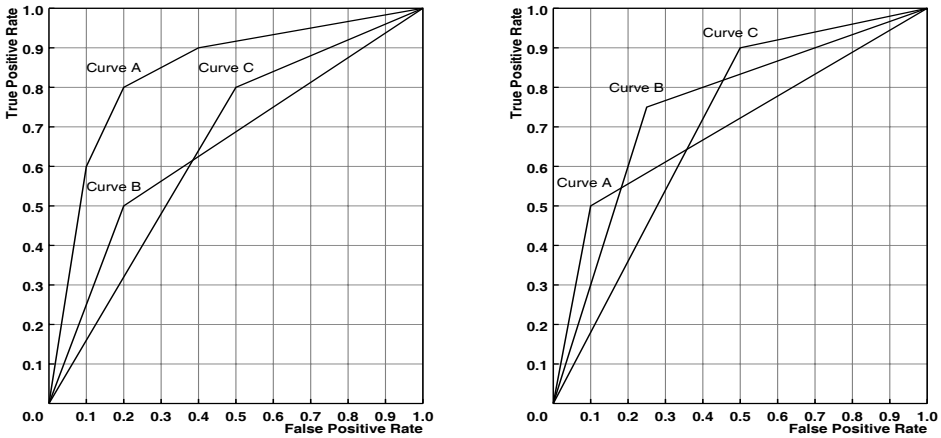

Figure 4: In (a), curve A is more "northwest" than curves B and C and thus dominates them. In (b), no single ROC curve dominates; in fact, each of the curves A, B and C dominates over a particular region of the combined ROC graph.

and then, depending on the prevailing conditions of *signal:noise* ratio and error-cost ratio, you can select the optimal detector's output for that range of conditions.

5.2 ROC convex hull (ROCCH)

The ROC convex hull (ROCCH) is a constructed "outer envelope" which includes that part of every ROC curve which dominates all other curves. If a single curve dominates all others over the entire graph, then it alone is the ROCCH. Otherwise, to construct the ROC convex hull you simply trace the outer envelope, making sure never to cave inwards or dip downwards, even if all other curves do at that point. If a potential dip occurs, then correct for that dip in the way shown previously in Figure 3. Thus, the ROCCH represents the best possible performance of a group of detectors, if you take the maximum accuracy of each and interpolate between different detectors whenever necessary to correct for any lulls.

An ROCCH which required the straight-line correction is pictured in Figure 5. Remember that in the correction, the straight line joining two or more detectors at their outermost points is an interpolation which can be achieved in practice if you alternate detector(s') response(s) proportionately between the two beneficial threshold points (the endpoints of the line segment, which for a graph of multiple ROC curves may lie on different empirical ROC curves). This holds even when the single point of a binary detector lies outside of the ROCCH constructed thus far; straight lines can be used to join it to the outermost points of the ROCCH lying on either side of it so that the remaining curve is always increasing and always convex. Don't forget that the points (0,0) and (1,1) may also be used in constructing the ROCCH.

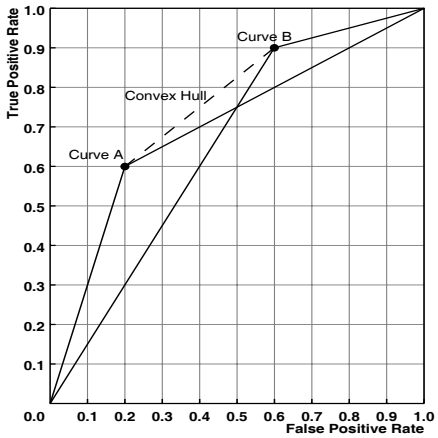

Figure 5: The convex hull dominates all curves on the graph. It either coincides with the single dominating curve, if one exists, or as shown here, it coincides with those parts of curves which dominate in a particular region. Wherever a "lull" exists in moving from one partial-dominator to another, the straight-line interpolation should be used in order to maximize total detector accuracy.

If you have several ROC curves, the best way to compare them easily and surely is to build the ROCCH and see which curves dominate over which regions of the graph. This method is robust in the face of imprecise class distribution and cost distribution estimates (Provost and Fawcett, 1997, 2001). As an aside, the "ROC convex hull method" was coined by Provost and Fawcett (1997); they make use of the QuickHull algorithm, which is provided by Barber et al. (1996). You can use the ROCCH to guide the construction of a hybrid classifier which outperforms any single classifier (assuming there is no single detector whose curve dominates all other curves); Provost and Fawcett (2001) discusses this concept.

5.3 Iso-performance lines

If you know either the *signal:noise* ratio (the class distribution) or the error-cost ratio (the cost distribution), or both, then you can find the slope of a line which represents the ideal bias to give a detector, in order to minimize cost of errors (see Section 4 for details on how to calculate such an ideal slope). Given the slope of such a line, you can see where it intersects a given ROC curve or an ROC convex hull; the point of intersection should be used as the ideal operating point (or decision threshold). Remember that the slope at an operating point is the same as the slope of the line segment that extends out to the left from it; that is, a line segment is associated with the upper endpoint (the upper endpoint has the same slope-at-a-point as the line segment stretching out below it). Of course if the ROC curve is smooth, then the slope is simply the slope of the line tangent to a point on the ROC

curve.

Wherever the ideal slope meets the ROCCH, you may use any of the detectors that constitute the ROCCH at that point, or any of the two (or more) whose end-points contribute to the making of the ROCCH envelope (whenever a straight line segment must be drawn to keep the envelope ever increasing and convex). Figure 6(a) shows a calculated iso-performance line which intersects both curve A and curve B (coinciding with the convex hull, in fact). For scenarios corresponding to this iso-performance line, the alternation of responses between A and B will be needed.

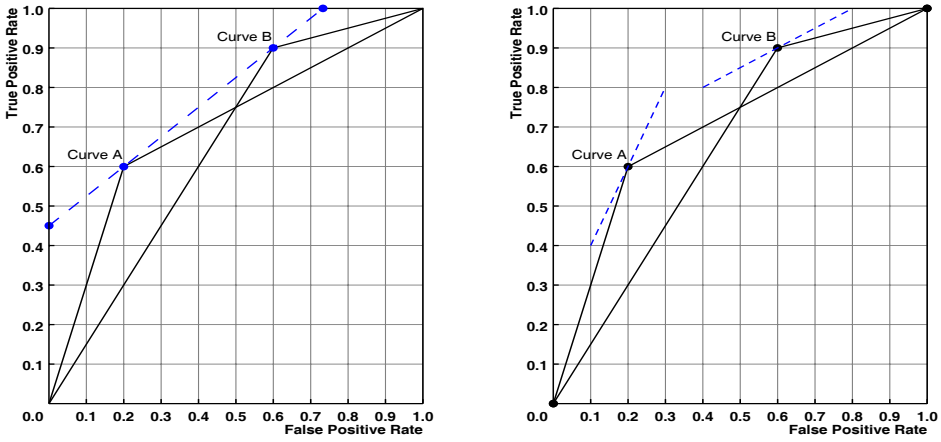

Figure 6: In (a), an iso-performance line coincides with the ROCCH at a stage between the ROC curves of two different detectors. In practice it would be most beneficial to utilize both detectors, alternating response between the two to achieve better accuracy than either alone. In (b), two different iso-performance lines are shown; each one is better served by a single detector acting alone.

Figure 6(b) shows two different iso-performance lines (or rather, portions of the lines). In each case, one curve is clearly better for that particular iso-performance line, because only one of the two curves meets the iso-performance line at a tangent on the convex hull. In this case you would choose the detector corresponding to the curve which the iso-performance line meets at a tangent on the convex hull (the envelope which lies most "northwest" along or outside of the ROC curves on the graph).

If the slope does not intersect the curve at all, i.e., it is steeper or flatter than any part of the curve, then it would be cheaper not to use a detector at all, but rather to say always "Yes" it's a signal (given a very flat slope < 1) or to always say "No" it's just a noise (given a very steep slope > 1). The only other recourse is to improve the signal detector (if possible) so that it meets the iso-performance line having the ideal slope.

By the way, if you have an estimated range of *signal:noise* and error-cost ratios,

then you will have a range of ideal biases which may be needed in reality (at different times, under changing conditions). These correspond to a range of slopes, and hence a family of iso-performance lines (Provost and Fawcett, 1997). Even given a range of slopes, you may be able to narrow down the choice of detectors to use by discarding any which don't intersect with the ROCCH and which don't intersect with the ideal range of slopes.

Keep in mind that when the *signal:noise* ratio or error-cost ratio are bound to change, or when they cannot be estimated in advance, it is good to keep around all classifiers which make up the portions of the ROC convex hull. If, however, the conditions are relatively stable, then you can use only the classifier (or classifiers) which coincide with the iso-performance line (or family of lines) needed.

6 Conclusion

6.1 Summary

This report has introduced the form and meaning of the ROC curve, as it is used to visualize and compare the accuracy and bias of one or more signal detectors tested on evaluation data. Though it has been just a starting point, and primarily for nonparametric analyses at that, it is hoped that at least the following points have been learned.

The primary advantages of the ROC are thus:

- It summarizes the entire (two-part) accuracy of a signal detector, over all operating thresholds
- It is independent of domain-specific particulars such as the *signal:noise* ratio and the error-cost ratio; this enables the most general comparison of detectors
- It provides a convenient visualization of the accuracy of a single detector, or of multiple detectors in comparison with one another
- It provides a single-number measure for accuracy
- It separates the adjustable detector bias from the discrimination capacity of the detector
- It can be used to help set an ideal decision threshold
- It can be used to help estimate the total rate of errors and the total cost of detection, when additional information is given
- It can be used to construct an improved accuracy result involving the combination of multiple detectors, each of which dominates the ROC-curve space for part of the graph

The ROC is particularly helpful when:

- You want to compare multiple detectors together on the same testing data set
- You lack some of the details of the final detection setting, such as event-class distribution and error-cost distribution

- You want to separate the detector's bias from the measurement of accuracy
- You want an entire view of the detector's accuracy throughout all of its possible biases

Caveats to using the ROC curve are:

- The result is only as good as the source data is representative, and as the ground-truth labels are correct
- Knowing whether, and how to generalize from one evaluated test set to another is not straightforward; using cross-validation may help, but you may never know everything about a future detection setting
- You should use the ROC-curve analysis methods properly; meaning, using nonparametric or parametric as appropriate; and being careful in the choice of whether, and how much curve-smoothing to apply

In general, the validity of generalization beyond a single ROC curve relies upon the following assumptions:

- The testing data sample is representative of the application domain, at least within the "signal" and "noise" categories
- The composition of the application domain does not change drastically over time; or, the sample composition represents the long-run composition of the domain
- The amount of data within each category (of signal and noise) is sufficient for the purposes of generalization to wider contexts
- The values of the testing data are continuously varying; that is, borderline cases are represented (if these occur in reality)
- The observation interval (or event) is defined in a reasonable way which leads to meaningful scoring and prediction of future error rates
- The label of "ground truth" is a gold standard (the best possible)

To the extent that these assumptions are not met, generalization should be withheld or at least considered cautiously.

6.2 Further Reading

The body of literature on ROC curves is vast, and still growing. Work began in earnest in the 1950s, and one still finds continuing work in applications and in statistics, as measures for smoothing ROC curves, estimating error, and calculating statistical significance are developed. Several citations appropriate to nonparametric analyses have already been given; this section will introduce sources especially for the other kind of analyses, parametric.

Parametric analyses of ROC curves have progressed a great deal, and much has been written about them. (Parametric analyses rely on distributional assumptions

about the underlying data, and they are not mentioned in the manual in detail.) Three suggested works which treat parametric ROC-curve analyses are Green and Swets (1966), Egan (1975), and Wickens (2002). Each of these books provides a comprehensive introduction to the theory of signal detection including the use of the ROC curve. Green and Swets (1966) emphasizes applications to psychophysics, which is the scientific study of sensory psychology. Note that John A. Swets is one of the biggest names associated with the use and advocation of the ROC curve; Swets et al. (2000) provides a good introduction, as well as a case for the increased use of the ROC curve in diagnostic situations. The book by Egan (1975) is more general than Green and Swets (1966); Egan emphasizes decision theory aspects, meaning the development of rules to inform decision making which satisfy a particular decision goal. Finally, Wickens (2002) presents the widespread aspects of the signal detection problem in a "from first principles" manner which might appeal to some readers. Each successive book also summarizes ideas felt to be important at the time of writing; the later books introduce some topics not present in the earlier books, while the earlier books add important historical context.

As a brief historical note, it is worth mentioning that the ROC curve was introduced during the Second World War to aid in the detection task of identifying enemy ships and planes on radar. In postwar years, its use widened as groups learned about, and then further developed, signal detection theory and its use of the ROC curve. Two key papers summarize the early signal detection theory at its fullest; these are Peterson et al. (1954) and Meter and Middleton (1954). They represent the two fields which contributed most to the theory: electronic communications, and mathematical statistics (in particular, hypothesis testing).

References

Barber, C. B., Dobkin, D. P., and Huhdanpaa, H. (1996). The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software (TOMS)*, 22(4):469–483.

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

Green, D. M. and Swets, J. A. (1966). *Signal detection theory and psychophysics*. Wiley, New York, NY.

Meter, D. V. and Middleton, D. (1954). Modern statistical approaches to reception in communication theory. *Transactions of the IRE Professional Group on Information Theory*, PGIT-4:119–141.

Peterson, W. W., Birdsall, T. G., and Fox, W. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, PGIT-4:171–212.

Provost, F. J. and Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions. In Heckerman, D., Mannila, H., Pregibon, D., and Uthurusamy, R., editors, *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*, pages 43–48, Menlo Park, California. AAAI Press.

Provost, F. J. and Fawcett, T. (2001). Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231.

Provost, F. J., Fawcett, T., and Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In Shavlik, J., editor, *Proceedings of the Fifteenth International Conference on Machine Learning (ICML-98)*, pages 445–453, San Francisco, CA. Morgan Kaufmann.

Swets, J. A., Dawes, R. M., and Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest (a supplement to Psychological Science)*, 1(1):1–26.

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of diagnostic systems: methods from signal detection theory*. Academic Press Series in Cognition and Perception. Academic Press, New York, NY.

Wickens, T. D. (2002). *Elementary signal detection theory*. Oxford University Press, Oxford.