ROC Confidence Bands : An Empirical Study[0]

Sofus A. Macskassy SMACSKAS@STERN.NYU.EDU
Foster Provost FPROVOST@STERN.NYU.EDU
New York University, Stern School of Business, 44 W. 4th Street, New York, NY 10012

Saharon Rosset SROSSET@US.IBM.COM
IBM T.J. Watson Research Center Yorktown Heights, NY 10598

Abstract

This paper is about constructing confidence bands around an ROC curve such that $(1-\delta)$% of the ROC curves traced by data sets of size r will fall completely within the bands. We introduce to the machine learning community three methods from the medical field that are applicable to generate such bands. We then evaluate these methods on the simple case of "binormal" distributions—the scores for positive and the score for negative instances are drawn from two normal distributions. We show that none of the methods generate appropriate bands and investigate two types of variances problems. We show that widening the bands does not produce the proper bandwidths but that fitting a normal distribution to the observed drawn samples and drawing samples from this distribution (parametric bootstrap) does generate bands that are much closer to the desired coverage although still not perfect. We tested the original methods as well as parametric bootstrap on the covertype data set from the UCI ML-repority. The original methods perform the same as in the synthetic case, whereas the parametric bootstrap technique did not yield the expected results. This is primarily due to not being able to generate a good fit for the score distributions. Whether it is possible to fit well-behaving parametric distribution to learned models is an open question we leave to the machine learning community to answer.

1. Introduction

In this paper we address the problem of creating confidence bands around ROC curves. When creating confidence bands, it is necessary to specify exactly what one is confident will be contained by the bands. For ROC curves there are various possibilities, and we consider one that is important for machine learning evaluations.

Many machine learning studies plot ROC curves to illustrate the possible tradeoffs of true-positive and false-positive rates that would be expected if a learned model were to be applied to data drawn from the same domain (distribution) as the training and testing data. We would like show the region where we are confident such a curve would lie. More specifically, since the variance of an ROC curve depends on the number of data on which it is based, we would like to plot a region that is expected to contain $(1-\delta)$% of the ROC curves traced by data sets of a given number r examples. This setting is notably different from the initial setting for the methods used in this paper, which were designed to place a confidence band on where the "true" ROC curve lies. The repercussions, as we shall see, are that the methods to not generate the proper bands. We will address two variance issues and propose methods to overcome them.

We first introduce the machine learning community to several existing methods, mostly from the medical literature, for computing confidence bands on ROC curves. This is a contribution itself, because rarely if ever do machine learning researchers plot confidence bands on ROC curves, and never do they use the more sophisticated of these methods. We then assess how well these methods work.

Since there has been almost no research on the assessment of confidence bands from ROC curves, and no research in a machine learning context (with the exception of the workshop paper by Macskassy and Provost (2004), that we here extend considerably), we start with a simple setting. We assume that it is desired to compute an ROC curve for a learned model (in the aforementioned setting), rather than for a learning algorithm. The latter is an important extension, but the former simpler question should be treated first. Furthermore, for purposes of ROC analysis, a learned model can be abstracted to the class-conditional score distributions it produces. We evaluate the various methods under the "binormal" assumption—that these score distributions are normally distributed, because some existing techniques make this assumption.

None of the methods produce appropriate confidence

[0]Macskassy, S.A. Provost, F.J., Rosset, S. "ROC Confidence Bands : An Empirical Study," CeDER Working Paper CeDER-05-12, Stern School of Business, New York University, NY, NY 10012. March 2005.

bands, either being too tight or too wide. We will explain why these results are as expected below. We point out two potential variance problems in the estimates, focusing on a technique for estimating the empirical distribution using bootstrap sampling. We introduce modifications that address these problems, showing that one of them produces more accurate confidence bands.

The remainder of the paper is outlined as follows: Section 2 describes relevant techniques for generating confidence bands, followed by a description of the methods used in this paper. Section 4 describes our synthetic data generation model, followed by the evaluation of these methods in Section 5. We finish in Section 6 with a discussion of open issues and other concluding remarks.

2. Overview of Existing Relevant Techniques

Prior work in machine learning on creating confidence intervals for ROC curves has for the most part created one-dimensional confidence intervals (cf. (Bradley, 1997; Provost et al., 1998; Fawcett, 2003)), which are not the focus of this paper–and generally are not useful for creating confidence bands around ROC curves (due in part to problems of multiple comparisons).

Use of the bootstrap (Efron & Tibshirani, 1993) as a robust way to evaluate expected performance has previously been suggested in related machine learning work—for evaluating cost-sensitive classifiers (Margineantu & Dietterich, 2000). In this work, the bootstrap was used to draw predictions $p(i, j)$, where $p(i, j)$ is the probability that an instance of class j was predicted to be in class i. Using these sample predictions, it was possible to generate final costs based on a cost-matrix. Repeated estimated costs were used to generate confidence bounds on expected cost.

Medical researchers have examined the use of ROC curves extensively and have introduced many techniques for creating confidence boundaries (intervals or bands). The problem domains and tasks in medical research are generally different from those of machine learning, in that they often consider only small data sets (where one instance is the test result from a patient). Further, it is often assumed that these data are ordinal in nature—e.g., that it is 'ratings' data with a small scale such as 'definitely diseased','probably diseased','possibly diseased',' 'possibly non-diseased','probably non-diseased','definitely non-diseased' (Beck & Shultz, 1986; Swets, 1988; Zweig & Campbell, 1993). We focus here on those methods that are directly applicable to the generation of confidence bands for continuous score distributions.

Creating a confidence region in ROC space restricts both FP and TP rates to the region $(0, 1)$. This restriction can cause difficulties when using intervals based on normal distributions. One solution is to transform the points to logit space[1], generate the confidence intervals in that space, and then convert them back into ROC space (Zou et al., 1997). An alternative transformation also used is that of converting to and from probit space[2] as done in the LABROC4 algorithms (Metz et al., 1998b; Metz et al., 1998a). Both of these bodies of work assume an underlying binormal distribution and focus on creating either one-dimensional confidence intervals, or joint confidence regions. We use LABROC4 to generate confidence bounds under the binormal distribution, as described in Section 3.3.3.

One method for generating simultaneous, or joint, confidence bands on ROC curves (Ma & Hall, 1993) is based on the Working-Hotelling hyperbolic confidence bands for simple regression lines (Working & Hotelling, 1929). Under the binormal model, an ROC curve can be parameterized as $\mathrm{TP} = \Phi(a - b\Phi^{-1}(\mathrm{FP}))$, where $\Phi(z)$ is the standard-normal cumulative distribution function (Dorfman & Alf, 1969). Using this parametrization, the Working-Hotelling bands can then be applied to ROC curves to generate simultaneous confidence bands. We describe our use of this method in Section 3.3.3.

The method of *simultaneous joint confidence regions* uses the distribution theory of Kolmogorov (Conover, 1980) to generate separate confidence intervals for TP and FP rates (Campbell, 1994). This is done by finding the Kolmogorov $(1 - \delta)$ confidence band for TP (tp$\pm d$) and FP (fp$\pm e$). By an independence assumption, the rectangle with width $2e$ and height $2d$, centered at a given point, should contain points at the given threshold with confidence $(1 - \delta)^2$. We describe our use of this method in Section 3.3.1.

The method of *fixed-width* simultaneous confidence bands is a non-parametric method, which generates simultaneous confidence bands by displacing the entire ROC curve "northwest" and "southeast" along lines with slope $b = -\sqrt{(m/n)}$, where m is the number of true positives and n is the number of true negatives (Campbell, 1994). This slope is an approximation of the ratio of the standard deviations for TP and FP—a property which tries to take into account the curvature of the ROC plot rather than using a displacement along one of the two axes as is done by the majority of methods described above. Campbell uses the bootstrap to create an empirical distribution from which to estimate the distance the curve should be displaced, thereby generating a "fixed-width" band across the complete curve (fixed-width with respect to the aforementioned slope). We describe how we use this method in Section 3.3.2.

[1] $logit(p) = \log(\frac{p}{(1-p)}); logit^{-1}(p') = \frac{1}{1+exp(-p')}$.

[2] $probit(p) = \Phi(p); probit^{-1}(p') = \Phi^{-1}(p')$, where $\Phi(z)$ is the cumulative normal distribution function.

3. ROC Confidence Calculations

In this section we describe our methodology for generating confidence bands for a classification model or modeling algorithm. We use three of the methods from the medical field: simultaneous joint confidence regions (SJR), Working-Hotelling based bands (WHB), and fixed-width confidence bands (FWB). FWB requires a set of ROC curves. These can be generated by evaluating the model on multiple testing sets or by resampling one test set. The resulting ROC curves will be used to generate confidence bands about an average curve.

3.1. Parametric vs non-parametric bands

Of course, parametric confidence calculations (intervals, bands, etc.) should be better when the distributional assumptions are correct—better in the sense that they should require fewer data to get the same tightness. In the case of ROC curves, the binormal-based approach has another advantage: a parametric confidence band is not of constant width, and thus takes into account the increased or reduced uncertainty at the ends of the scales (when FP, TP are both close to 0 or 1).[3] The major drawback of parametric confidence bands is, of course, their behavior when the underlying parametric assumptions are violated. In those cases, depending on the nature and extent of the violation, the bands may simply be meaningless.

A Kolmogorov-Smirnov non-parametric (SJR) band is of constant width, as it is based on a bound on the maximum distance possible between the true (mean) curve and the estimated curve. This bound does accommodate different width for the band in different regions of ROC space (more or less powerful models). As we will see, and as would be expected, the SJR bands typically are very conservative. It is possible, in principle, to build analytic non-parametric bands with different width in different regions; however this involves complicated theory and is unlikely to be practical.

Resampling-based bands (like bootstrap bands) tend to give variable-width confidence bands, but are not based on assumed distributional assumptions. They can be considered a compromise between parametric and non-parametric curves, in that they do consider the data to "estimate" a distribution on which to base the bands, but do not assume an a priori fixed parametric model. Resampling-based curves require considerable computation; however, this is becoming increasingly less relevant.

[3] Whether the uncertainty will be reduced or increased depends on the power of the scoring model and the marginal class distribution of the data. As Stein shows (Stein, 2002) the variance of an ROC curve is driven by the number of examples of the minority class. For a highly unbalanced class distribution, the ends of the scale will be much more balanced. On the other hand, for an initially balanced distribution, the ends of the scale for a powerful scoring model will be highly unbalanced.

3.2. Creating the Distribution of ROC Curves

The bootstrap (Efron & Tibshirani, 1993) is a standard statistical technique that creates multiple samples by randomly drawing instances, with replacement, from a host sample (the host sample is a surrogate for the true population). Each such set of samples can then be used to generate an ROC curve. For our setting, we can repeatedly draw r samples to generate a distribution of ROC curves. To our knowledge there is only one previous body of work which has applied the bootstrap to generate multiple ROC curves to use for estimating confidence bands (Campbell, 1994). Section 5.1 contains the details on how we use the bootstrap in our study.

3.3. Generating Confidence Bands

This section describes three methods to generate confidence bands across the complete ROC curve. The semantics for these confidence bands is that we would expect an ROC curve based on scores drawn from the same G^+ and G^- distributions that were used to generate the bands to fall completely within these bands with the specified probability (frequency). This is the basis for evaluating all the methods.

3.3.1. Simultaneous Joint Confidence Regions (SJR)

The simultaneous joint confidence region (SJR) uses the Kolmogorov-Smirnov (KS) (Conover, 1980) one-sample test statistic to identify a global confidence interval for TP and FP independently (Campbell, 1994). The KS statistic is used to test whether two sampled sets come from the same underlying normal distribution by considering the maximal vertical distance in their respective estimated cumulative density functions. For our purpose, that means the maximal vertical (horizontal) distance allowed from the given ROC curve to another ROC curve without rejecting H_0—*i.e.*, the confidence interval along FP (TP). Using the KS one-sample test allows us to identify these two distances, using the number of instances in each sample—*i.e.*, the number of true positives, m, and the number of true negatives, n. For sufficiently large set sizes (> 35), these distances are defined as follows.

We look up d and e, the critical distances along TP and FP respectively, at confidence level $(1-\delta)$. These identify the simultaneous joint confidence region for a given observed point $(\mathrm{fp}, \mathrm{tp})$ to be $(\mathrm{fp} \pm d, \mathrm{tp} \pm e)$ at confidence level $(1-\delta)^2$. Note that while the confidence level is theoretically $(1-\delta)^2$, we empirically test it as though it is at the $(1-\delta)$ level. We show that it generally is too conservative and that

	δ				
Set Size	0.20	0.15	0.10	0.05	0.01
> 35	$\frac{1.07}{\sqrt{n}}$	$\frac{1.14}{\sqrt{n}}$	$\frac{1.22}{\sqrt{n}}$	$\frac{1.36}{\sqrt{n}}$	$\frac{1.63}{\sqrt{n}}$

Table 1. Kolmogorov-Smirnov (KS) critical values for rejecting H_0 for set sizes > 35.

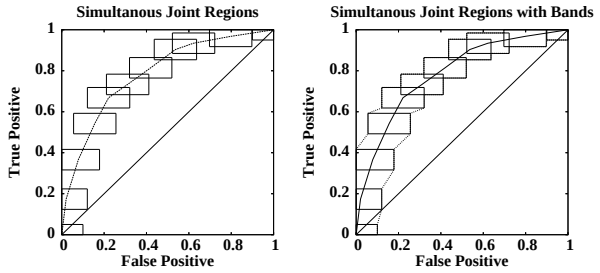

Figure 1. Transforming SJR into confidence bands.

$(1-\delta)^2$ would therefore be too loose.

The way we generate the confidence bands using these regions is by generating a confidence region for each distinct point on the ROC curve constructed from the scored samples in $\mathcal{D}$. We use the upper (lower) points of the confidence region to define the upper (lower) confidence band, cropped to stay within ROC space. Figure 1 illustrates this transformation.

3.3.2. FIXED-WIDTH BANDS (FWB)

The *fixed-width bands* (FWB) method works by identifying a slope, $b < 0$, along which to displace the original ROC curve to generate the confidence bands (Campbell, 1994). In other words, the upper (lower) confidence band would consist of all the points of the original observed ROC curve displaced "northwest" ("southeast") of their original location. This creates a confidence band with a fixed width across the entire curve. The question is what slope to use and what distance to displace the curve. While the ideal slope would be the ratio of the standard deviations associated, respectively, with TP and FP, we here adopt the same approximation as that used in the original work and use the slope $b = -\sqrt{(m/n)}$.

The way we generate the confidence bands using this method is by sweeping along the discrete points of the original ROC curve and adding upper (lower) boundary points by moving a distance d in each direction along the line with slope $b = -\sqrt{(m/n)}$. Figure 2 illustrates this transformation.

As with our study, the original work used the bootstrap to identify the distance to displace the curve to generate

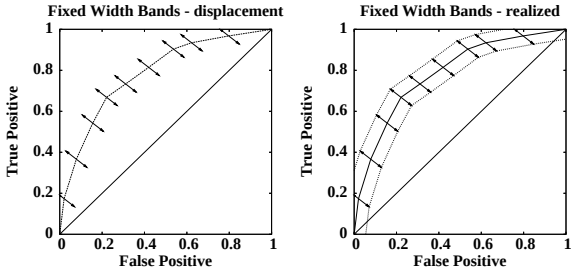

Figure 2. Displacing curve to generate FWB confidence bands.

the confidence bands. Given sample D, we generate bootstrap sample $D*$ and calculate the maximum distance along slope b from the ROC curve generated by D to the ROC curve generated by $D*$. We need the maximum distance because this is the width needed in order for $D*$ to be completely within the band. Sampling many $D*$'s, we can then find the distance needed in order to keep $1-\delta$ of all the curves completely within the generated bands. Note that the distances we sample are the distances along the given slope b.

3.3.3. SIMULTANEOUS WORKING-HOTELLING BANDS (WHB)

We adapt a method for using Working-Hotelling hyperbolic bands (Working & Hotelling, 1929) to generate simultaneous confidence bands on an ROC curve (Ma & Hall, 1993). The confidence bands are fitted to a regression line,

$$y = a - b \cdot x, \tag{1}$$

and are of the form

$$l(x, \pm k) = a - b \cdot x \pm k \cdot \sigma(x), \tag{2}$$

where $k \geq 0$ is a constant which we define below, and

$$\sigma(x) = \sqrt{\sigma_a^2 - 2\rho\sigma_a\sigma_b \cdot x + \sigma_b^2 \cdot x^2}, \tag{3}$$

as defined by the covariance matrix Σ:

$$\Sigma = \begin{pmatrix} \sigma_a^2 & \rho\sigma_a\sigma_b \\ \rho\sigma_a\sigma_b & \sigma_b^2 \end{pmatrix} \tag{4}$$

We use maximum-likelihood estimation (MLE) to generate a regression line to fit the ROC curve. We use a modified implementation of the LABROC4 algorithm (Metz et al., 1998a) to do so.[4] The LABROC4 algorithm works by first

[4]The LABROC4 fortran source code was acquired from a public web-site and modified to tailor its I/O to work with our ROC analysis toolkit. The toolkit, which we will release to the public later this year, is written in Java 1.5.

grouping continuous data into 'bins' or 'runs' of instances either with the same model score and/or same label. Then it uses an ordinal ('rating method') algorithm (Dorfman & Alf, 1969) to create a smooth binormal ROC curve. The covariance matrix is also calculated as part of the algorithm.

There are various constants, k, available at confidence level $(1-\delta)$, depending upon the type of band being generated. The original work describes two types of bands—pointwise and *simultaneous unrestricted*. To generate confidence bands, we use the wider simultaneous unrestricted Working-Hotelling bands (WHB), where, k_δ is determined by the chi-square distribution with 2 degrees of freedom:

$$k_\delta = \sqrt{-2ln(\delta)} \tag{5}$$

The confidence bands generated by this method are parametric in that they will calculate a TP-interval for any given FP. This is in contrast to the first two methods which have discrete bands based on the original ROC curve.

4. Data Generator

In order to facilitate a controlled experiment, we used a synthetic data generator such that we had complete control of G^+ and G^-. Prior work has shown that popular machine learning methods do not induce models which generate two normally distributed sets of scores but rather these scores take on score distributions which are a closer fit to asymmetric Laplace distributions or asymmetric Gaussian distributions (Bennett, 2003). The work further indicated that the score distribution for positive instances is generally "fatter" than the distribution for the negative instances. For this paper, we use two normal distributions, the positive being "fatter" than the negative. This is deliberate such that we can study the behavior of the confidence calculations in a close to ideal setting to evaluate whether they in fact generate the proper bounds when their assumptions are met (which we will see is not the case).

4.1. Synthetic World

We make the assumption that the only difference between G^+ and G^- are their model parameters. Our synthetic world therefore takes five parameters:

1. $P(+)$, the probability that an instance is from the positive distribution
2. the two model parameters for G^+, θ^+ and σ^+.
3. the two model parameters for G^-, θ^- and σ^-.

Each random sample drawn from this world has a probability of $P(+)$ for being drawn from G^+.

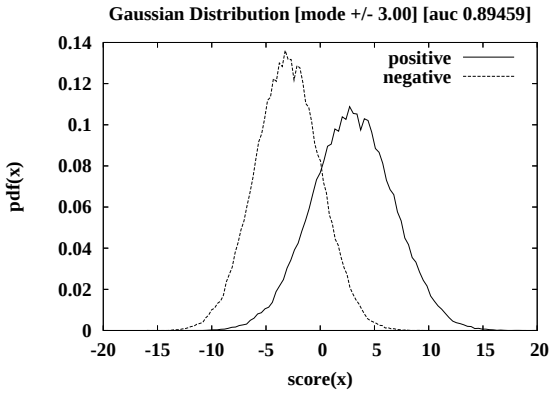

Figure 3. Example distribution used in study below.

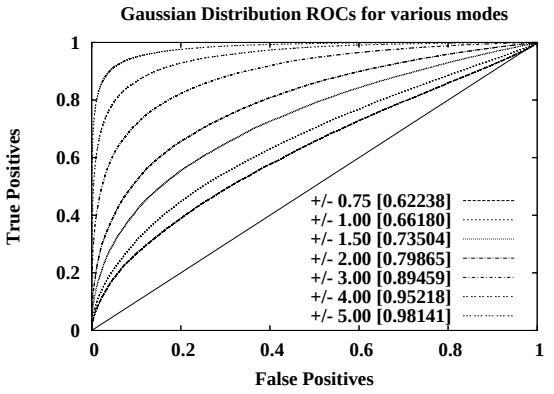

Figure 4. ROC curves generated for distribution as we vary θ.

For the study below, we fix $P(+) = 0.5$, $\sigma^+ = 3.75$, and $\sigma^- = 3.0$. We tested a range of values of θ's, where we set $\theta^+ = \{0.75, 1.00, 1.50, 2.00, 3.00, 4.00, 5.00\}$, and $\theta^- = -\theta^+$. Figure 3 shows this distribution when using $\theta = \pm 1.5$. Figure 4 shows the resulting ROC curves for all values of θ. These were generated by plotting the points $(\text{cdf}_{G^-}(x), \text{cdf}_{G^+}(x))$, for $x \in \{-\infty, \ldots, \infty\}$.

We investigate sensitivity of the confidence calculations to the difference in the modes of the distributions. We investigate this dimension because it will target sensitivity of the confidence calculations to the separability of G^+ and G^-. The closer the modes, the closer the true ROC curve is to the random line ($x = y$), and the further the modes are from each other, the "fatter" the true ROC curve is. The modes shown above were selected to yield a range of AUCs from 0.55 to 0.95.

5. Evaluation

We will examine how well the confidence bands fully contain ROC curves based on evaluation samples of r examples each, drawn from the same distribution(s).

5.1. Bootstrap-based Evaluation

To generate and evaluate confidence bands, we use the following method based on a bootstrapped empirical sampling distribution.

1. Build a synthetic world, $\mathcal{W}$, consisting of two distributions, G^+ and G^- with modes θ and $-\theta$ respectively, and $P(+)$ the probability that a randomly drawn sample comes from G^+. We set $P(+) = 0.5$ as defined above, with $\sigma^+ = 3.75$ and $\sigma^- = 3.0$.

2. Fix a sampling size, r, and sample from $\mathcal{W}$ a confidence-generation set, R, of size r.

3. Generate confidence bands:

   (a) **FWB:** Generate r_{fit} "fitting sets", F_i of size r by repeated sampling with replacement from R. For each F_i, generate an ROC curve, $\text{roc}(F_i)$, for the model. The result is a set of ROC curves, $\text{roc}_F = \{\text{roc}(F_i)\}$. Generate confidence bands, C_b, based on δ and roc_F.

   (b) **WHB,SJR**: Generate confidence bands, C_b, based on $\text{roc}(R)$, the roc curve based on R.

4. Generate r_{eval} "verification" sets, V_j, of size r by repeated sampling from $\mathcal{W}$. For each such sample, generate a verification ROC curve, $\text{roc}(V_j)$. The result is a set of ROC curves, $\text{roc}_V = \{\text{roc}(V_j)\}$.

5. Evaluate C_b. This is done by calculating the percentage of ROC curves in roc_V that fall completely within the generated confidence bands, C_b.

6. Repeat steps (2)–(5) 10 times to account for variability in the generated confidence calculations.

This methodology has five parameters: (1) the synthetic world, which is defined by G^+, G^-, and $P(+)$, (2) the ROC-generation size, r, (3) the number of sampling runs, r_{fit}, used to generate roc_F in step 3a to generate the confidence bands using FWB, and (4) the number of sampling runs, r_{eval}, used to generate roc_V, and (5) the confidence δ.

We fix r_{fit} and r_{eval} to 1000 and $\delta = 0.1$. We then examine the sensitivity of the confidence calculations to the ROC-generation size, $r \in \{25, 50, 100, 250, 500, 1000, 2500, 5000, 10000, 25000\}$ and the synthetic world used. We described the synthetic worlds we will use in Section 4.1.

5.2. Evaluating the Confidence Bands

Figure 5 shows the coverage for the 3 band methods. As we can see, JSR is too wide (albeit the best of the three), whereas WHB is far too wide and FWB is too tight.

None of the methods generate the proper bands.[5] This is not surprising and should be expected due to the subtle difference in setting between our "machine learning" setting and the setting for which the methods were designed. We would like to place a confidence band on the performance of the model in practice. These methods place a confidence band on where the "true" ROC curve lies.[6] Although we would not be able to assess the latter in an actual machine learning context, we can since here we know the true ROC curves (defined by the cumulative distribution functions). As we will clarify next, bands on where the true curve lies should be narrower, so the SJR and WH bands are even worse, but the FWB bands perform quite well—and can be recommended for this task (at least based on our synthetic world).

5.3. A variance problem with the bootstrap

For our setting, the FWB bands are too tight. Potentially, there are two types of variance problems with the bootstrap-based setting. The estimated variance about the curves could be incorrect, either because the original sample is not adequate (e.g., if the sample by chance contains all examples with the same score, the bootstrap will yield no variance), or because with the bootstrap there will be (by design) duplicate scores. Duplicate scores tend to lead to larger steps in the ROC curve (which is a step function for a finite sample), and thus to larger variance. However, comparisons of the variance from the bootstrap and from the actual distribution show that the bootstrap estimates are very close to the actuals.

Nevertheless, when we generated the fitting sets F_i in step 3a in Section 5.1 from $\mathcal{W}$ rather than from R, we see bands that converged to containments of $89 \pm 0.02\%$, very consistent across all sample sizes and all values of θ. The only difference is the use of the bootstrap.[7]

The second type of variance is in the observed data. Clearly, the observed data has a variance about the true curve, and therefore $\text{roc}(R)$ will not lie directly on the true curve. Since the bootstrap is estimating a confidence band about $\text{roc}(R)$, the bands generated will be off by the same amount as $\text{roc}(R)$ is from the true curve. Figure 6 illustrates the problem. The variances about the true and sam-

[5] We tested with other values of δ with similar results.

[6] We would therefore expect the bands to be too tight, as is the case for FWB, and not too wide as is the case with WHB and JSR.

[7] Conversely, we also saw that FWB, when centered on R, did in fact contain the true curve $\approx 90\%$ of the cases.

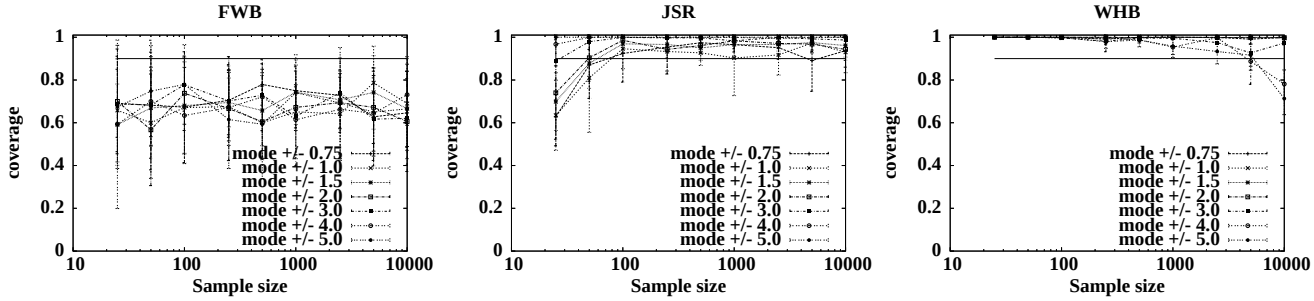

Figure 5. Coverage of bands at $\delta = 0.1$. The vertical line shows the expected coverage. As we can see, none of the methods generate the proper bands.

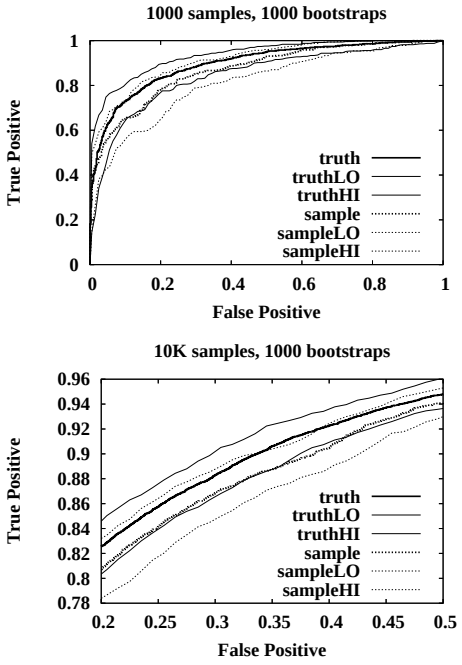

Figure 6. Variance problem with initial sample R. Variance about curves is correct, but the observed curve is off the true curve and the estimated bands are therefore off the proper region.

pled curves are very similar. However, because the sample is so far off from the true curve, the bands about it are clearly inappropriate for the purpose of estimating bands around expected future curves.

5.4. Accounting for the sample variance

There are two ways in which we might account for the sample variance. Either widen the band to take the variance into account or estimate what the true ROC curve should be and use the original bandwidth. We will take a look at both approaches in this section.

Number of	Absolute values of θ						
Samples	0.75	1.0	1.5	2.0	3.0	4.0	5.0
25	0.919	0.856	0.774	0.876	0.746	0.851	0.633
50	0.954	0.944	0.934	0.983	0.962	0.961	0.831
100	0.951	0.950	0.964	0.974	0.924	0.887	0.845
250	0.978	0.964	0.985	0.978	0.883	0.931	0.894
500	0.981	0.951	0.948	0.990	0.988	0.951	0.954
1000	0.990	0.981	0.961	0.982	0.974	0.968	0.957
2500	0.965	0.988	0.983	0.956	0.990	0.950	0.980
5000	0.975	0.993	0.958	0.962	0.992	0.951	0.981
10000	0.983	0.992	0.983	0.985	0.977	0.883	0.971
25000	0.992	0.988	0.982	0.990	0.991	0.942	0.976

Table 2. Coverages of FWB using a $(1 - \delta^2)$ bandwidth. These bands are too wide.

5.4.1. Widening the band

We first look at widening the band. Let us consider the true ROC curve (R_T), the sample ROC curve R_M from which we'll calculate the bands (B_M) of width d, and the ROC curves sampled subsequently (R_{M*}) which should with probability $(1 - \delta)$ lie within B_M.

We know that R_T falls outside B_M with probability no more than δ. Assuming that the variance about R_T is well estimated from R_M (which we have shown is the case), then R_{M*} will fall outside a band of width d around R_T with a probability no more than δ. Therefore R_{M*} will fall outside a $2d$ band around R_M with a probability no more than δ^2. In our case, we want $\delta^2 = 0.1$ and therefore need a $2d$ band around R_M with $\delta = 0.316$.

Table 2 shows the coverages we get from applying this technique, using $\delta = 0.1$ (and therefore using double the widths generated by using $\delta_M = 0.316$). We see that in general these bands are too wide. We next turn to estimating the true ROC curve as a way to overcome the sampling variance.

Number of Samples	Absolute values of θ 0.75	1.0	1.5	2.0	3.0	4.0	5.0
25	0.805	0.776	0.854	0.768	0.664	0.679	0.854
50	0.894	0.824	0.830	0.896	0.900	0.775	0.704
100	0.855	0.886	0.775	0.744	0.890	0.912	0.732
250	0.907	0.847	0.847	0.825	0.837	0.841	0.845
500	0.892	0.836	0.739	0.812	0.846	0.855	0.933
1000	0.873	0.873	0.803	0.832	0.771	0.916	0.900
2500	0.930	0.814	0.832	0.899	0.750	0.864	0.929
5000	0.855	0.914	0.773	0.858	0.892	0.876	0.859
10000	0.780	0.938	0.927	0.901	0.824	0.926	0.891
25000	0.933	0.933	0.932	0.893	0.944	0.943	0.910

Table 3. Coverages of FWB using parametric bootstrap and estimate of the true curve. The coverages are not consistently correct, but they are so the the ones closest to the proper coverage.

5.4.2. ESTIMATING THE TRUE ROC CURVE

In this approach, we try to estimate the true ROC curve. The true curve is generated by plotting $(\text{cdf}_{G^-}(x), \text{cdf}_{G^+}(x))$, for $x \in \{-\infty, \ldots, \infty\}$. However we don't know G^+ and G^-. If we could estimate these, then we could estimate the true curve. For example, we could assume that they are binormal (which we know is true for our synthetic world). Under this assumption, we can estimate θ and σ from the observed samples and use these to generate the true curve. Further, estimating $P(+)$ from the observed data we can now generate a new world, $\mathcal{W}'$ from which to estimate the distributions around the true curve. Thus, we generate $\mathcal{W}'$ both to estimate the true ROC curve as well as to draw samples from. We call this sampling technique *parametric bootstrap*.

Table 3 shows the coverages we get by using this technique. While the bands are too tight for small sample sizes, the coverages quickly do become respectable although still not as well fitted as we would have hoped. The bands do end up too wide with a few exceptions; however they are in much closer proximity to the coverage we desire than any of the other methods. On the other hand, parametric bootstrap requires that we can estimate the true score distributions G^+ and G^-.

5.5. Real Data

Estimating the score distributions produced by machine learning classifiers on real data is critical to parametric bootstrap. To our knowledge, there is almost no published work even mentioning these score distributions, let alone estimating them (the one exception being the work of Bennett, discussed above). Let us now examine the score distributions for models learned with the covertype data set from the UCI machine learning repository (Blake & Merz, 1998). Covertype examples are described by 54 features, 10 being numerical and the rest being binary. To produce a binary classification problem (for ROC analysis) we chose the two classes with the most instances (yielding 57.2% base accuracy).

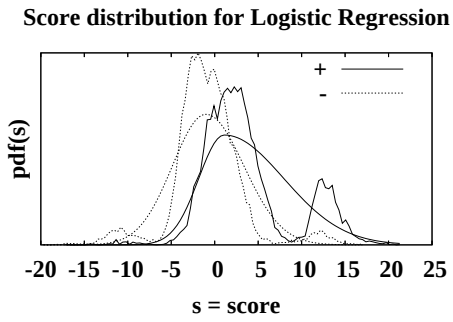

Figure 8. Sample logistic regression score distribution with $r = 25000$. These distributions are clearly not unimodal.

We randomly sampled 100 instances (50 of each class) and built various learned models using Weka[8] (Witten & Frank, 2000)—logistic model trees (LMT) (Landwehr et al., 2003), J48, naive Bayes trees (NBT) (Kohavi, 1996), logistic regression (LR), and Naive Bayes (NB). We then generated prediction scores for the remaining $490,000$ instances. The log-odds scores, $\log \frac{P(+|x)}{P(-|x)}$, were used as the base population R from which to draw predictions. Figure 7 shows the distributions of scores generated for each method for positives and negatives. Although LMT, LR and NB have nice smooth distributions, they are clearly not binormal and the distributions of J48 and NBT are even further from gaussian distributions. The naive Bayes distributions are more-or-less in line with observations made by Bennett (Bennett, 2003) (who studied naive Bayes for text classification): they are asymmetric bell-shaped distributions, but the negative distribution is fatter than the positive (which should not be important here).

For parametric bootstrap, we fit the scores to asymmetric Gaussian distributions. We used the evaluation methodology presented above using the same values for the roc-generation sampling size r. For each run, we sampled a fitting set F_i of r points from R and then sampled our verification sets V_i from $R \cap F_i$.

Using the original methods yielded the same results as before, where FWB was too tight, while SJR and WHS-s were too wide. The coverages we get from using the parametric bootstrap and ROC estimation technique are shown in Table 4.

[8] We use version 3.4.2. Weka is available at `http://www.cs.waikato.ac.nz/~ml/weka/`

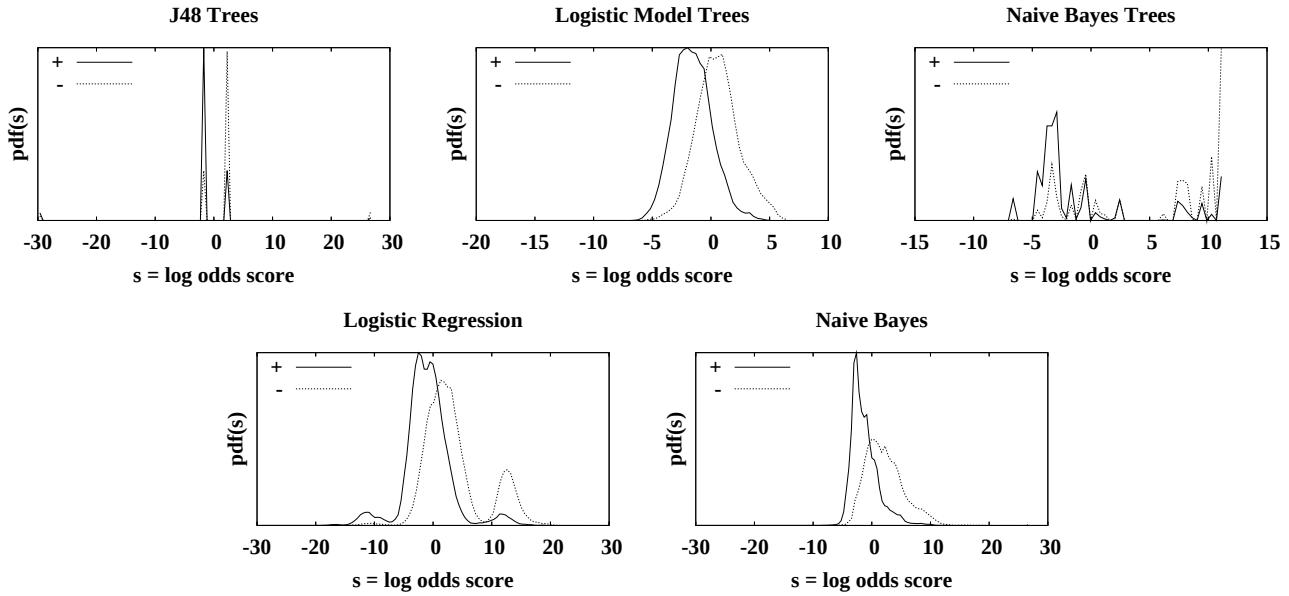

Figure 7. Score distributions of 5 Machine Learning methods on the covertype data set.

Number of Samples	J48	LMT	NBTREE	LR	NB
25	0.605	0.587	0.550	0.787	0.797
50	0.722	0.693	0.743	0.739	0.883
100	0.557	0.816	0.884	0.779	0.827
250	0.121	0.862	0.848	0.613	0.640
500	0.000	0.823	0.813	0.692	0.574
1000	0.000	0.791	0.487	0.340	0.511
2500	0.000	0.751	0.229	0.213	0.100
5000	0.000	0.727	0.000	0.002	0.008
10000	0.000	0.696	0.000	0.000	0.000
25000	0.000	0.424	0.000	0.000	0.000

Table 4. Coverages of FWB using parametric bootstrap to generate bands based on prediction scores from the 5 machine learning methods on the covertype data sets

We look first to the distributions to see if the problem lies in badly estimated distributions. This turns out to be the case, as is clear in Figure 8. While the empirical scores form a nice smooth distribution, the distribution is not uni-modal. These peaks in the score distribution translate directly to peaks in ROC space, which the smoothed asymmetric gaussian distribution did not take into account. In order for the parametric bootstrap technique to work, we must obviously be able to genearet a better fit than the asymmetric gaussian. Whether this is possible is still an open question.

6. Discussion

In this paper we assessed various methods for generating confidence bands for ROC curves, in the context of machine learning evaluations. For computing a confidence band on where the "true" ROC curve for a model lies, a resampling technique (FWB) from the medical literature performs well, but (somewhat surprisingly) two other techniques (WHB and JSR) are far too conservative–even when the underlying parametric assumptions are correct.

However, machine learning studies often are interested in the future performance of the model. None of the methods are designed to produce accurate confidence bands for this task, and none do. After showing that the variance estimated by bootstrap sampling is reasonable, we introduce two techniques to address the variance introduced by producing bands about the ROC curve generated from a particular test set. One method, parametric bootstrap, attempts to estimate where the true ROC curve lies, and then put a band around it. This method yielded the best bands on synthetic data for which the parametric assumption (a binormal score distribution) was appropriate.

We also evaluated this methodology on the UCI Covertype data set. We generated models from 5 different learning methods and used those models to generate a large set of scored predictions. We then used our methodology to generate, and evaluate, confidence bands based on these prediction scores. The parametric bootstrap did not perform well. We showed that the bad performance was due to

not being able to fit the score distributions well enough. Whether it is possible to generate better parametric estimates of the scoring distributions is an open question.

Acknowledgments

We would like to thank Tom Fawcett for his pointers to related work and for many discussions about ROC curves, an anonymous early reviewer for directing us to additional medical literature we were unaware of, Michael Littman for initial discussions on ROC evaluations, Haym Hirsh for his feedback early in the design stages and Matthew Stone who initially suggested using the bootstrap for evaluating ROC curves.

This work is sponsored in part by the National Science Foundation under award number IIS-0329135. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the National Science Foundation or the U.S. Government.

References

Beck, J. R., & Shultz, E. K. (1986). The use of relative operating characteristic (ROC) curves in test performance evaluation. *Archives of Pathology and Laboratory Medicine*, *110*, 13–20.

Bennett, P. N. (2003). Using Asymmetric Distributions to Improve Text Classifier Probability Estimates. *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Toronto, Canada: ACM Press.

Blake, C. L., & Merz, C. J. (1998). UCI Repository of machine learning databases.

Bradley, A. P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *7*, 1145–1159.

Campbell, G. (1994). Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. *Statistics in Medicine*, *13*, 499–508.

Conover, W. J. (1980). *Practical nonparametric statistics*. New York: Wiley. 2nd edition.

Dorfman, D. D., & Alf, E. (1969). Maximum-likelihood estimation of parameters of signal-detection theory and determination of confidence intervals–rating method data. *Journal of Mathematical Psychology*, *6*, 487–496.

Efron, B., & Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Fawcett, T. (2003). *ROC Graphs: Notes and Practical Considerations for Data Mining Researchers*Technical Report HPL-2003-4). HP Labs.

Kohavi, R. (1996). Scaling Up the Accuracy of Naive-Bayes Classifiers: a Decision-Tree Hybrid. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*. Menlo Park, CA: AAAI Press/MIT Press.

Landwehr, N., Hall, M., & Frank, E. (2003). Logistic Model Trees. *Proceedings of the 16th European Conference on Machine Learning*.

Ma, G., & Hall, W. J. (1993). Confidence bands for receiver operating characteristic curves. *Medical Decision Making*, *13*, 191–197.

Macskassy, S., & Provost, F. (2004). Confidence Bands for ROC Curves: Methods and an Empirical Study. *Proceedings of the First Workshop on ROC Analasis in AI (ROCAI-2004) at ECAI-2004*.

Margineantu, D. D., & Dietterich, T. G. (2000). Bootstrap methods for the cost-sensitive evaluation of classifiers. *International Conference on Machine Learning, ICML-2000* (pp. 582–590).

Metz, C. E., Herman, B. A., & Roe, C. A. (1998a). Statistical Comparison of Two ROC-curve Estimates Obtained from Partially-paired Datasets. *Medical Decision Making*, *18*, 110–121.

Metz, C. E., Herman, B. A., & Shen, J.-H. (1998b). Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. *Statistics in Medicine*, *17*, 1033–1051.

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

Stein, R. (2002). *Benchmarking Default Prediction Models: Pitfalls and Remedies in Model Validation* (Technical Report #030124). Moody's KMV.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*, 1285–1293.

Witten, I. H., & Frank, E. (2000). In *Data Mining: Practical machine learning tools with Java implementations*. San Francisco: Morgan Kaufmann.

Working, H., & Hotelling, H. (1929). Application of the theory of error to the interpreation of trends. *Journal of the American Statistical Assocation*, *24*, 73–85.

Zou, K. H., Hall, W. J., & Shapiro, D. E. (1997). Smooth nonparametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, *16*, 2143–2156.

Zweig, M. H., & Campbell, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry*, *39*, 561–577.