Model Selection via the AUC

Saharon Rosset SROSSET@US.IBM.COM
IBM T.J. Watson Research Center, P.O. Box 218, Yorktown Heights, NY 10598

Abstract

We present a statistical analysis of the AUC as an evaluation criterion for classification scoring models. First, we consider significance tests for the difference between AUC scores of two algorithms on the same test set. We derive exact moments under simplifying assumptions and use them to examine approximate practical methods from the literature. We then compare AUC to empirical misclassification error when the prediction goal is to *minimize future error rate*. We show that the AUC may be preferable to empirical error even in this case and discuss the tradeoff between approximation error and estimation error underlying this phenomenon.

1. Introduction: ROC Analysis and the AUC

The term Receiver Operating Curve (ROC) has long been used in the signal processing (Egan, 1975) and medical (Hanley & McNeil, 1982) literature to describe a curve displaying the relationship between sensitivity and 1-specificity at all possible thresholds for a 2-class classification scoring model, when applied to independent (test) data. Sensitivity (or TP, for True Positive rate) and specificity (or TN, for True Negative rate) are defined as follows:

$$TP = \frac{\#number\ of\ observations\ predicted\ as\ positive}{\#positive\ observations}$$
$$TN = \frac{\#number\ of\ observations\ predicted\ as\ negative}{\#negative\ observations}$$

The area under the ROC curve (AUC) is a one-number measure of a model's discrimination performance, i.e., the extent to which a model successfully separates the positive and the negative observations and "ranks" them correctly. This number is considered of great interest in various engineering, scientific and medical domains. In recent years, there has been a surge of interest in the AUC as an evaluation measure in the Data Mining and Machine Learning communities. Its statistical properties have been investigated in some recent papers: (Provost & Fawcett, 1997) illustrate the "robustness" of ROC analysis, and the AUC in particular, against changing class balance; (Ling et al., 2003) define a rigorous discrimination measure, under which the AUC is provably superior to the empirical misclassification rate as an evaluation measure, in that it is less prone to ties when evaluating non-equivalent models; (Cortes & Mohri, 2003) investigate the relationship between error rate minimization and AUC maximization by analyzing the range of possible AUCs when the error rate is fixed. They also discuss algorithms that directly maximize AUC.

In this paper we offer two new statistical insights about the AUC. We aim to increase understanding of this evaluation measure and its advantages, and to present a new tool for its analysis. In section 2 we tackle the issue of comparing AUC scores of two models when using the same evaluation or test set. We derive an exact expression for the moments of the difference between the scores, subject to some simplifying assumptions required to make the calculation feasible. We use our results to evaluate the performance of practical methods suggested in the AUC literature in the 80's. In section 3 we consider the situation when our underlying goal is the standard classification goal, i.e., to *minimize error rate*. We show that even in that case, the AUC may be a better model selection criterion than empirical error rate because it is more stable (incurs smaller estimation error), despite the fact that it is biased (incurs larger approximation error).

2. Significance Testing of the Difference between AUC Scores

If we intend to use AUC as a comparison and discrimination method for scoring models, we would like to be able to say not only which model performs better for

Appearing in *Proceedings of the 21st International Conference on Machine Learning*, Banff, Canada, 2004.

the given test set, but also whether its performance is significantly better than that of other models.

In this section we first develop limited theory-based exact moment calculations for the difference between two AUC scores based on the same test set. Because knowledge of the underlying probability structure is required, our results are not applicable in real-life situations. However, they can be used to test the performance of practical estimators for these moments in simulations, where the underlying structure is known.

We then introduce two approximate significance tests (Hanley & McNeil, 1983; DeLong et al., 1988), developed in the 1980's, in the context of medical experiments. We examine the usefulness of these approximate tests by comparing them to the exact theoretic derivation for some synthetic examples. The non-parametric test suggested by (DeLong et al., 1988) performs better on our examples and we second the authors of that paper in preferring their method.

In what follows, we assume we have a test set of size n: $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$ with $y_i \in \{-1, +1\}$. Assume further the test set has n_+ positive examples and n_- negative examples. Denote $p_i = P(y_i = 1|x_i) = f(x_i)$, and let $m_1(x)$ and $m_2(x)$ be two scoring models.

2.1. Exact Moments

We now derive the explicit formulae for the mean and variance of the difference between two AUCs, conditioned on the set of test x values. For every scoring model m_k we define several properties of the underlying score distribution and we calculate the moments as a function of those. Specifically, let $I_{k,ij}$ to be an indicator for the event that scoring model m_k gives a higher score to observation i in our test set than to observation j. We then define:

$$\begin{aligned}
p_{k,ij} &= E(I_{k,ij}) \\
p^{(1)}_{k,ijl} &= P\{m_k(x_l) > m_k(x_j) > m_k(x_i)\} = \\
&= E\{I_{k,ij} \cdot I_{k,il} \cdot I_{k,jl}\} \\
p^{(2)}_{k,ijl} &= P\{\min(m_k(x_l), m_k(x_j)) > m_k(x_i)\} = \\
&= E\{I_{k,ij} I_{k,il} I_{k,jl} + I_{k,ij} I_{k,il}(1 - I_{k,jl})\} \\
p^{(3)}_{k,ijl} &= P\{m_k(x_l) > \max(m_k(x_j), m_k(x_i))\} = \\
&= E\{I_{k,ij} I_{k,il} I_{k,jl} + (1 - I_{k,ij}) I_{k,il} I_{k,jl}\}
\end{aligned}$$

We also define the corresponding q quantities as $1-p$. That is, $q_{k,ij} = 1 - p_{k,ij}$, $q^{(1)}_{k,ijl} = 1 - p^{(1)}_{k,ijl}$ etc.

2.1.1. Assumptions

The derivation of the exact distribution of the difference between two AUC scores in the most general case is extremely complex. To make it feasible within this framework we use two simplifying assumptions about the probability structure underlying the process.

First, we assume the processes underlying the "binary order switches" for the two models are independent: $Pr\{I_{1,ij} = 1 \cap I_{2,ij} = 1\} = Pr\{I_{1,ij} = 1\} \cdot Pr\{I_{2,ij} = 1\}$. This assumption is very sensible if the models are "independent" — if they use completely different information to calculate the scores, and the dependence is only through the use of the same test data. In comparing "similar" models this assumption may be violated. Such a violation would tend to decrease the variance of the difference between the scores for the two models, hence increase the chance of correct discrimination. So if this assumption is violated towards positive correlation between the binary switches then our results below give a "conservative approximation" of the real probability of identifying the better model.

Second, we assume that for each model, the "binary switch" behaviors for disjoint pairs of observations are independent:
$Pr\{I_{1,ij} = 1 \cap I_{1,uv} = 1\} = Pr\{I_{1,ij} = 1\} \cdot Pr\{I_{1,uv} = 1\}$ if i, j, u, v are all different. This assumption is problematic in general, since for practically any real-life model it should seem likely that the bias in a model's prediction at a certain x covariate vector would be related to the bias in its neighborhood. For example: in a k-NN model, if at a certain x-covariate value most of the neighbors are class-0, it is highly likely that the same would apply to other x-values in the "neighborhood". Thus knowing that a certain observation has taken part in a "switch" would increase the likelihood of the same being true for its neighbors. However, due to the fact that our derivation is not applicable for real-life situations but only for hand-crafted examples, we prefer to make this assumption, to make calculations manageable.

Finally, we use our moment calculations to estimate probabilities of correct discrimination, implicitly assuming normal distribution. The difference between AUCs is clearly not exactly normal, however it is asymptotically normal. To see this, consider that AUC scores are well established to be asymptotically normal and therefore their difference is too, unless they are perfectly correlated.

2.1.2. Moment calculations

It is well known that the AUC is equivalent to the Mann-Whitney statistic and can be represented as:

$$AUC(m) = \frac{1}{n_+ n_-} \sum_{y_i=1} \sum_{y_j=-1} I\{m(x_i) > m(x_j)\} =$$

$$= \frac{1}{2 \cdot n_+ \cdot n_-} \sum_{i=1}^{n} \sum_{j>i} (y_i - y_j) I_{m,ij}$$

The difference between the two AUC scores on the same data set can thus be expressed as a function of the binary order switches :

$$2n_+n_- \cdot (AUC(m_2) - AUC(m_1)) = \tag{1}$$
$$= \sum_{i=1}^{n} \sum_{j=i+1}^{n} [I_{1,ij} - I_{2,ij}][y_j - y_i]$$

To derive the mean and variance of this expression, we should note that we have here 2 different, *independent*, levels of stochasticity, one in "selecting" the model scores (which determine the "order switches") and the other in drawing the actual y-values for the test set. For the mean we thus get:

$$E[\cdot n_+ \cdot n_- \cdot (AUC(m_2) - AUC(m_1))] =$$
$$= \frac{1}{2} \sum_{i=1}^{n} \sum_{j=i+1}^{n} E[I_{1,ij} - I_{2,ij}] E[y_j - y_i] =$$
$$= \sum_{i=1}^{n} \sum_{j=i+1}^{n} [p_{1,ij} - p_{2,ij}][p_j - p_i]$$

To calculate the variance we take advantage of the formula for two level variance:
$V_{X,Y}(f(X,Y)) = V_X[E(f(X,Y)|X] + E_X[V(f(X,Y)|X)]$

In our case that leads us to:

$$V[2n_+n_-(AUC(m_2) - AUC(m_1))] =$$
$$= V\{\sum_{i=1}^{n} \sum_{j=i+1}^{n} [I_{1,ij} - I_{2,ij}][y_j - y_i]\} =$$
$$= E_Y \left[V\{\sum_{i=1}^{n} \sum_{j=i+1}^{n} [I_{1,ij} - I_{2,ij}][y_j - y_i] | \mathbf{y}\} \right] +$$
$$+ V_Y \left[E\{\sum_{i=1}^{n} \sum_{j=i+1}^{n} [I_{1,ij} - I_{2,ij}][y_j - y_i] | \mathbf{y}\} \right]$$

A tedious analysis of each one of the two expressions follows, which eventually leads us to the following exact (if somewhat inelegant) variance calculation:

$$V[n_+n_-(AUC(m_2) - AUC(m_1))] =$$
$$= \sum_{i<j} [p_i q_j + p_j q_i][p_{1,ij} q_{1,ij} + p_{2,ij} q_{2,ij}] -$$
$$-2 \sum_{i<j<l} [p_i q_j p_l + q_i p_j q_l] \cdot$$
$$\cdot [p^{(1)}_{1,ijl} - p_{1,ij} p_{1,jl} + p^{(1)}_{2,ijl} - p_{2,ij} p_{2,jl}] +$$
$$+2 \sum_{i<j<l} [p_i q_j q_l + q_i p_j p_l] \cdot$$
$$\cdot [p^{(2)}_{1,ijl} - p_{1,ij} p_{1,il} + p^{(2)}_{2,ijl} - p_{2,ij} p_{2,il}] +$$
$$+2 \sum_{i<j<l} [p_i p_j q_l + q_i q_j p_l] \cdot$$
$$\cdot [p^{(3)}_{1,ijl} - p_{1,il} p_{1,jl} + p^{(3)}_{2,ijl} - p_{2,il} p_{2,jl}] +$$
$$+ \sum_{i<j} (p_{1,ij} - p_{2,ij})^2 [p_j q_j + p_i q_i] +$$
$$+2 \sum_{i<j<l} [(p_{1,ij} - p_{2,ij})(p_{1,il} - p_{2,il}) p_i q_i +$$
$$+ (p_{1,il} - p_{2,il})(p_{1,jl} - p_{2,jl}) p_l q_l +$$
$$+ (p_{1,ij} - p_{2,ij})(p_{1,jl} - p_{2,jl}) p_j q_j]$$

2.1.3. Example: Calculating the Moments

We illustrate the use of the above formulae for the moments of the difference between two AUC scores. Let us take the simple case $p_i = i/n$, and take two models whose scores are exactly the p_i's, except for possible deviations of fixed size c, which occur with a fixed probability q_k that depends on the model. Note that we have no explicit x covariates at all in this synthetic example. Thus the distribution of the scores for $k = 1, 2$ is:

$$m_k(i) = \begin{cases} \frac{i}{n} & \text{w.p } 1 - 2q_k \\ \frac{i}{n} + c & \text{w.p. } q_k \\ \frac{i}{n} - c & \text{w.p. } q_k \end{cases}$$

(These scores are not valid probability estimates, of course, but are legitimate as scores for ranking.)

The better model is the model with the smaller q because it has, on average, a more correct ranking (as well as a lower misclassification rate with a threshold of 0.5). We would like to utilize our formulae to calculate the moments of the distribution of $n_+n_-(AUC(m_2) - AUC(m_1))$. For this we need to evaluate the generic probabilities defined at the beginning of section 2.1 (in all the calculations below we assume implicitly that $i < j < l$):

$$p_{k,ij} = \begin{cases} q_k(2 - 3q_k) & \text{if } j - i < c \cdot n \\ q_k^2 & \text{if } c \cdot n < j - i < 2c \cdot n \\ 0 & \text{if } j - i > 2c \cdot n \end{cases}$$
$$p^{(1)}_{k,ijl} = \begin{cases} q_k^2(1 - 2q_k) & \text{if } max(j - i, l - j) < c \cdot n \\ 0 & \text{otherwise} \end{cases}$$

etc.

Using the asymptotic normality of the AUCs themselves we can then approximate $Pr(AUC(m_2) -$

Table 1. Moments of the re-scaled difference between two AUC scores in the setup of section 2.1.3.

N	q_1	q_2	Mean	Variance	Φ
50	0.1	0.2	8.36	476	0.65
200	0.1	0.2	127.86	28577	0.78
400	0.1	0.2	507.46	226097	0.86
50	0.1	0.3	15.99	592	0.74
200	0.1	0.3	246.39	35602	0.90
400	0.1	0.3	979.19	281769	0.97

$AUC(m_1) > 0$). Table 1 displays the results for various values of n, q_1 and q_2, with c fixed at 0.2005 (i.e., $0.2 + \epsilon$ to prevent ties). The column titled Φ refers to the normal-tail probability approximation for the event $AUC(m_2) < AUC(m_1)$.

2.2. Empirical Evaluation of Methods from AUC Analysis

As mentioned in section 2.1.2, the AUC is equivalent to the Mann-Whitney 2-sample statistic. A well known derivation exists for the moments of this statistic under the "alternative" that the two classes do not follow the same distribution (Lehmann, 1975). The mean of the AUC is p_1 and the variance is:

$$\frac{p_1(1-p_1) + (n_+ - 1)(p_2 - p_1^2) + (n_- - 1)(p_3 - p_1^2)}{n_+ n_-}$$

with p_1, p_2, p_3 now representing various probabilities which depend on the probability structure of the scores of the 2 classes under the given model, as follows:
$p_1 = \Pr$ {A random positive case attains a higher score than a random negative case}
$p_2 = \Pr$ {Two random positive cases attain a higher score than a random negative case}
$p_3 = \Pr$ {A random positive case attains a higher score than two random negative cases}
These probabilities can be estimated directly from the data, or calculated using assumptions about the parametric "prior" distribution of the classes.

What we want, however, is to test the significance of the difference between the AUCs for two different models. For this, we need to approximate the covariance between the two AUCs calculated on the same test data, to obtain the variance of the difference:
$V(AUC(m_1)\text{-}AUC(m_2)) = V(AUC(m_1)) + V(AUC(m_2))$ -
- $2Cov(AUC(m_1), AUC(m_2))$

(Hanley & McNeil, 1982) discussed estimation of the variances. They suggested various methods —- mostly parametric methods, based on assuming that the scores for the two classes have some known "prior" distribution. They concluded empirically that the best approach is to assume that the "prior" score distributions are Exponential. Following this, (Hanley & McNeil, 1983) suggested methods of estimating the correlation between two AUC scores on the same test set. They did not give explicit formulae for calculating the correlation, and just supplied a tabulation of the correlation as a function of the correlations between scores for cases from the same class in the two models and the AUC scores themselves. If, for example, the correlation between the scores which the two models give class-1 observations is 0.87, the correlation for class-0 observations is 0.83 and the two AUC scores are 0.7 and 0.8, the table tells us that the correlation between the AUC scores will be 0.81. Given estimates for the variances of the AUC scores, we can use this to estimate the covariance. (DeLong et al., 1988) suggested the use of U-statistic methods to approximate this covariance without any parametric assumptions (the Mann-Whitney statistic, hence the AUC, is a U-statistic, of course). They employed a method suggested by (Sen, 1960) for approximating the full covariance matrix for the AUC scores. This method assures that the estimates are consistent. We can therefore expect to have low bias for these estimates, but can expect to have increased variance, because without parametric assumptions the dependence of the estimates on the data naturally increases.

Table 2. Simulation results of variance-covariance estimation methods.

n, q_2	HM	DDC	Actual
50, 0.2	635 ± 286	464 ± 276	476
200, 0.2	39013 ± 8113	29415 ± 7915	28577
400, 0.2	311579 ± 45112	236852 ± 42915	226097
50, 0.3	729 ± 290	571 ± 290	592
200, 0.3	48478 ± 9340	37641 ± 8613	35602
400, 0.3	390504 ± 50421	304247 ± 50689	281769

We use these methods to estimate the variance of the difference between the AUCs using data generated according to the distribution described in section 2.1.3. We then compare them to the exact calculation of table 1. Table 2 shows the results of multiple experiments in estimating the variance using the two covariance estimation methods — Hanley and McNeill's (HM's) "parametric" method and DeLong et al.'s (DDC's) "asymptotic" method. We have performed 1000 simulations with each method for each (n, q_1, q_2) triplet. The table presents the empirical mean and 95% CI of the estimator for $Var(AUC(m_1) - AUC(m_2))$ using the two covariance approximation methods.

If indeed our goal is to find an estimation method which is more consistent and less biased for $Var(AUC(m_1) - AUC(m_2))$, we can see (not surprisingly) that the DDC estimator's average value tends to be much nearer to the actual variance than the HM estimator's average. The HM estimator usually overestimates the variance as a result of under-estimating the covariance. This is consistent with the results of DDC who also compared their method to HM's on a real-life example and got a larger covariance estimate (hence smaller variance estimate) using their own method. The surprising result, however, is that the empirical variance of the DDC estimator is usually not much larger than that of the HM estimator, and sometimes even smaller. This seems to be another indication that the DDC estimator is more appropriate. However we feel more experimentation is required to establish the unequivocal superiority of DDC to HM in practical situations.

3. Using AUC to Evaluate Classification Performance

In this section we consider the use of AUC as a one-number summary of classification performance — i.e., when our prediction goal is simply to correctly classify the data, and our model selection task is to find the model with the *minimal future error rate*. We illustrate below that in many situations — both on simulation data and real data — the AUC score succeeds more in identifying the better model for future minimum error rate classification than the empirical Misclassification Rate (MC) on the test set. We conclude that while we cannot formalize a set of rules indicating when AUC will be a better discrimination method than MC, we can certainly recommend the complementary use of AUC as an evaluation criterion.

Assume as before that we have two scoring models m_1, m_2 and a test set of size n. Recall our representation of the difference between the two AUC scores above (1). We can derive a similar expression for MC when using a classification threshold t:

$$MC(m_2) - MC(m_1) = \sum_{i=1}^{n}[I\{m_1(x_i) \leq t\} - I\{m_2(x_i) \leq t\}]y_i \tag{2}$$

Since we are looking to *minimize* MC (as opposed to maximize AUC), a positive difference will lead us to select m_1. This is an "unbiased" representation of our ultimate goal, which is to select the better classification model. That is, we would ideally like to select m_1 if its *future* performance is better than that of m_2, i.e., if $E[MC(m_2) - MC(m_1)] > 0$.

We first present experiments which illustrate that AUC often does a better job than MC in selecting the better classification model among two candidates, then discuss the underlying reasons.

3.1. Simulations and real data experiments

We experiment with two of the simplest and most popular algorithms for creating classification models: Naive Bayes and K-Nearest Neighbors (K-NN). We create multiple pairs of models and demonstrate that AUC consistently identifies the better model with a higher probability across different pairs of models and large numbers of randomly generated test sets. For simplicity all the examples in this section were constructed assuming equal prior probabilities for the two classes and equal misclassification costs. As both algorithms are actually probability approximation algorithms the threshold for classification is always 0.5. [1]

It should be noted that the issue of ties is not addressed, so cases with equal scores were randomly ordered and given different ranks. This does not impede the validity of the results, however, as it is quite easy to show that it can only harm the average performance of AUC as a discrimination technique.

For the Naive Bayes simulations a 2-dimensional input vector is used. All x-values are independently drawn in [0,1]. The probability of y to be 1 is set to:

$$\begin{aligned} &Pr(y = 1|\mathbf{x}) = \\ &= \frac{1}{2}(x_1 + x_2)^2 \cdot I\{x_1 + x_2 \leq 1\} + \\ &+[1 - \frac{1}{2}(2 - x_1 - x_2)^2] \cdot I\{x1 + x2 > 1\} \end{aligned}$$

Twenty training sets of size 1000 each were drawn. For each training set, 100 test sets of size 100 each were drawn. As Naive Bayes is based on using discrete x-

[1] Some recent papers, (e.g. (Lachiche & Flasch, 2003)) suggest that 0.5 may not be the optimal threshold for Naive Bayes models, even in balanced class situations. However, our interest is purely in comparing test set and population performance. So, even if our threshold is indeed sub-optimal for classification, it has no bearing on the validity of our comparisons, as long as the same threshold is used for the testing and for population-level evaluation. The classification threshold in this view is a part of the Naive Bayes model specification, not a separate parameter to be optimized. The fact that AUC is oblivious to that threshold is an illustration of the "bias" in using AUC for model selection, as discussed in section 3.2.

Table 3. Comparison of AUC and MC performance on Naive Bayes models.

	MC	AUC
AVG. % m_2 BETTER	90.05	98.60
MIN. % m_2 BETTER	69	95
# TIMES m_2 BETTER ON ALL 100	1	8

Table 4. Comparison of AUC and MC performance on Nearest Neighbor models.

	MC	AUC
AVG. % m_2 BETTER	69.7	92.6
MIN. % m_2 BETTER	61	86
# TIMES m_2 BETTER ON AT LEAST 80	0	20

values, the training set x-values were discretized using a standard χ^2 method . It is easy to see that the real Bayes classifier (which gives $y = 1$ iff $x_1 + x_2 \leq 1$) cannot be expressed as a Naive Bayes classifier. The best Naive Bayes classifier should clearly be the one using both x_1, x_2 for the model. Classifiers with only x_1 or x_2 will generally be "under-fitted" (lack vital model flexibility), while classifiers with additional superfluous predictors will generally be "over-fitted" (contain extra flexibility which is used to model noise). The simulation compared the Naive Bayes classifier using only x_1 to the classifier using both x_1 and x_2.

Thus, two models were created from each of the twenty training sets and were used to score the 100 test sets. Each of the models was then evaluated on each of the test sets using both MC and AUC. The result which is of interest is the percentage of test sets on which the better model (m_2) gives a better overall evaluation score than the worse model (m_1). As expected, each of the twenty different pairs of models indicated more than half the time that m_2 was better than m_1, and m_2 was much better than m_1 in terms of future classification performance in all twenty cases. The overall average MC rates of the two models combined over all experiments: m_1 : 33.5%, m_2 : 27.5%.

The results are summarized in table 3. The first row shows the average percentage of test sets on which m_2 was better than m_1 (averaged over the twenty different pairs of models, each pair generated using a different training set). The second row shows the minimal number of times out of 100 that m_2 was better (taken over the twenty experiments) and the third row shows the number of times m_2 got a perfect record over m_1 by being better on all 100 test sets. These results indicate clearly that AUC succeeded much better than MC in tracking down the better model — AUC was wrong on average just 1.4% of the time while MC was wrong on average 9.95% of the time. AUC also did better in the "worst case" sense — its worst result (95%) was much better than MC's worst result (69%).

We performed a similar simulation experiment for K-NN. The dimension used here was $d = 10$ and the true model used here was even simpler: $Pr(y = 1|x) = x_1$. m_1 used K=10 nearest neighbors, while m_2 used $K = 50$. As before, we drew twenty training sets with 1000 observations in each and used 100 test sets of size 100 for each. m_2 was the better model, incurring an aggregated error rate of 26.5% over all test sets compared to 28.5% for m_1. Table 4 shows the results of our experiments in a similar format to table 3. The results here are even more conclusive in favor of AUC than in table 3 — we even have a "threshold" of 80% correct discrimination between the models which AUC achieved on all twenty pairs of models but MC achieved on none.

We also performed some simulations comparing pairs of models where one model was a Naive Bayes model and the other was a K-NN model. We omit them for space considerations. Their overall results were similar in spirit to the results we have shown, i.e., that AUC did a better job of identifying the better model. However, they were less conclusive. In particular, it seemed that AUC tended to prefer K-NN models to Naive Bayes models *more than* MC. So, in situations where the Naive Bayes model is a slightly better classification model than the K-NN model, the AUC still tends to prefer the K-NN model most of the time. This illustrates the "bias" in using AUC to select classification models, which we discuss in section 3.2

Finally, we performed experiments on a real-life data set. We used the "Adult" data-set available from the UCI repository (Blake & Merz, 1998). We used only the first ten variables in this data-set, to make a large-scale experiment feasible, and compared performance of Naive Bayes models using different subsets of these ten predictors. We had a total of 17000 cases for our experiments, of which we used 1700 for "training" the various models. We tested the models on 100 small test samples randomly drawn from the rest of the data. Table 5 shows the setup for the different experiments and their outcome. In it we can see the models and their "population" error rate (for the 15300 observations not used for training).

The results confirm our general intuition about the discrimination performance of AUC — in cases 1,2,3

Table 5. Comparison of AUC and MC performance in selecting the better Naive Bayes classification model for the Adult dataset.

m_1 FEATURES	m_1 ACCURACY	m_2 FEATURES	m_2 ACCURACY	# MC SELECTS m_2	# AUC SELECTS m_2
3	0.771	5	0.785	66	95
3	0.771	8	0.804	73.5	100
5	0.784	7	0.815	82.5	100
6	0.816	10	0.797	26	75
8	0.803	10	0.797	33	39.5

it is much more conclusive than MC in its choice of the better model. In case 5 the two methods perform approximately the same. In case 4 we see that AUC and MC select different models as the better model — here the "bias" of AUC controls the decision.

3.2. Discussion: Estimation-Approximation Tradeoff

In comparing AUC and MC, equations (1) and (2) show the similarity in the structure of the model comparisons . In each summand we have a first term which compares the decisions of the two models. (2) compares a thresholding decision while (1) compares an ordering decision. If the two decisions differ, they are confronted with the evidence from the test set — which determines whether 1 will be added to or subtracted from the sum. For example: for MC, if $m_1(x_i) \leq t$ and $m_2(x_i) > t$, the sum will increase by 1 if $y_i = 1$ (evidence in favor of m_2) and will decrease by 1 if $y_i = -1$ (evidence in favor of m_1). The MC score (2) compares these decisions on the n test-set cases and checks them against the test-set y values when the two models disagree. The AUC score (1) compares decisions on the $n(n-1)/2$ pairs of cases and checks the y-values when the decisions regarding the ordering of these cases by their scores disagree.

So in general we can say that Misclassification Rate sums over the results of $O(n)$ comparisons and AUC sums over the results of $O(n^2)$ comparisons (these comparisons are not independent, of course). This gives us a sense of the advantage of AUC over the MC — it is in fact the reduced variance of the resulting decision, obtained by using more information to make it. On the other hand, the fewer comparisons we make in (2) are the ones which actually determine the classification performance of the models. Thus, the mean result of (2) is guaranteed to be the right one, when our goal is to minimize the misclassification rate. The comparisons in (1) check slightly different attributes of the scoring model. A scoring model may have one behavior with regard to the "threshold crossings" tested in (2) and a different one with regard to the "binary switches" tested in (1). So, AUC may be "biased" in that its result may not reflect the true difference in classification accuracy. Consequently, models for which AUC should be effective as a comparison method would be ones where the scoring behavior is "consistent" in the sense that threshold crossings are as common as can be expected given the number and magnitude of binary order switches and vice versa. When this consistency does not exist we can expect that AUC will not always outperform MC in selecting the better classification model. Examples of both scenarios can be found above.

Intuitively it seems reasonable that most of the algorithms for creating scoring models, which fit the data to some non-trivial structures, should not display inconsistent behavior between threshold crossings and binary order switches, so the amount of "bias" they introduce in our representation should not be large and the decreased "variance" should control the process. The experimental results we presented above — for simulation studies and real data — illustrate that in most of the scenarios we have checked, the researcher would indeed do well to select models by their AUC score, even if the future goal is to obtain the best possible misclassification rate performance.

As one more concrete illustration of what we mean by reduced "variance" from using the AUC, we analytically calculated the moments of the MC difference for our example of section 2.1.3 (this is not a complicated calculation, given (2)) . The results, together with previously displayed moments for AUC, are presented in table 6. A concrete concept of "variance" is to compare the magnitude of the difference in means to the standard deviation. The columns labeled Φ give the corresponding normal tail probabilities for this ratio, i.e $1 - \Phi(\frac{mean}{std})$. Reduced variance indeed leads to higher probability of correct discrimination for the AUC on all these examples.

It is interesting to compare our conclusion with two recent related papers. From a theoretical perspective,

Table 6. Expansion of table 1 to include the moments and tail probability for misclassification rate.

N	q_1	q_2	AUC-Mean	AUC-Variance	AUC-Φ	MC-Mean	MC-Variance	MC-Φ
50	0.1	0.2	8.36	476	0.65	0.42	5.36	0.57
200	0.1	0.2	127.86	28577	0.78	1.63	20.93	0.64
400	0.1	0.2	507.46	226097	0.86	3.22	41.69	0.69
50	0.1	0.3	15.99	592	0.74	0.85	6.98	0.63
200	0.1	0.3	246.39	35602	0.90	3.25	27.25	0.73
400	0.1	0.3	979.19	281769	0.97	6.45	54.38	0.81

(Ling et al., 2003) show that AUC tends to select the same models as empirical error, but is less prone to ties. They conclude that AUC is "statistically consistent and more discriminating" than empirical error. The fundamental difference between their work and ours is that we are ultimately concerned with the prediction performance of our models, while they perform combinatorial analysis of test-set performance. On the other hand, (Perlich et al., 2003) compare AUC and MC in the context of a large scale empirical study. They conclude that in most cases the both criteria tend to select the same models, although they note some interesting exceptions (e.g. pruning decision trees improves MC but not AUC). Their experiments illustrate the "bias" effect we discussed above and verify that the two measures are usually — but not always — consistent. Note that they utilize large test sets and therefore the "variance" effects we consider are much less relevant.

4. Conclusion

Our aim in this paper is to improve understanding and usefulness of AUC as a tool for model selection and discrimination. We have introduced a theoretical method for calculating the moments of AUC differences, as a means to better understand the underlying processes and to investigate the correctness of practical methods. We have also illustrated the usefulness of AUC as a stable classification evaluation measure. This implies that in high-uncertainty situations, such as having a small amount of independent data for model selection, the AUC may be the better performance measure for discriminating between models than empirical error, even when the ultimate goal is to classify well.

Some interesting questions remain unanswered. In particular, we would like to be more specific about the "bias-variance" tradeoff in using the AUC to evaluate classification performance. Can we define more clearly and rigorously situations where the AUC is preferable? Where the misclassification rate is preferable?

References

Blake, C., & Merz, C. (1998). Repository of machine learning databases. `www.ics.uci.edu/~mlearn/MLRepository.html` UC Irvine, Dept. of ICS.

Cortes, C., & Mohri, M. (2003). AUC optimization vs. error rate minimization. *NIPS-03*.

DeLong, E. R., DeLong, D., & Clarke-Pearson, D. L. (1988). Comparing the area under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, 837:845.

Egan, J. (1975). *Signal detection theory and roc analysis*. Academic Press.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 29:36.

Hanley, J. A., & McNeil, B. J. (1983). A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, 839:843.

Lachiche, N., & Flasch, P. (2003). Improving accuracy and cost of two-class and multi-class probabilisitc classifiers using ROC curves. *ICML-03*.

Lehmann, E. L. (1975). *Nonparametrics : Statistical methods based on ranks*. Holden Day.

Ling, C., Huang, J., & Zhang, H. (2003). AUC: a statistically consistent and more discriminating measure than accuracy. *IJCAI-03*.

Perlich, C., Provost, F., & Simonoff, J. S. (2003). Tree induction vs. logistic regression: A learning-curve analysis. *JMLR*, *4*, 211:255.

Provost, F., & Fawcett, T. (1997). Analysis and visualization of classifier performance: Comparison under imprecise class and cost distribution. *KDD-97*.

Sen, P. (1960). On some convergence properties of U-statistics. *Calcutta Stat. Assoc. Bulletin*.