Bootstrap Inference on the Area Under the ROC Curve

Lealand Morin[1]
Department of Economics
College of Business Administration
University of Central Florida
Lealand.Morin@business.ucf.edu
https://business.ucf.edu/person/lealand-morin/

January 16, 2019

Abstract

There are several approaches to quantify the variability of the area under the receiver operating characteristic curve (AUROC) to measure the predictive performance of a binary classification model. These methods typically involve calculating the variance of the estimate and calculating a z-statistic, under specified distributional assumptions to characterize sampling variability.

Bootsrap methods are available but they are limited to the pairs bootstrap, which does not impose the null hypothesis.

In contrast, this paper proposes an approach to impose the null hypothesis of a particular AUROC, by shifting the sampling distribution toward a distribution that has the specified AUROC under the null hypothesis.

Under this approach, distance is measured as the Kullback-Leibler divergence between the full sample and other potential samples. The extreme values of the AUROC statistics from distributions with the specified value of divergence provide an interval for forecasting the AUROC to be expected in future applications of the model.

In essence, the procedure involves re-weighting the observations to those of the closest data generating process that satisfies the null hypothesis of a particular AUROC value.

The key element is a recursive algorithm for reallocation of sampling weights in the empirical distribution. The algorithm solves for a set of fixed points that characterize the optimal distribution, from which the desired AUROC values are calculated.

[1]I am grateful for helpful feedback from seminar participants at the University of Central Florida and at the meeting of the 2018 Canadian Econometric Study Group in Ottawa. All errors are my own.

Simulation exercises show the size and power are better than existing alternatives.

KEY WORDS: Bootstrap, ROC curve, AUC, AUROC, classification models, predictive modeling, Mann-Whitney statistic, Wilcoxon statistic, Kullback-Leibler divergence.

1 Introduction

The `pROC` has a bootstrap technique built into the package following the recommendations contained in (Carpenter). The paper by Carpenter is aimed at a health audience and not specifically derived for the AUROC. It boils down to a version of the pairs bootstrap. This DGP does not satisfy the null hypothesis as the population AUROC statistic is the AUROC realized in the sample.

Essentially, this amounts to adding noise to the standard error approach set out in DeLong and Clarke-Pearson (1988), and is made computationally efficient in Sun and Xu (2014). Under this approach, the variance of the AUROC is calculated as functions of the parameters of the exponential distributions to calculate a z-statistic. Hanley and McNeil (1982) presents the formulation of the variance of this statistic. Under this approach, the statistics are calculated from the empirical distributions.

However, this approach does not alow for asymmetry in the sampling distribution of the AUROC statistic. If one wants to use a sampling distribution close to the actual distribution, asymmetry and all, the bootstrap approach in `pROC` is the only currently available option. It is a reasonable approach but the approach presented in this paper is an improvement because it imposes a realistic sampling distribution AND also imposes the null hypothesized value of the AUROC statistic.

How do I do this? By finding the nearest distribution to the EDF from the sample such that the null hypothesis is satisfied.

2 Bootstrap

- Set up problem. Make sure to specify the distance metric so that the true distribution is under the null hypothesis.
- Proposition 1. The choice set is nonempty. There exists a distribution that satisfies the null hypothesis. That is, there is some distribution that is a re-weighting of the distribution that can generate any value of the AUROC, as long as the support of the distributions overlap.
- Secondly, the distribution satisfies the FOC's. This problem may or may not have a unique solution. There could be multiple distributions with the same conditions. Think of a pathological case with a few, maybe two, observations.
- This problem is well defined, since I restrict the distribution to a finite set of at most n points, restricted to the unique realizations in the sample of size n. Tightness is not

an issue, since the samples follow from the DGP and each of the bootstrap DGPs will have the same support as the individual samples.

- Convergence properties: As the sample size increases, the sequence of bootstrap distributions will converge to equal weighting. This is because the empirical CDF will converge to the populations CDF. The sequence of distributions will satisfy the null hypothesis in that the calculated AUROC is exactly the hypothesized AUROC value under the bootstrap DGP.
- A result of this is that the pairs bootstrap does not satisfy the null hypothesis.
- The sequence of distributions will converge to something. The something that this converges to will satisfy the null hypothesis as well. In particular, if the null is true, the distribution converges to the population CDF.
- Under alternatives, it would be nice if it converged to something that was well defined. Although this doesn't matter so much, since divergence is another way to get power.
- The algorithm for the solution is a contraction mapping. Therefore the repeated application of the FOC iteration will push the distribution toward the optimal weighting for the bootstrap sample.
- Derive the result for the discrete case for implementation with the sample.

3 Example: Power Law Distribution

Consider the case of power law distribution. Suppose that y is distributed on support $[y_{min}, \infty)$ with power parameter γ. Similarly, suppose that x is distributed on support $[x_{min}, \infty)$ with power parameter α.

Under this scenario, the AUROC will be as follows.

Case 1: $y_{min} \geq x_{min}$

$$1 - \frac{\gamma - 1}{(\gamma - 1) + (\alpha - 1)} \left(\frac{x_{min}}{y_{min}} \right)^{\alpha - 1} \tag{1}$$

Case 2: $y_{min} \leq x_{min}$

$$\frac{\alpha - 1}{(\gamma - 1) + (\alpha - 1)} \left(\frac{y_{min}}{x_{min}} \right)^{\gamma - 1} \tag{2}$$

Note that the AUROC is well-defined as long as $\gamma > 1$ and $\alpha > 1$. Under this case, the probability mass defining AUROC does not diverge. No moments of the power law distribution are necessary for the distribution of the AUROC to be well-defined. However, the sampling distribution of the AUROC is asymmetric.

This implies that testing the null hypothesis is more complicated that simply shifting the distribution or comparing the number of standard deviations of the calculated statistic from the null value.

References

Bamber, D. (1947): "The area above the ordinal dominance graph and the area below the receiver operating characteristic graph," *Journal of Mathematical Psychology*, 12(12), 387–415.

DeLong, E., D. D., and D. Clarke-Pearson (1988): "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach," *Biometrics*, pp. 837–845.

Demidenko, E. (2012): "Confidence intervals and bands for the binormal ROC curve revisited," *Journal of Applied Statistics*, 39, 67–79.

Hanley, J. (1989): "Receiver operating characheristic (ROC) methodology: the state of the art," *Critical Reviews in Diagnostic Imaging*, 29(3), 307.

Hanley, J., and B. McNeil (1982): "The Meaning and Use of the Area under a Receiver Operating (ROC) Curve," *Radiology*, 143, 29–36.

Kendall, M., and J. Gibbons (1990): *Rank Correlation Methods*. Arnold, London.

Lehmann, E., and H. D'Abrera (2006): *Nonparametrics: statistical methods based on ranks*. Springer, New York, NY.

Mann, H., and D. Whitney (1947): "On a test of whether one of two random variables is stochastically larger than the other," *The Annals of Mathematical Statistics*, 18(1), 50–60.

Newson, R. (2006): "Confidence intervals for rank statistics: Somers' D and extensions," *The Stata Journal*, 6(3), 309–334.

Peterson, W, B. T., and W. Fox (1954): "The theory of signal detectability," *IRE Professional Group on Information Theory*, 4(4), 171–212.

Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, and M. Müller (2011): "pROC: an open-source package for R and S+ to analyze and compare ROC curves," *BMC Bioinformatics*, 12, 77.

Somers, R. H. (1962): "A new asymmetric measure of association for ordinal variables," *American Sociological Review*, 27, 799–811.

Sun, X., and W. Xu (2014): "Fast Implementation of Delong's Algorithm for Comparing the Areas under Correlated Receiver Operating Characteristic Curves," *IEEE Signal Processing Letters*.

Van Meter, D., and D. Middleton (1954): "Modern statistical approaches to reception in communication theory," *IRE Professional Group on Information Theory*, 4(4), 119–145.

Wald, A. (1949): "Statistical Decision Functions," *The Annals of Mathematical Statistics*, pp. 165–205.

Wilcoxon, F. (1945): "Individual comparisons by ranking methods," *Biometrics Bulletin*, 1(6), 80–83.

Appendix A Area Under the ROC Curve

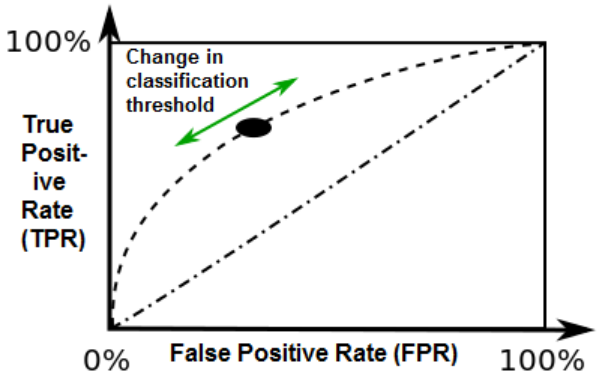
The ROC curve is a parametric curve, plotting the true positive rate against the false positive rate. The curve is parameterized by the discrimination threshold, the value of the classification variable, above which observations are classified as positive. A particular point on this curve shows the performance of the classification, plotting the observations correctly classified as positive, against those that are classified as positive but are truly negative.

The result can be used as a measurement of the frequency of correct interpretation of a signal for predicting a binary element of a message. In fact, the use of this curve for a measure of signal quality dates back to the measurement of the quality of radio transmissions for military use. Early academic references include Peterson and Fox (1954), Van Meter and Middleton (1954) and Wald (1949) for uses in this context. A later review can be found in Hanley (1989), in which the statistic is used to assess the quality of medical diagnoses. Recently, this curve has been used by the business community to assess the discriminatory power of variables used in classification models for predictive modeling.

Figure 1 shows an example of the Receiver Operating Characteristic curve. It plots the True Positive Rate (TPR) on the vertical axis against the False Positive Rate (FPR) on the horizontal axis. These proportions are calculated for a particular value of the classification threshold, which the user can vary to make a classification more selective or more inclusive. The observations with classification variable above the classification threshold are classified as positive and the rates are then calculated according to whether that classification is correct in each case. In the bottom left corner, only the strongest signals are used to identify positive outcomes and, for an effective model, the ratio of correct to incorrect positive identifications should be greater than equal, implying a steeper slope. As the threshold is increased, progressively weaker signals are used to identify positive outcomes and the performance should decline, resulting in a declining slope. At the other extreme, the classification threshold will include nearly the entire sample, for which the accuracy will have declined enough to coincide with the sample-wide event rates.

The better the model, the steeper will be the initial slope and the closer the curve will be to the upper boundaries. The area under this curve is the statistic that indicates the performance value of the classification variable. The diagonal line indicates the classifications that are no better than random allocation to positives and negatives. In this locus, the event rate is simply the average event rate for the entire sample. A model with a curve below this

Figure 1: Receiver Operating Characteristic Curve



Receiver Operating Characteristic Curve: True Positive Rate vs. False Positive Rate, for a particular value of the classification threshold. Observations with classification variables above the classification threshold are classified as positive. The true positive rate is the proportion of positive values of the classification that are correctly classified as negative, i.e. that lie above the threshold. Conversely, the true negative rate is the proportion of negative values of the classification that are incorrectly classified as positive, also above the threshold. The ROC curve is the parametric curve that is followed as the classification threshold varies from the highest to the lowest values in the sample.

line is not effective in classifying, unless used in reverse order, when the curve will then lie within the upper triangular region.

The area under this curve can be computed directly by integration:

$$A = \int_{-\infty}^{\infty} TPR(t)[-FPR'(t)]dt, \tag{3}$$

in which the true and false positive rates are defined as

$$TPR(t) = \int_{t}^{\infty} g_y(t)dt, \quad FPR(t) = \int_{t}^{\infty} f_x(t)dt. \tag{4}$$

The observations of the classification variable corresponding to positive outcomes is denoted by y and that for negative outcomes is denoted by x. The functions $g_y(\cdot)$ and $f_x(\cdot)$ are the corresponding density functions. A brief series of calculations will reveal this to be equivalent to the following probability

$$A = Prob\{y > x\}. \tag{5}$$

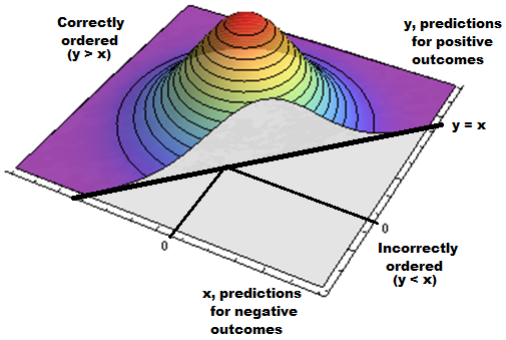
This can be interpreted as an evaluation of a pairwise comparison of the correct ordering of predictions for all pairs of predictions that could be made in the sample. Specifically, if one were to pick a pair of predictions, drawn randomly from predictions corresponding to pairs of the positive (y) and the negative (x) outcomes, the AUROC is the probability that these predictions are correctly ordered. The sample analogue is calculated as follows

$$\hat{A} = \hat{\Pr}\{y > x\} = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} I_{\{y_j > x_i\}}. \tag{6}$$

This calculation is provided in Hanley and McNeil (1982), while an implementation of the calculation is made available in Robin, Turck, Hainard, Tiberti, Lisacek, Sanchez, and Müller (2011). This is an implementation of the approach taken in DeLong and Clarke-Pearson (1988). A more efficient method of computation is proposed by Sun and Xu (2014), which is a faster approach to calculating these quantities.

A more intuitive approach to the calculation of the AUROC is to visualize it as the probability mass that satisfies a certain condition. Consider the following joint distribution formed by the two groups of classification variables. Combine the marginal distributions of the classification variables corresponding to the positive and negative outcomes by simply multiplying the marginal distributions. The resulting joint density will represent the probability density of all possible (x, y) pairs from the sample. Now restrict this space to the region defined by the indicator in equation (6), which is the region above the $y = x$ line. The area under the receiver operating curve is the volume under this "receiver operating surface" shown in Figure 2.

Figure 2: Volume Under the "ROC Surface"



Volume Under the "ROC Surface": With the AUROC defined in the form $A = Prob\{y > x\}$, it is clear that it can be visualized as the volume of the probability mass under the surface of the joint distribution of classification variables from positive and negative outcomes.

A.1 Examples

Figure 2 also illustrates the calculation of the AUROC for the specific case of the bi-normal model. The calculation is known to be equivalent to the normal CDF evaluated at the value of a test statistic for testing the difference between the two normal distributions. Similarly, the variability of the estimate results in confidence bounds that are also calculated by the normal CDF, which is specified in Demidenko (2012). However, it is noteworthy that the AUROC statistic is a function of the test statistic for a test of the difference between the pair of distributions of classification variables.

Another possibility is the bi-exponential model, in which both sets of classification variables are drawn from exponential distributions. This alternative is featured in Hanley and McNeil (1982) for a calculation of the variance of the AUROC. This approach allows for a

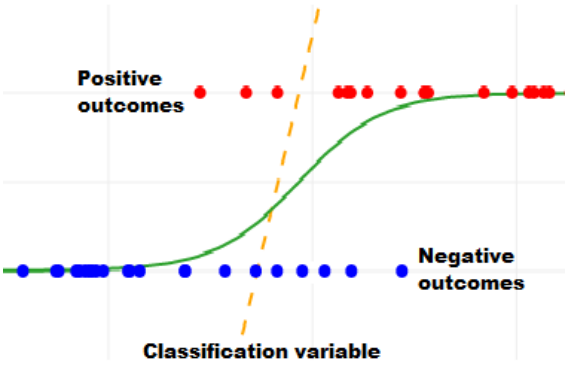


Figure 3: Predictive value of classification variables

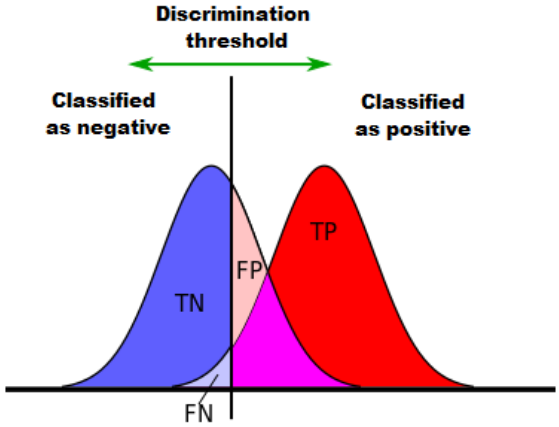


Figure 4: Difference in the distributions of variables

more generic formulation that is calculated directly from the empirical distributions of the classification variables from both the positive and negative outcomes.

Features of these two examples can be extended to the study of generic models for classification. However, parametric solutions are not preferred in the literature, since the classification model is likely to be much more unstructured than a parametric model would allow. A nonparametric solution is not only possible but preferred by users in industry. Regardless of the particular model employed in practice, the parametric examples can still be used to explain an important correspondence between modeling methods. The mapping presented next is generalized to all classification models.

A.2 Equivalence to Ranking Statistics

There is an important similarity between the act of creating a classification model and that of comparing two distributions. A difference between the distributions is a necessary condition for a classification variable to have predictive power for the binary outcome. Conversely, if the classification variable produces predictions with an AUROC above 0.5, then the distributions must be different. For this reason, an important connection can be made between nonparametric statistics and the AUROC, and this connection is useful for finding ways to further characterize the AUROC.

Because of this close connection between statistical methodologies, a number of methods in the nonparametric literature are useful for analyzing the AUROC. Nonparametric ranking statistics are described in Lehmann and D'Abrera (2006) and Kendall and Gibbons (1990), in reference to the literature on the comparison of distributions. One method that stands out as especially relevant is the Mann-Whitney U-statistic, which is presented in Mann and Whitney (1947), for comparing two distributions nonparametrically,

$$U = \sum_{i=1}^{m} \sum_{j=1}^{n} I_{\{y_j > x_i\}}. \tag{7}$$

This is the same as AUROC, without the normalization by the count of pairs mn.

This equivalence is leveraged for computational advantage, since the U-statistic is calculated as follows. First, sort the observations in increasing order. Then, assign ranks to the observations, with 1 for the smallest score and the sample size for the observation with the largest score. Finally, take the sum of the ranks corresponding to the positive outcomes and subtract half of the quantity $n_1(n_1+1)$, where n_1 is the number of positive observations in the sample. A similar statistic was proposed in Wilcoxon (1945) for similar purposes. Bamber (1947) made the connection between the AUROC and the Mann-Whitney U-statistic. Other approaches for evaluating classification models are similarly inspired by tools such as Somers' d (Somers (1962)) and Kendall's τ (see Kendall and Gibbons (1990)) A simulation-based approach to modeling the variability of these statistics is found in Newson (2006), which involved jackknifing the numerator and denominator of the ratio in Kendall's τ and taking a Taylor expansion.

Appendix B Distance to Distribution with AUROC A_0

The objective is to minimize distance $D(\mathbf{u}\otimes\mathbf{v}, \mathbf{f}\otimes\mathbf{g})$, which could be either of the functions $CHI(h_1,h_2)$ or $KLD(h_1,h_2)$ or any other measure of distance between distributions, such that the chosen distribution corresponds to a specified AUROC statistic. The optimization problem is formalized as follows.

$$\min_{\mathbf{u},\mathbf{v}} D(\mathbf{u}\otimes\mathbf{f}, \mathbf{v}\otimes\mathbf{g}) \tag{8}$$

subject to

$$\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} u_i v_j I_{\{y_j > x_i\}} = A_0, \tag{9}$$

and the conditions on the probabilities

$$\sum_{i=1}^{m} u_i = 1, \quad \sum_{j=1}^{n} v_j = 1, \quad \{u_i \geq 0\}_{i=1}^{m}, \quad \{v_j \geq 0\}_{j=1}^{n}. \tag{10}$$

Stated as a Lagrange optimization problem, this becomes

$$\begin{aligned}
&\min_{\mathbf{u},\mathbf{v}} Q(\mathbf{u},\mathbf{v},\mathbf{f},\mathbf{g},\gamma_x,\gamma_y,\delta_x,\delta_y) && (11)\\
&\quad = \min_{\mathbf{u},\mathbf{v}} D(\mathbf{u}\otimes\mathbf{f}, \mathbf{v}\otimes\mathbf{g}) && (12)\\
&\quad - \lambda\left[\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} u_i v_j I_{\{y_j > x_i\}} - A_0\right] && (13)\\
&\quad - \gamma_x\left[\sum_{i=1}^{m} u_i - 1\right] - \gamma_y\left[\sum_{j=1}^{n} v_j - 1\right] - \sum_{i=1}^{m}\delta_{x,i}u_i - \sum_{j=1}^{n}\delta_{y,j}v_j && (14)
\end{aligned}$$

The first order conditions for this problem are

$$\frac{dD(\mathbf{u},\mathbf{f})}{du_i} = \lambda \sum_{j=1}^{n} v_j I_{\{y_j > x_i\}} + \gamma_x + \delta_{x,i}, i = 1, \ldots, m, \tag{15}$$

and

$$\frac{dD(\mathbf{v},\mathbf{g})}{dv_j} = \lambda \sum_{i=1}^{m} u_i I_{\{y_j > x_i\}} + \gamma_y + \delta_{y,i}, j = 1, \ldots, n. \tag{16}$$

For $D(f_1, f_2) = KLD_1(f_1, f_2)$ the distance function has terms of the form

$$D(\mathbf{u}_i, \mathbf{f}_i) = u_i \ln \left\{ \frac{u_i}{f_i} \right\} \tag{17}$$

so the derivative term is

$$\frac{dD(\mathbf{u},\mathbf{f})}{du_i} = \ln \left\{ \frac{u_i}{f_i} \right\} - u_i \ln f_i - 1. \tag{18}$$

Note that the nonnegativity constraints are not binding in the case of the Kullback-Leibler metric, since the probability weights in $\mathbf{u}$ and $\mathbf{v}$ appear within the natural log function. Setting these to zero would result in an infinite distance, meaning that they provide absolute information for discriminating between distributions. As a result, the Lagrange multipliers $\{\delta_{x,i}\}_{i=1}^{m}$ and $\{\delta_{y,j}\}_{j=1}^{n}$ are all equal to zero.

The first order conditions can be solved to isolate u_i and v_j so that they satisfy the following fixed points.

$$u_i = \exp \left\{ 1 + \ln f_i + u_i \ln f_i + \lambda \sum_{j=1}^{n} v_j I_{\{y_j > x_i\}} + \gamma_x \right\}, \tag{19}$$

and

$$v_j = \exp \left\{ 1 + \ln g_j + v_j \ln g_j + \lambda \sum_{i=1}^{m} u_i I_{\{y_j > x_i\}} + \gamma_y \right\}. \tag{20}$$

Note that this system of equations has dimension equal to the respective sample sizes, which could make the solution prohibitively costly to compute. To this end, one can iterate on the fixed point implied by the above:

$$u_i^{(t+1)} = k_x f_i^{1+u_i^{(t)}} \exp \left\{ \lambda \sum_{j=1}^{n} v_j^{(t)} I_{\{y_j > x_i\}} \right\}, \tag{21}$$

and

$$v_j^{(t+1)} = k_y g_j^{1+v_j^{(t)}} \exp \left\{ \lambda \sum_{i=1}^{m} u_i^{(t)} I_{\{y_j > x_i\}} \right\}, \tag{22}$$

where the constants k_x and k_y serve to normalize to unit probability mass.

Finally, the λ could, in principle, be solved for numerically, on each iteration. However, it is sufficient to choose λ so that it is small enough not to overstep the root of the fixed point. In addition, λ determines the direction of the iterations, toward either higher or lower A_0. To this end, it is sufficient to set the step size as

$$\lambda = \eta(\hat{A}^{(t)} - A_0), \tag{23}$$

for a small η, so that the step size is declining at each iteration, as the AUROC $\hat{A}^{(t)}$ moves in the direction of the specified value of A_0. The iterations continue until the first order conditions are satisfied and the distance $|\hat{A}^{(t)} - A_0|$ is near zero, such that both conditions are met, up to a chosen tolerance.