How Random is a Classifier given its Area under Curve?

Chris Zeinstra[1], Raymond Veldhuis[1], Luuk Spreeuwers[1]

Abstract: When the performance of a classifier is empirically evaluated, the Area Under Curve (AUC) is commonly used as a one dimensional performance measure. In general, the focus is on good performance (AUC towards 1). In this paper, we study the other side of the performance spectrum (AUC towards 0.50) as we are interested to which extend a classifier is random given its AUC. We present the *exact* probability distribution of the AUC of a truely random classifier, given a finite number of *distinct* genuine and imposter scores. It quantifies the "randomness" of the measured AUC. The distribution involves the restricted partition function, a well studied function in number theory. Although other work exists that considers confidence bounds on the AUC, the novelty is that we do not assume any underlying parametric or non-parametric model or specify an error rate. Also, in cases in which a limited number of scores is available, for example in forensic case work, the exact distribution can deviate from these models. For completeness, we also present an approximation using a normal distribution and confidence bounds on the AUC.

Keywords: Random Classifier, AUC, Exact Distribution, Approximation.

1 Introduction

The trade off between the False Match Rate (FMR) and True Match Rate (TMR) of a classifier while varying the decision threshold is commonly reported in a receiver operating characteristic (ROC) curve [Fa06]. There exist several one dimensional classifier performance measures that can be derived from its ROC curve, for example, the Equal Error Rate and the Area under Curve [HM82]. In this study, we consider the Area Under Curve (AUC) measure. An ideal classifier has AUC=1, whereas a random classifier has AUC=0.50. The AUC is equal to the probability that a randomly chosen genuine score is larger than a randomly chosen imposter score [HT01]. Also, the AUC can be interpreted as the Wilcoxon-Mann-Whitney statistic [MW47] when ordering the genuine and imposter scores produced by the classifier [HM82], [MG02].

In any empirical performance evaluation, only a finite number of genuine and imposter scores is available. Under the assumption that genuine and imposter scores are drawn from unknown probability densities, ultimately the AUC is also a random variable, having a probability distribution on its own. If we could replicate the experiment having the exact same number of genuine and imposter scores, we most likely would have obtained a different ROC curve and AUC. In particular, this implies that the performance evaluation might yield an AUC value that is not identified as being produced by a random classifier. This could occur in the case of a subject anchored approach to evidence evaluation in which the available number of scores is limited, see [Me06] for a general framework.

[1] University of Twente, Faculty of EEMCS, SCS Group, P.O.Box 217, 7500 AE Enschede, The Netherlands, {c.g.zeinstra,r.n.j.veldhuis,l.j.spreeuwers}@utwente.nl

The probability distribution of the AUC of a random classifier is easily derived for trivial cases. More precisely, we assume that (a) this classifier draws genuine and imposter scores randomly from the *same* probability distribution and (b) the drawn scores are *distinct*. The last condition is a necessary technicality; if we for example assume that scores come from a continuous interval, this condition is typically met. Suppose we construct a ROC curve based on 1 genuine score g and n imposter scores i_k, $k = 1, \ldots, n$. We have $n+1$ possible orderings of the scores:

$$g < i_1 < \ldots < i_{n-1} < i_n \text{ to } i_1 < i_2 < \ldots < i_n < g. \tag{1}$$

Since g and i_k come from the same distribution, each sequence in (1) has equal probability $\frac{1}{n+1}$. If l ($l = 1, \ldots, n+1$) is the position of g in any sequence in (1), then its AUC is equal to $\frac{l-1}{n}$. Hence, each possible AUC has equal probability. The one-to-one mapping in this trivial 1 genuine/n imposter case between sequences and the AUC does not hold in general. For example, both i_1, g_1, g_2, i_2 and g_1, i_1, i_2, g_2 yield AUC=0.50, and the situation becomes rapidly complex when m and n attain values found in practice.

The contribution of this paper is the exact probability distribution of the AUC of the random classifier for *any* finite number of genuine and imposter scores. Also, we present an approximation. This work can be used in the situation when we want to determine the probability that a random classifier produces the measured AUC; this is of interest when the measured AUC is low or the total number of scores is limited.

The remainder of this article is structured as follows. In Section 2, we present related work. Since the general approach involves the restricted partition function, we present its definition in Section 3. In Section 4, we present two theorems regarding respectively the probability distribution of the AUC and an approximation. Section 5 presents some examples of the exact and an application of the approximation. In Section 6, we discuss the two theorems. Finally, in Section 7 we present our conclusion.

2 Related Work

As indicated before, this work fits in a larger framework that studies whether two AUC's are significantly different by constructing confidence intervals. This is not only of importance in decision theory, but also for clinical medicine and psychology studies in which treatments are compared. We present some of these studies here.

For example, the work of [CM04] analytically derives exact and estimated confidence intervals based on a statistical and combinatorical analysis, using a fixed error rate and the number of genuine and imposter scores. Our work only uses the number of genuine and imposter scores, assuming that they are drawn from the same probability distribution. Another approach is the use of parametric models to construct confidence intervals. For example score distributions have been modeled as normal [HSZ09], binormal [MHS98], exponential [To77], and Gamma [PA95], from which expressions for the confidence intervals can be derived. Their main issue is the influence of the parametric assumption on the estimation of confidence intervals. To cater for that situation, several non-parametric methods have been explored, including Wilcoxon-Mann-Whitney and De-Long non-parametric interval [DDCP88]. The work of [QH08] compares nine non-parametric approaches in different simulation scenarios (moderate to good AUC and different combinations of genuine and imposter scores). They found that their own empirical likelihood approach [QZ06] has a good coverage in different scenarios. Several studies have shown that methods can be negatively influenced by the number of considered scores. For example, [OL98] found that asymptotic methods are less accurate in this situation; the study of [Ha10] shows how estimates for the AUC can differ significantly from the true value.

In summary, these studies emphasise on one hand the restriction of our work (random classifier) and on the other hand its uniqueness (exact distribution, depending on the number of genuine and imposter scores).

3 Partition functions

The partition function is an essential function in number theory, a branch of mathematics that studies properties of integers [An98]. A partition of a positive integer k is a decomposition of k as a sum of positive integers. The partition function p counts the number of different partitions of a positive integer, disregarding any permutations in the order of the terms. For example $p(5) = 7$, since

$$5 = 5 = 4+1 = 3+2 = 3+1+1 = 2+2+1 = 2+1+1+1 = 1+1+1+1+1. \quad (2)$$

It is customary to order the terms in a partition from the largest to the lowest value. This can be written more formally as $k_1 + \ldots + k_r = k,$ and $k_1 \geq k_2 \geq \cdots k_r$. Also, by convention, the domain of p is extended by including $p(0) = 1$ and $p(k) = 0$ for $k < 0$.

There exist different "restricted" versions of the partition function. In particular, one can limit the number and value of the terms of a partition. Let $p(n,m;k)$ be the number of partitions of k which have at most m terms, each having maximum value n. In the sequel, we refer to this function as "the" restricted partition function. For example, $p(4,2;5) = 2$, since the maximum value is 4 and the maximum number of terms is 2:

$$5 = 4+1 = 3+2. \quad (3)$$

The restricted partition function has a generating function:

$$\sum_{k=0}^{nm} p(n,m;k)q^k = \binom{m+n}{m}_q, \quad (4)$$

in which

$$\binom{m+n}{m}_q = \frac{\prod_{j=1}^{m+n}(1-q^j)}{\prod_{j=1}^{m}(1-q^j)\prod_{j=1}^{n}(1-q^j)} \quad (5)$$

is the Gaussian binomial coefficient [An74]. It generalises the binomial coefficient as for $\lim_{q \nearrow 1}$, (5) reverts to the standard binomial coefficient $\binom{k+l}{l}$. As an example, we expand

$p(4,2;k)$ for $k=0,\cdots,8$:

$$\sum_{k=0}^{8} p(n,m;k)q^k = \binom{6}{2}_q = \frac{\prod_{j=1}^{6}(1-q^j)}{\prod_{j=1}^{2}(1-q^j)\prod_{j=1}^{4}(1-q^j)} = \frac{(1-q^5)(1-q^6)}{(1-q)(1-q^2)}. \tag{6}$$

It is straightforward to verify that (6) is equal to $1+q+2q^2+2q^3+3q^4+2q^5+2q^6+q^7+q^8$. In particular, we observe that $p(4,2;5)=2$ (the factor of q^5), a result that was also demonstrated by (3).

4 Exact and Approximative Distribution

We have the following theorem on the distribution of AUC.

Theorem 1. Given m genuine and n imposter scores, all distinct, the possible values for AUC are

$$\{\frac{k}{mn}|k \in \{0,\ldots,mn\}\}. \tag{7}$$

Moreover, if the genuine and imposter scores are drawn from the *same* score distribution, then the probability distribution of the AUC is given by

$$\mathrm{p}\left(\mathrm{AUC}=\frac{k}{mn}\right) = \frac{p(n,m;k)}{\binom{n+m}{n}}, \tag{8}$$

where $p(n,m;k)$ is the restricted version of the partition function.

Proof. Having m genuine and n imposter scores, this divides the TMR (resp. FMR) space into $m+1$ (resp. $n+1$) points with distance $\frac{1}{m}$ (resp. $\frac{1}{n}$). Since we have distinct scores, whenever the threshold increases and passes a score, the corresponding operating point in ROC space will either move to the left with a step size $\frac{1}{n}$ or down with a step size $\frac{1}{m}$. Hence, the AUC can be seen as a sum of blocks of equal area of $\frac{1}{mn}$, showing that (7) holds.

Given the set of ROC curves for which the number of blocks under the curve is k, we can assign to each ROC curve a sequence $k_1,k_2,\ldots,k_r$ where k_1 is the number of blocks between $TMR=0$ and $TMR=\frac{1}{m}$, until k_r, being the number of blocks between $TMR=\frac{r-1}{m}$ and $TMR=\frac{r}{m}$. By construction, (a) $k_1+\ldots+k_r=k$, (b) the size of k_i is restricted to n, (c) r is limited to m, and (d) $k_1 \geq k_2 \geq \cdots k_r$.

The reverse relation also holds: given a sequence $k_1,k_2,\ldots,k_r$ with properties (a)-(d), we can construct the corresponding ROC curve uniquely as follows. Place k_1 blocks to the right between $TMR=0$ and $TMR=\frac{1}{m}$, until k_r blocks to the right between $TMR=\frac{r-1}{m}$ and $TMR=\frac{r}{m}$.

The properties (a)-(d) of a sequence $k_1,k_2,\ldots,k_r$ make it a restricted partition of k. Since there is a one-to-one correspondence between a ROC curve and a restricted partition, we conclude that the number of ROC curves with $AUC=\frac{k}{mn}$ is equal to $p(n,m;k)$.

Given that the total number of ROC curves is $\binom{n+m}{n}$, all being equiprobable due to the same score distribution assumption, we conclude that (8) holds. □

We can also approximate (8) with the normal distribution.

Theorem 2. Given m genuine and n imposter scores, the distribution of the AUC has an asymptotic normal distribution if $m \to \infty$ and $n \to \infty$, in particular

$$\lim_{\substack{m\to\infty \\ n\to\infty}} \mathrm{p}(\mathrm{AUC} \geq x) = 1 - \Phi\left(\frac{(x-\frac{1}{2})mn}{\sigma_{mn}}\right). \tag{9}$$

Here Φ is the cumulative standard normal distribution and

$$\sigma_{mn} = \sqrt{\frac{mn(m+n+1)}{12}}. \tag{10}$$

Proof. According to Theorem 4 of [Ta86], we have, using our notation

$$\lim_{\substack{m\to\infty \\ n\to\infty}} \mathrm{p}\left(\frac{k-\frac{1}{2}mn}{\sigma_{mn}} \leq t\right) = \Phi(t), \tag{11}$$

with k related to AUC as $\mathrm{AUC} = \frac{k}{mn}$. Using this relation in (11) we observe that

$$\lim_{\substack{m\to\infty \\ n\to\infty}} \mathrm{p}\left(\frac{(\mathrm{AUC}-\frac{1}{2})mn}{\sigma_{mn}} \leq t\right) = \Phi(t), \tag{12}$$

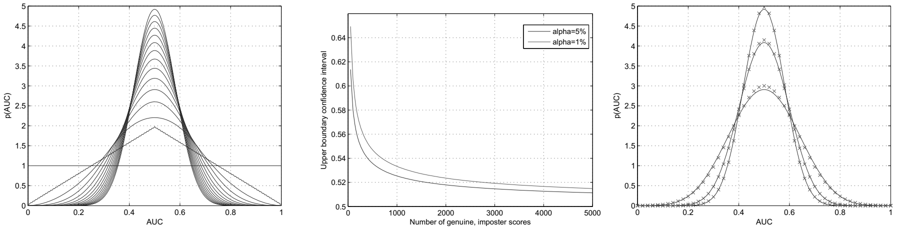
defining $x = \frac{1}{2} + \frac{t\sigma_{mn}}{mn}$ and reversing the inequality in (12) we conclude that (9) holds. □

5 Examples

In this section, we provide three examples of the exact distribution and one application that uses the approximation.

5.1 The 1 genuine/n imposter case

It is straightforward to show that $p(n,1;k) = \frac{(1-q)\cdots(1-q^{n+1})}{(1-q)\cdots(1-q^n)(1-q)} = \frac{1-q^{n+1}}{1-q} = \sum_{k=0}^{n} q^k$. Hence, $p(n,1;k) = 1$ for $k = 0,\ldots,n$. Moreover, $\mathrm{p}(\mathrm{AUC} = \frac{k}{mn}) = \frac{1}{\binom{n+1}{n}} = \frac{1}{n+1}$. This is in accordance with the example discussed in the Introduction.



a) b) c)

Fig. 1: a) p(AUC) for $m = 1, \cdots, 15$ genuine and $n = 100$ imposter scores. Graphs are scaled such that they can be interpreted as continuous probability distributions. b) The upper limit of 95% and 99% confidence intervals as a function of equal number of genuine and imposter scores. c) p(AUC) for $m = 5, 10, 15$ genuine and $n = 100$ imposter scores in blue, together with the approximation given by (9) in red.

5.2 The 2 genuine/Even n imposter case

Suppose n is even, then it can be shown that (5) can be written as

$$p(n,2;k) = (1+q+q^2+\cdots+q^n)(1+q^2+q^4+\cdots+q^n). \tag{13}$$

A straightforward calculation gives a staircase like shape:

$$\begin{array}{ll} p(2,n;2k) = p(2,n;2k+1) = k+1 & \text{if } 2k \leq n-1, \\ p(2,n,k) = \frac{n}{2}+1 & \text{if } k = n, \\ p(2,n;k) = p(2,n;2n-k) & \text{if } k \geq n+1. \end{array} \tag{14}$$

5.3 The 1-15 genuine/100 imposter case

In this example, we plot p(AUC) for $m = 1, \cdots, 15$ genuine and $n = 100$ imposter scores in Figure 1a. In particular, we see respectively the uniform and staircase like shapes appearing for $m = 1$ and $m = 2$.

5.4 Confidence bounds

Theorem 2 can be used to construct a two sided $1-\alpha$ confidence interval $[\frac{1}{2} - x_\alpha, \frac{1}{2} + x_\alpha]$ around the AUC of a random classifier that depends on the number of genuine and imposter scores. Rewriting (9) shows that x_α is given by $x_\alpha = z_\alpha \sqrt{\frac{m+n+1}{12mn}}$, with z_α defined implicitly as $\Phi(z_\alpha) = 1 - \frac{\alpha}{2}$.

In Figure 1b we have chosen $m = n$, respectively $\alpha = 5\%$ $(z_\alpha = 1.96)$ and $\alpha = 1\%$ $(z_\alpha = 2.33)$ and plotted the upper limit of confidence intervals as a function of the number of genuine and imposter scores. This illustrates the asymptotic behaviour of the approximation; for smaller numbers of scores, the AUC of a random system can still deviate much from AUC=0.50.

6 Discussion

Figure 1a visualises the dependency of $\mathrm{p}(AUC)$ on the number of scores. Especially, we observe that for a lower number of scores, the probability that a random system has an AUC that differs significantly from 0.50 is non trivial. This is of relevance in, for example, the case of a subject anchored approach to evidence evaluation.

Although Theorem 1 provides an exact result, it can be challenging to calculate the value of the restricted partition function. One needs to resort to data structures to accommodate for values that are larger than those can fit into an IEEE-754 64 bit integer representation. This may result in an increased calculation time due to the lack of an efficient mapping from primitive operators to single machine instructions. Moreover, if we would be interested in the cumulative probability $\mathrm{p}(\mathrm{AUC} \geq x)$, then a repeated calculation is not optimal as one could better use its generating function (4) for the simultaneous calculation of $p(n,m;k)$ over a range of values of k.

The result of Theorem 2 is an approximative result, and it is instructive to see how well it approximates the true probability distribution for finite values of m and n. We show the exact and the approximation for three cases: $m = 5,\ 10,\ 15$, and $n = 100$ in Figure 1c. Even for moderate values of m and n the approximation seems satisfactory. Furthermore, if the number of genuine and imposter scores are equal (k) and $k \to \infty$, the distribution becomes centered around AUC=0.50.

Although our work only considered approximative confidence bounds, we can also construct exact confidence bounds, especially when the number of scores is low.

7 Conclusion

In this paper, we have presented an exact formula for the probability distribution of the AUC of a random classifier, given a finite number of distinct genuine and imposter scores. This work can be used in the situation when we want to determine the probability that a random classifier produces the measured AUC; this is of interest when the measured AUC is low or the total number of scores is limited, masking the true nature of the classifier. The exact probability distribution involves the restricted partition function and can be approximated by a normal distribution. We used this approximation to derive confidence intervals for the AUC as a function of the number of genuine and imposter scores.

References

[An74] Andrews, George E.: Applications of basic hypergeometric functions. SIAM review, 16(4):441–484, 1974.

[An98] Andrews, G. E.: The Theory of Partitions. Cambridge Mathematical Library. Cambridge University Press, 1998.

[CM04] Cortes, Corinna; Mohri, Mehryar: Confidence intervals for the area under the ROC curve. In: Nips. pp. 305–312, 2004.

[DDCP88] DeLong, Elizabeth R.; DeLong, David M.; Clarke-Pearson, Daniel L.: Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics, pp. 837–845, 1988.

[Fa06] Fawcett, Tom: An introduction to ROC analysis. Pattern Recognition Letters, 27(8):861–874, 2006.

[Ha10] Hanczar, Blaise; Hua, Jianping; Sima, Chao; Weinstein, John; Bittner, Michael; Dougherty, Edward R.: Small-sample precision of ROC-related estimates. Bioinformatics, 26(6):822–830, 2010.

[HM82] Hanley, James A.; McNeil, Barbara J.: The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1):29–36, 1982.

[HSZ09] Hsieh, Hsin-Neng; Su, Hsiu-Yuan; Zhou, Xiao-Hua: Interval estimation for the difference in paired areas under the ROC curves in the absence of a gold standard test. Statistics in medicine, 28(25):3108–3123, 2009.

[HT01] Hand, David J.; Till, Robert J.: A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine Learning, 45(2):171–186, 2001.

[Me06] Meuwly, D.: Forensic Individualisation from Biometric Data. Science & Justice, 46(4):205–213, 2006.

[MG02] Mason, Simon J.; Graham, Nicholas E.: Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: Statistical significance and interpretation. Quarterly Journal of the Royal Meteorological Society, 128(584):2145–2166, 2002.

[MHS98] Metz, Charles E.; Herman, Benjamin A.; Shen, Jong-Her: Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data. Statistics in medicine, 17(9):1033–1053, 1998.

[MW47] Mann, Henry B.; Whitney, Donald R.: On a test of whether one of two random variables is stochastically larger than the other. The Annals of Mathematical Statistics, pp. 50–60, 1947.

[OL98] Obuchowski, Nancy A.; Lieber, Michael L.: Confidence intervals for the receiver operating characteristic area in studies with small samples. Academic Radiology, 5(8):561–571, 1998.

[PA95] Pham, T.; Almhana, J.: The generalized gamma distribution: its hazard rate and stress-strength model. IEEE Transactions on Reliability, 44(3):392–397, 1995.

[QH08] Qin, Gengsheng; Hotilovac, Lejla: Comparison of non-parametric confidence intervals for the area under the ROC curve of a continuous-scale diagnostic test. Statistical Methods in Medical Research, 17(2):207–221, 2008.

[QZ06] Qin, Gengsheng; Zhou, Xiao-Hua: Empirical likelihood inference for the area under the ROC curve. Biometrics, 62(2):613–622, 2006.

[Ta86] Takács, Lajos: Some asymptotic formulas for lattice paths. Journal of Statistical Planning and Inference, 14(1):123–142, 1986.

[To77] Tong, Howell: On The Estimation of Pr$\{Y < X\}$ for Exponential Families. IEEE Transactions on Reliability, 26(1):54–56, 1977.