8-2019

Why Area Under the Curve in Hypothesis Testing?

Griselda Acosta
University of Texas at El Paso, gvacosta@miners.utep.edu

Eric Smith
University of Texas at El Paso, esmith2@utep.edu

Vladik Kreinovich
University of Texas at El Paso, vladik@utep.edu

Comments:
Technical Report: UTEP-CS-19-88

Why Area Under the Curve in Hypothesis Testing?

Griselda Acosta[1], Eric Smith[2], and Vladik Kreinovich[3]
[1]Department of Electrical and Computer Engineering
[2]Department of Industrial, Manufacturing, and
Systems Engineering
[3]Department of Computer Science
University of Texas at El Paso
500 W. University
El Paso, TX 79968, USA
gvacosta@miners.utep.edu, esmith2@utep.edu, vladik@utep.edu

Abstract

To compare two different hypothesis testing techniques, researchers use the following heuristic idea: for each technique, they form a curve describing how the probabilities of type I and type II errors are related for this technique, and then compare areas under the resulting curves. In this paper, we provide a justification for this heuristic idea.

1 Formulation of the Problem

Type I and type II errors. There are many different techniques for hypothesis testing, i.e., for deciding, based on the observation, whether the original (*null*) hypothesis is valid or whether this hypothesis has to be rejected (and the alternative hypothesis has to be considered true); see, e.g., [3]. In hypothesis testing, we can have two different types of errors:

- a type I error (also known as False Negative) is when the correct null hypothesis is erroneously rejected, while
- a type II error (also known as False Positive) is when the false null hypothesis is erroneously accepted.

The probability of the type I error is usually denoted by α and the probability of the type II error is usually denoted by β.

In different situations, we have different requirements on the allowed probabilities of these two errors. For example, in early cancer diagnostics, when the null hypothesis means no cancer, type I error is not that critical – it simply

means that a healthy patient has to go through an extra testing to make sure that he/she is healthy. On the other hand, a type II error means missing a potentially dangerous disease – which can lead to grave consequences. In such situations, it is desirable to minimize the probability of type II errors as much as possible – even when this leads to a larger type I error.

On the other hand, in law enforcement, we do not want to have too high a probability of type I errors – that would mean SWAT teams breaking into the houses of innocent people in the middle of the night, that would mean massively arresting people who have not done anything wrong.

Depending on the situation, we can adjust the given technique – by changing some appropriate parameters – to increase or decrease α and β. In the ideal world, we should have both errors as small as possible, but this is not possible if all we have is a finite sample. Thus:

- if we decrease α, the probability β increases, and,
- vice versa, if we decrease β, the probability α decreases.

In particular, based on the finite sample, the only way to make sure that we do not have any type I errors is to never reject the null-hypothesis. In this case, however, every time the null hypothesis is false, it will still be accepted. In other words, when the probability α of the type I error is 0, then the probability β of the type II error will be 1.

Vice versa, the only way to get $\beta = 0$ is to never accept the null hypothesis – but in this case, we will have $\alpha = 1$.

How can we compare two hypothesis testing techniques? To get a full description of the quality of a given hypothesis testing technique, we need to indicate, for each $\alpha > 0$, what probability β we can achieve with this technique, and, vice versa, for each β, what probability α we can achieve with this technique. In other words, the perfect description of this quality is a curve that describes how β depends on α – and vice versa. We have a curve $\beta = f(\alpha)$ that describes the dependence of the smallest possible β on the given value α.

For $\alpha = 0$, as we have mentioned earlier, we have $\beta = f(0) = 1$. The larger α we allow, the smaller β can be – so the dependence $f(\alpha)$ is decreasing, and it reaches the value $f(1) = 0$ for $\alpha = 1$.

How do we compare the two hypothesis testing techniques? If the required value α is given, we select the technique for which the corresponding value β is the smallest – and, vice versa, if the value β is given, we select the technique for which the corresponding value α is the smallest.

This requires that we implement all possible hypothesis testing techniques, and every time select a technique depending on the specifics of a situation. In reality, however, there are dozens and dozens of different hypothesis testing techniques. In practice, it is often not realistically possible to implement all of them on the available computational device. In such situations, we select one of the techniques and use it – with varying parameters – for all possible practical situations. (Alternatively, we can select two or more techniques – and for each

given values of α or β, select the best of these. This is equivalent to selecting a *hybrid* technique: e.g., technique A for values of α which are smaller than some threshold α_0 and a technique B for all other values α.)

In all such cases, we select one of the techniques (either one of the original ones or one of the hybrid ones). The question is: how do we select? One technique may be better for some α, another may be better for another α. The usual way to select one of the available techniques is to select the one for which the Area Under the Curve (AUC) $\int\limits_0^1 f(\alpha)\,d\alpha$ is the smallest possible.

Comment. Instead of the dependence of β on α, we can plot the dependence of $1-\beta$ on α: $1-\beta = g(\alpha)$. For this new function, the area under its curve is equal to 1 minus the area under the f-curve:

$$\int_0^1 g(\alpha)\,d\alpha = \int_0^1 (1-f(\alpha))\,d\alpha = 1 - \int_0^1 f(\alpha)\,d\alpha.$$

Thus, for these functions, minimizing the area under the f-curve is equivalent to maximizing the area under the g-curve.

Why? In practice, the AUC criterion seems to lead to reasonable results. A natural question is: why? Alternatively, we could, e.g., take different values $f(\alpha)$ with different weights $w(\alpha)$ and compare the weighted values $\int w(\alpha)\cdot f(\alpha)\,d\alpha$; so why AUC?

In this paper, we provide a possible explanation for the empirical efficiency of the Area Under the Curve criterion.

2 Our Explanation

Analysis of the situation. In practice, we usually have bound on both types of error, i.e., we have bounds α_0 and β_0 for which we would like to have $\alpha \le \alpha_0$ and $\beta \le \beta_0$.

Can we achieve this requirement by using a hypothesis testing technique with a given curve $f(\alpha)$? If we cannot achieve the desired values β_0 for some $\alpha < \alpha_0$, not all is lost: we may still be able to get the desired probability of the type II error if we allow higher type I errors. Thus, to test whether the given requirements can be achieved, we should take the largest allowed value α_0 of the type I error and check whether for this value, we can get $\beta \le \beta_0$, i.e., whether we have $f(\alpha_0) \le \beta_0$.

This inequality corresponds to the point (α_0, β_0) being above the curve $\beta = f(\alpha)$. If the point (α_0, β_0) is below this curve, this means that for this hypothesis testing technique, the corresponding requirement cannot be satisfied.

For each technique, some requirements can be satisfied, some cannot. A natural measure of the technique's quality is the frequency with which this technique succeeds – i.e., in more precise terms, the probability that this technique will succeed.

To formalize this idea, we need to select a probability distribution on the set of all pairs (α_0, β_0). To estimate the probability that a given pair of probabilities (α_0, β_0) can be achieved, we need to select some probability distribution on the set of all such pairs.

On the unit square $[0, 1] \times [0, 1]$, there are many possible probability distributions. Such situations are ubiquitous in applications of probabilistic methods. In such situations of uncertainty, a reasonable idea is not to pretend that we have less uncertainty than we do – and thus, put of all probability distributions consistent with our knowledge, to select the probability distribution with the largest uncertainty. A natural measure of the distribution's uncertainty is its entropy [1, 2]. Thus, the idea is to select the probability distribution for which the entropy $S = -\int \rho(x) \cdot \ln(\rho(x))\, dx$ is the largest possible; see, e.g., [1].

This indeed explains the AUC. It is known that among all possible probability distributions located on the unit square, the uniform distribution has the largest entropy. For the uniform distribution, the probability that the randomly selected requirements can be implemented by this technique – i.e., that randomly selected pair (α_0, β_0) will be under the curve $\beta = f(\alpha)$ – is equal to the area under this curve.

Thus, when comparing two techniques, we should indeed select the one for which the area under the f-curve is the smallest possible – or, equivalently, the technique for which the area under the g-curve is the largest possible.

Acknowledgments

This work was supported in part by the National Science Foundation grants 1623190 (A Model of Change for Preparing a New Generation for Professional Practice in Computer Science) and HRD-1242122 (Cyber-ShARE Center of Excellence).

References

[1] E. T. Jaynes and G. L. Bretthorst, *Probability Theory: The Logic of Science*, Cambridge University Press, Cambridge, UK, 2003.

[2] H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing Statistics under Interval and Fuzzy Uncertainty*, Springer Verlag, Berlin, Heidelberg, 2012.

[3] D. J. Sheskin, *Handbook of Parametric and Nonparametric Statistical Procedures*, Chapman and Hall/CRC, Boca Raton, Florida, 2011.