Open Science Framework DOI:10.31219/osf.io/***

Consistency and classification of metrics for binary classifiers

K. Dirland
<***@***>

A. S. Lundervold
<***@***>

P.G.L. Porta Mana
<pgl@portamana.org>

(or any permutation thereof)

Draft. 4 March 2022; updated 2 April 2022

abstract

[Luca] I find it very difficult to structure the paper: there seems to be issues at several levels in the development and use of binary classifiers (and classifiers in general) within machine-learning.

Here are some relevant points:

- There should be a distinction between "inference" (or forecast, prediction, guess) and "decision" (or action, choice). In particular, the possible situations we may be uncertain about and the possible decisions available may be completely different things. A clinician, for example, may be uncertain about "cancer" vs "non-cancer", while the choices are about "drug treatment 1" vs "drug treatment 2" vs "surgery".
- Probability theory & decision theory say that in order to make self-consistent decision we need two things: (a) the probabilities for the possible situations, (b) the utilities of the decisions given each possible situation.
- A useful machine-learning algorithm should therefore give us one of two things:
  - either the *probabilities* of the uncertain situations ("cancer" vs "non-cancer" in the example above),
  - or the final decision ("drug treatment 1" vs "drug treatment 2" vs "surgery" in the example above).

  Current machine-learning classifiers do not give us either: the output in the example above would be "cancer" vs "non-cancer", often without probabilities.
- So there are two possible solutions to the problem above:
  - We must build a classifier that outputs probabilities. The 0–1 outputs of current classifiers cannot properly interpreted as probabilities, for various reasons.
  - We must build a classifier that output *decisions*: so not "cancer" vs "non-cancer", but "drug treatment 1" vs etc..

1 Valuation metrics, amounts of data, inferences, and decisions

Let's consider the simple example of a binary classifier and several dilemmas that appear in its development, choice, and use.

At the moment of evaluating different classifier algorithms, or different hyperparameter settings for one algorithm, we are avalanched by a choice of possible evaluation scales: accuracy, area under curve,

F_1-measure, mean square contingency[1] also known as Matthews correlation coefficient[2], precision, recall, sensitivity, specificity, and many others[3]. Only vague guidelines are usually given to face this choice. A thorough analysis and discussion of several such scales was given by Goodman & Kruskal (1954; 1959; 1963; 1972).

We can also ask: are all these scales well-founded and self-consistent? is it possible that the use of any of them leads to contradictions? The literature abounds with studies showing that some scale X may imply hidden contradictions with the data or the assumptions used for our inference, and is therefore worse than some other scale Y. See for example Baker & Pinsky (2001), Lobo et al. (2008), Hand & Christen (2018), Zhu (2020) and Goodman & Kruskal's papers cited above for instances of criticisms of area under the curve, F_1-measure, Matthews correlation coefficient, and other scales.

If we have many more data for one class than for the other – a common predicament in medical applications – we must face the "class-imbalance problem": the classifier ends up classifying all data as belonging to the more numerous class[4], which may be an undesirable action if the misclassification of cases belonging to the less numerous class entails high costs.

The three points above turn out to be tightly related and to have a common solution. We show that

1. the admissible valuation scales for a binary classifier form a two-dimensional family; that is, the choice of a specific scale corresponds to the choice of two numbers. Such choice is problem-dependent and cannot be given a priori.
2. admissible scales are only those that can be expressed as a linear function of the elements of the population-normalized confusion matrix. Scales such as the F_1-measure or the Matthews correlation coefficient are therefore inad***

2 Overview of decision theory

[5]

Decision theory makes a distinction between

[1] Yule 1912 denoted "r" there. [2] Matthews 1975. [3] Sammut & Webb 2017. [4] Sammut & Webb 2017; Provost 2000. [5] Russell & Norvig 2022 ch. 15; Jeffrey 1965; North 1968.

a. the possible situations we are uncertain about

b. the possible choices we can make.

This distinction is important, in fact in some cases the numbers of possible uncertain situations

Appendix: broader overview of binary classification

Suppose we have a population of units or individuals characterized by a possibly multidimensional variable Z and a binary variable $X \in \{0, 1\}$. Different joint combinations of (X, Z) values can appear in this population. Denote by $F(X{=}x, Z{=}z)$, or more simply $F(x, z)$ when there is no confusion, the number of individuals having specific joint values $(X{=}x, Z{=}z)$. This is the absolute frequency of the values (x, z). We can also count the number of individuals having a specific value of $Z{=}z$, regardless of X; this is the marginal absolute frequency $F(z)$. It is easy to see that

$$F(z) = F(X{=}0, z) + F(X{=}1, z) \equiv \sum_x F(x, z) \; . \tag{1}$$

Analogously for $F(x)$.

Select only the subpopulation of individuals that have a specific value $Z{=}z$. In this subpopulation, the *proportion* of individuals having a specific value $X = x$ is $f(x \mid Z = z)$. This is the conditional relative frequency of x given that z. It is easy to see that

$$f(x \mid z) = \frac{F(x, z)}{F(z)} \; . \tag{2}$$

Now suppose that we know all these statistics about this population. An individual coming from this population is presented to us. We measure its Z and obtain the value z. What could be the value of X for this individual? We know that among all individuals having $Z{=}z$ (and the individual before us is one of them) a proportion $f(x \mid z)$ has $X{=}x$. Thus we can say that there is a probability $f(x \mid z)$ that our individual has $X{=}x$. And this is all we can say if we only know Z.

For this individual we must choose among two actions $\{a, b\}$. The utility of performing action a if the individual has $X{=}x$, and given any other known circumstances, is $U(a \mid x)$; similarly for b. If we knew the value of X, say $X = 0$, we would simply choose the action leading to maximal utility:

$$\begin{aligned}
&\text{if}\quad U(a \mid X{=}0) > U(b \mid X{=}0) \quad \text{then choose action } a\,,\\
&\text{if}\quad U(a \mid X{=}0) < U(b \mid X{=}0) \quad \text{then choose action } b\,,\\
&\text{else}\quad \text{it does not matter which action is chosen.}
\end{aligned} \tag{3}$$

But we do not know the actual value of X. We have probabilities for the possible values of X given that $Z{=}z$ for our individual. Since X is uncertain, the final utilities of the two actions are also uncertain; but we can calculate their *expected* values $\bar{U}(a \mid Z{=}z)$ and $\bar{U}(b \mid Z{=}z)$:

$$\begin{aligned}
\bar{U}(a \mid z) &\coloneqq U(a \mid X{=}0)\; f(X{=}0 \mid z) + U(a \mid X{=}1)\; f(X{=}1 \mid z)\,,\\
\bar{U}(b \mid z) &\coloneqq U(b \mid X{=}0)\; f(X{=}0 \mid z) + U(b \mid X{=}1)\; f(X{=}1 \mid z)\,.
\end{aligned} \tag{4}$$

Decision theory shows that the optimal action is the one having the maximal expected utility. Our choice therefore proceeds as follows:

$$\begin{aligned}
&\text{if}\quad \bar{U}(a \mid z) > \bar{U}(b \mid z) \quad \text{then choose action } a\,,\\
&\text{if}\quad \bar{U}(a \mid z) < \bar{U}(b \mid z) \quad \text{then choose action } b\,,\\
&\text{else}\quad \text{it does not matter which action is chosen.}
\end{aligned} \tag{5}$$

The decision procedure just discussed is very simple and does not need any machine-learning algorithms. It could be implemented in a simple algorithm that takes as input the full statistics $F(X, Z)$ of the population, the utilities, and yields an output according to (5).

Our main problem is that the full statistics $F(X, Z)$ is almost universally not known. Typically we only have the statistics $F_{\mathrm{s}}(X, Z)$ of a sample of individuals that come from the population of interest or from populations that are somewhat related to the one of interest. This is where probability theory steps in. It allows us to assign probabilities to all the possible statistics $F(X, Z)$. From these probabilities we can calculate the *expected* value of the conditional frequencies $f(x \mid z)$. Decision theory says that this expected value should then be used, in this uncertain case, in eq. (4) in place of the unknown $f(x \mid z)$. The decision procedure (5) can then be used again.

Probability theory says that in this particular situation the probability of a particular possible statistics $F(X, Z)$ is the product of two factors having intuitive interpretations:

- the probability of observing the statistics $F_s(X, Z)$ of our data sample, assuming the full statistics to be $F(X, Z)$. With some combinatorics it can be shown that this probability is proportional to

$$\exp\left[\sum_{X,Z} F_s(X, Z) \ln F(X, Z)\right] \tag{6}$$

  The argument of the exponential is the cross-entropy between $F_s(X, Z)$ and $F(X, Z)$; this is the reason of its appearance in the loss function used for classifiers[6].

  This factor tells us how much the possible statistics *fits* the sample data; it gives more weight to a better fit.

- the probability of the full statistics $F(X, Z)$ for reasons not present in the data, for example because of physical laws, biological plausibility, or similar.

  This factor tells us whether the possible statistics should be especially considered, or maybe discarded, for reasons that go beyond the data we have seen; in other words, whether the hypothetical statistics would *generalize* well beyond the sample data.

The final probability comes from the balance between these "fit" and "generalization" factors. Note that the first factor becomes more important as the sample size and therefore $F_s(X, Z)$ increases; the sample data eventually determine what the most probable statistics is, if the sample is large enough.

A similar probabilistic reasoning applies if our sample data come not from the population of interest but from a population having at least the same *conditional* frequencies as the one of interest, either $f(X \mid Z)$ or $f(Z \mid X)$.

Bibliography

("de X" is listed under D, "van X" under V, and so on, regardless of national conventions.)

Baker, S. G., Pinsky, P. F. (2001): *A proposed design and analysis for comparing digital and analog mammography special receiver operating characteristic methods for cancer screening*. J. Am. Stat. Assoc. **96**[454], 421–428. DOI:10.1198/016214501753168136.

Bridle, J. S. (1990): *Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition*. Neurocomputing **68**, 227–236. DOI: 10.1007/978-3-642-76153-9_28.

[6] Bridle 1990; MacKay 1992.

Goodman, L. A., Kruskal, W. H. (1954): *Measures of association for cross classifications*. J. Am. Stat. Assoc. **49**[268], 732–764. DOI:10.1080/01621459.1954.10501231. See corrections Goodman, Kruskal (1957; 1958) and also Goodman, Kruskal (1959; 1963; 1972).

— (1957): *Corrigenda: Measures of association for cross classifications*. J. Am. Stat. Assoc. **52**[280], 578. DOI:10.1080/01621459.1957.10501415. See Goodman, Kruskal (1954).

— (1958): *Corrigenda: Measures of association for cross classifications*. J. Am. Stat. Assoc. **53**[284], 1031. DOI:10.1080/01621459.1958.10501492. See Goodman, Kruskal (1954).

— (1959): *Measures of association for cross classifications. II: Further discussion and references*. J. Am. Stat. Assoc. **54**[285], 123–163. DOI:10.1080/01621459.1959.10501503. See also Goodman, Kruskal (1954; 1963; 1972).

— (1963): *Measures of association for cross classifications. III: Approximate sampling theory*. J. Am. Stat. Assoc. **58**[302], 310–364. DOI:10.1080/01621459.1963.10500850. See correction Goodman, Kruskal (1970) and also Goodman, Kruskal (1954; 1959; 1972).

— (1970): *Corrigenda: Measures of association for cross classifications. III: Approximate sampling theory*. J. Am. Stat. Assoc. **65**[330], 1011. DOI:10.1080/01621459.1970.10481142. See Goodman, Kruskal (1963).

— (1972): *Measures of association for cross classifications, IV: Simplification of asymptotic variances*. J. Am. Stat. Assoc. **67**[338], 415–421. DOI:10.1080/01621459.1972.10482401. See also Goodman, Kruskal (1954; 1959; 1963).

Hand, D., Christen, P. (2018): *A note on using the F-measure for evaluating record linkage algorithms*. Stat. Comput. **28**[3], 539–547. DOI:10.1007/s11222-017-9746-6.

Jeffrey, R. C. (1965): *The Logic of Decision*. (McGraw-Hill, New York).

Lobo, J. M., Jiménez-Valverde, A., Real, R. (2008): *AUC: a misleading measure of the performance of predictive distribution models*. Glob. Ecol. Biogeogr. **17**[2], 145–151. DOI:10.1111/j.1466-8238.2007.00358.x, https://www2.unil.ch/biomapper/Download/Lobo-GloEcoBioGeo-2007.pdf.

MacKay, D. J. C. (1992): *The evidence framework applied to classification networks*. Neural Comput. **4**[5], 720–736. http://www.inference.phy.cam.ac.uk/mackay/PhD.html, DOI:10.1162/neco.1992.4.5.720.

Matthews, B. W. (1975): *Comparison of the predicted and observed secondary structure of T4 phage lysozyme*. Biochim. Biophys. Acta **405**[2], 442–451. DOI:10.1016/0005-2795(75)90109-9.

North, D. W. (1968): *A tutorial introduction to decision theory*. IEEE Trans. Syst. Sci. Cybern. **4**[3], 200–210. DOI:10.1109/TSSC.1968.300114, https://stat.duke.edu/~scs/Courses/STAT102/DecisionTheoryTutorial.pdf.

Provost, F. (2000): *Machine learning from imbalanced data sets 101*. Tech. rep. WS-00-05-001. (AAAI, Menlo Park, USA). https://aaai.org/Library/Workshops/2000/ws00-05-001.php.

Russell, S. J., Norvig, P. (2022): *Artificial Intelligence: A Modern Approach*, Fourth Global ed. (Pearson, Harlow, UK). First publ. 1995.

Sammut, C., Webb, G. I., eds. (2017): *Encyclopedia of Machine Learning and Data Mining*, 2nd ed. (Springer, Boston). DOI:10.1007/978-1-4899-7687-1. First publ. 2011.

Yule, G. U. (1912): *On the methods of measuring association between two attributes*. J. R. Stat. Soc. **75**[6], 579–652. DOI:10.1111/j.2397-2335.1912.tb00463.x.

Zhu, Q. (2020): *On the performance of Matthews correlation coefficient (MCC) for imbalanced dataset*. Pattern Recognit. Lett. **136**, 71–80. DOI:10.1016/j.patrec.2020.03.030.