Analysis of Binary Classification Metrics Using Empirical Distributions of Prediction Scores

A. Chtcheprov[1], **S. Krovvidy**[1], **and H. Vera**[1]
[1]Advanced Analytics Enablement, Fannie Mae, Washington, DC, USA

Abstract - *Analysis of empirical distribution functions of class prediction scores generated by underlying binary classifiers is a very powerful technique to evaluate Machine Learning algorithms. We develop a methodology of using empirical functions and introduce a family of evaluation metrics called score Distance Measures (DM) that characterize a spatial distribution and a separation of the prediction scores. The paper also implements the methodology to re-examine other classification metrics such as AUC, Recall, Precision, and Misclassification Error. We illustrate the technique with a number of examples. The methodology allows business and data science practitioners to compare evaluation measures and select the optimal Machine Learning models.*

Keywords: machine learning, binary classifier, empirical distribution, evaluation metric

1 Introduction

Binary classifiers are very popular Machine Learning (ML) methods for practical business applications across many industries. Given historical data, the goal is to build a quantitative model that would assign new data points to one of the classes (called positive and negative). Modern software implementations of binary classification methods generate class prediction scores. To select the best ML model, business practitioners apply certain comparison criteria called evaluation metrics. There are many metrics currently used in practical applications (see review in [1]). Another class of evaluation techniques compares separations of prediction score curves [2]. A *Decision Curve* Analysis assesses net benefits against threshold probability [3]. Our paper describes a methodology of using the prediction score distributions to evaluate binary ML classifiers. We introduce a family of evaluation measures called *DM* that characterize a distribution and a spatial separation of prediction score sets. We also apply the developed methodology to analyze and compare several exiting metrics. We run Monte-Carlo simulations to illustrate the performance of our framework.

2 Mathematical Formalism

Given a ML binary classifier, to characterize the distribution of scores generated for the positive and negative classes, we consider two random variables (r.v.) $S_+(\omega)$ and $S_-(\omega)$ defined on probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$. These scores are interpreted as probabilities that predictions would belong to the positive class. It is common to assume that $\Omega = \mathbb{R}$ and select a Borel σ -algebra $\mathcal{F} = \mathcal{B}(\mathbb{R})$ [4]. Let $F_+(s)$ and $F_-(s)$ be the cumulative distribution functions of r.v. $S_+(\omega)$ and $S_-(\omega)$. We also assume absolute continuity of $F_\pm(s)$, i.e. the existence of probability densities $f_\pm(s)$ such that

$$F_\pm(s) = \int_0^s f_\pm(x)dx \qquad (1)$$

In practical applications, distributions (1) are commonly represented as empirical distribution functions. Denote by N_+ and N_- the number of data points in the positive and negative classes. Let $\{s_i^+\}_{i=1}^{N_+}$ and $\{s_i^-\}_{i=1}^{N_-}$ be ordered sets of positive and negative scores generated by the underlying ML algorithm. These sets are considered as random samples generated according to the cumulative distribution functions of r.v. $S_+(\omega)$ and $S_-(\omega)$. The corresponding empirical probability densities can be expressed [5]

$$f_\pm^{(N_\pm)}(s) = \frac{1}{N_\pm}\sum_{i=1}^{N_\pm}\delta(s - s_i^\pm) \qquad (2)$$

In (2), $\delta(s)$ is a Dirac delta function. The notations for empirical probability densities (2) emphasize the dependence of these functions on the score sample sizes N_+ and N_-. From (2) we get the representations for empirical cumulative distributions [6]:

$$F_\pm^{(N_\pm)}(s) = \frac{1}{N_\pm}\sum_{i=1}^{N_\pm}\theta(s - s_i^\pm) \qquad (3)$$

In (3), $\theta(x)$ is a Heaviside step function. The empirical distributions are also step functions which are differentiable a.e. (almost everywhere). According to the Glivenko-Cantelli theorem, if the sample sizes increase the empirical distributions uniformly converge (in probability) to the real distributions [6]. The rate of conversion is estimated based on the Kolmogorov-Smirnov and Iterated Logarithm statistical laws [6]. To simplify notations, the upper indexes appeared in LHS of (2) and (3) to denote the empirical distributions will be dropped. The same notations $f_\pm$ and $F_\pm$ for both the real and empirical functions will interchangeably be used throughout the text.

3 Analysis of Binary Metrics

3.1 Score Distance Measures (DM)

Consider a family of metrics:

$$DM[K,f_+,f_-]=\int_0^1 K(s)[f_+(s)-f_-(s)]ds \qquad (4)$$

Functional (4) depends on probability densities $f_\pm(s)$ and kernel function $K(s)$. The family of metrics $DM[K,f_+,f_-]$ describes the differences of "moments" calculated for the positive and negative distributions and, thus, evaluates a spatial separation of the positive and negative score sets. Metrics (4) are called score *Distance Measures* (*DM*). Along with a Kullback-Leibler divergence [5], which characterizes the deviation of two distributions, measures (4) describe the spatial "concentrations" of the positive and negative scores near points $s=1$ and $s=0$ correspondingly. Assume that kernel $K(s)$ is a continuously differentiable function on interval [0, 1] (to guarantee the existence of integral (5) below). We also assume that $0\le K(t)\le 1$, $K(0)=0$, and $K(1)=1$. Integrating (4) by parts yields:

$$DM[K,f_+,f_-]=\int_0^1[F_-(s)-F_+(s)]\frac{dK(s)}{ds}ds \qquad (5)$$

If $f_+(s)=f_-(s)$, which would mean a very poor binary classifier, then $DM=0$. We also note that

$$-1\le DM[K,f_+,f_-]\le 1.$$

Indeed, since all functions in (4) are non-negative, we get:

$$-1=-\max[K(s)]\int_0^1 f_-(s)ds\le$$
$$-\int_0^1 K(s)f_-(s)ds\le DM[K,f_+,f_-]\le$$
$$\int_0^1 K(s)f_+(s)ds\le \max[K(s)]\int_0^1 f_+(s)ds=1.$$

If all positive scores are in the vicinity of $s=1$ and all negative scores are in the neighborhood of $s=0$ then *DM* is close to 1. To show that let us consider the extreme situation: $f_+(s)=\delta(s-1)$ and $f_-(s)=\delta(s)$. We have: $DM[K,\delta(s-1),\delta(s)]=1$. On the other hand, in another extreme case $f_-(s)=\delta(s-1)$ and $f_+(s)=\delta(s)$ we get: $DM[K,\delta(s),\delta(s-1)]=-1$. A negative sign means a reverse classifier (similar to the situation when $AUC<0.5$). Hence, without loss of generality, we always assume that $0\le DM_1[K,f_+,f_-]\le 1$ so *DM* can be used as a measure of ML performance.

Consider the following kernels:

$$K(s)=\left\{s,\ s^2,\ s^3,\frac{\ln(1+s)}{\ln(2)},\frac{(1+s)\ln(1+s)}{2\ln(2)}\right\},$$

where $\ln(x)$ is a natural logarithm. Using (5) we define the corresponding metrics for the above kernels:

$$DM_n=n\int_0^1 s^{n-1}[F_-(s)-F_+(s)]\,ds,\ n=1,2,3 \qquad (6a)$$

$$DM_{entropy}=\frac{1}{\ln(2)}\int_0^1[F_-(s)-F_+(s)]\frac{ds}{1+s} \qquad (6b)$$

$$DM_{log}=\frac{1}{2\ln(2)}\int_0^1[1+\ln(1+s)][F_-(s)-F_+(s)]\,ds \qquad (6c)$$

Geometrically, measure DM_1 is the area between curves $F_-(s)$ and $F_+(s)$. It also shows the average signed "distance" of the spatial separation between the positive and negative sets. This can be verified by direct calculations taking into account that DM_1 has another representation:

$$DM_1[K,f_+,f_-]=\iint_0^1(x-y)f_+(x)f_-(y)dxdy$$

It follows from (4) that all metrics (6) compute the difference between "average" values $\overline{K(s_+)}-\overline{K(s_-)}$. In practice *DM* metrics (6a) - (6c) are evaluated by numerical integration using (3). Similar to a computation of *AUC*, interval [0, 1] is partitioned by a sequence of M non-overlapping subintervals $\{[s_{i-1},s_i]\}_{i=1}^M$ and a corresponding numerical method, e.g. the trapezoid rule, is applied to compute the integrals.

Comparing (6) and *AUC* calculated as

$$AUC=\int_0^1[1-F_+(s)]f_-(s)ds=\int_0^1 F_-(s)f_+(s)ds \qquad (7)$$

we conclude that *AUC* is less sensitive to the degree of the score set separation. To explain this, imagine that all values of the positive score set belong to interval $[p,1]$ and the values of the negative score set are on interval $[0,q]$. Let $0\le q<p\le 1$. As long as there is a full separation of the score sets, $AUC=1$ regardless of the difference $(p-q)$ in contrast to *DM*. This is because *AUC* shows the probability that a score of a randomly selected record from the positive class will be higher than a score of a record randomly selected from the negative class [3]. This can also be concluded based on (7):

$$AUC=\int_0^1 F_-(s)f_+(s)ds=\int_0^1 ds\ f_+(s)\int_0^s f_-(x)dx$$

3.2 Recall & Precision

Recall as a function of the score threshold $t\in[0,1]$ is defined as $Re(t)=1-F_+(t)$. *Recall* is usually used together with other metrics, e.g. *Precision*. $Re(t)$ is a non-increasing function of t because its derivative $dRe(t)/dt=-f_+(t)\le 0$. Obviously, $Re(0)=1$ and $Re(1)=0$. *Recall* characterizes the distribution of the positive scores only. A graph $Re(t)$ starts at value 1 and eventually drops to 0. "Initial" values of $Re(t)$ may be equal to 1 on some interval $[0,t_R]$ where $F_+(s)=0$ before a graph $Re(t)$ starts decaying. For stronger classifiers points t_R are more shifted to the right

meaning that the positive score points are in the vicinity of $s = 1$. For weaker classifiers, the positive score points may be found in the neighborhood of $s = 0$ so values of t_R are closer (or equal) to zero. These observations suggest that the quality of the classifiers can be characterized by the length of interval $[0, t_R]$. A *Recall Unit Interval* (*RUI*) metric is defined in terms of a support of $F_+(s)$:

$$RUI = 1 - \lambda\{supp(F_+)\} = t_R \quad (8)$$

where $\lambda\{..\}$ is a Lebesgue measure. *RUI* can be used as a classifier evaluation metric along with *AUC*. The shapes of the $Re(t)$ plots suggest another metric, the *Area under the Recall Curve*, defined as

$$AURC = \int_0^1 Re(s)ds = \int_0^1 sf_+(s)ds = \overline{s_+} \quad (9)$$

According to (9), $0 \le AURC \le 1$ and is the average score $\overline{s_+}$ of the positive class set. In the extreme case $f_+(s) = \delta(s-1)$, $AURC = 1$. If $AURC < 0.5$, positive and negative class labels should be reversed. $AURC = 0.5$ corresponds to the score with a uniform probability density. In this case, $dRe(t)/dt = -1$ so the plot $Re(t)$ is a straight line connecting points (0, 1) and (1, 0) on the coordinate plane.

Let $\beta = N_+/N_-$ be the class imbalance ratio. *Precision* is expressed in terms of the cumulative distributions as follows:

$$Pr(t) = \frac{\beta[1-F_+(t)]}{\beta[1-F_+(t)]+1-F_-(t)} \quad (10)$$

$\Pr(0) = \frac{\beta}{1+\beta}$. Estimation of $\Pr(1)$ results in a 0/0 situation. We can define $\Pr(1)$ by continuity provided the limiting value is finite. The 0/0 situation makes the problem of estimation of *Precision* in the vicinity of $t = 1$ ill-posed. To define (10) at point $t = 1$ we apply the L'Hospital's rule:

$$\Pr(1) = \lim_{t\to 1}\frac{\beta f_+(t)}{\beta f_+(t)+f_-(t)} = \left[1 + \frac{f_-(1)}{\beta f_+(1)}\right]^{-1} \quad (11)$$

The first derivative of $Pr(t)$ is

$$\frac{dPr}{dt} = \beta^2\frac{f_-(t)[1-F_+(t)]-f_+(t)[1-F_-(t)]}{[1+\beta-\beta F_+(t)-F_-(t)]^2} \quad (12)$$

If $f_+(1) \neq 0$, which is usually the case in practice, then limit (11) exists. First, we observe that $\Pr(1) \le 1$. For weak classifiers, it is possible that $f_-(1) > 0$, thus, making $\Pr(1) < 1$. If $f_-(1) = 0$ (the case of stronger classifiers) $\Pr(1) = 1$. A *Precision Unit Interval* (*PUI*) metric characterizes interval $\{t: Pr(t) = 1\}$ and is defined as

$$PUI = \lambda\{t \in [0,1]: F_-(t) = 1\} \quad (13)$$

Second, we observe that *Precision* is very sensitive to the class imbalance. If $\beta \ll 1$ then both $\Pr(0)$ and $\Pr(1)$ may be very small. Third, the *Precision* curves may "oscillate" in the vicinity of $t = 1$. Such a behavior is due to the 0/0 situation described above. The same 0/0 situation also occurs for derivative (12). The shape of the *Precision* graphs suggests another metric, the *Area under the Precision Curve*

$$AUPC = \frac{\int_0^1 \Pr(t)dt - \Pr(0)}{1-\Pr(0)} = (1+\beta)\int_0^1 \Pr(t)\,dt - \beta \quad (14)$$

First, we observe that $-1 \le AUPC \le 1$. Indeed, for the ideal classifier $f_+(t) = \delta(t-1)$ and $f_-(t) = \delta(t)$, $AUPC = 1$. In another extreme case $f_+(t) = \delta(t)$ and $f_-(t) = \delta(t-1)$, $AUPC = -1$, meaning that the positive and negative class labels should be reversed. Second, for very weak classifiers $AUPC = 0$. This result follows from (10) and (14) if we substitute $F_+(s) = F_-(s)$ to get $\Pr(t) = \beta/(1+\beta) = \Pr(0)$ for all values of t.

3.3 Misclassification Error

As is known, the *Misclassification Error* $e(t)$ computed as the proportion of the incorrectly classified predictions may be a very misleading evaluation metric in case of a significant class imbalance, e.g. $\beta \ll 1$. We consider an *Error Ratio*

$$e_r(t) = \frac{e(t)}{e_{tr}} = \frac{N_-}{N_+} + F_+(t) - \frac{N_-}{N_+}F_-(t) \quad (15)$$

Differentiating (15) and equating the derivative to zero we find that the necessary condition of the minimum of $e_r(t)$ is $N_+f_+(t_{min}) = N_-f_-(t_{min})$, which is the same as the condition of the minimum for $e(t)$. In the ideal case $f_+(s) = \delta(s-1)$ and $f_-(s) = \delta(s)$, $e_r(t) = e(t) = 0$ for any t.

4 Preliminary Results

For a preliminary illustration of the methodology we perform Monte-Carlo simulations. In each experiment we create two score sets to model the positive and negative classes, as if the scores were generated by some ML method. We proceed as follows. First, random numbers from the normal distribution are drawn and points which are outside of interval [0, 1] are removed. Then we apply additional "sampling" to make sure all final samples have pre-defined sizes which are set to $N_+ = 8000$ and $N_- = 10000$. For the 1st experiment that simulates a strong classifier, the mean values of the normal distribution for the positive and negative score sets are taken 1 and 0 and standard deviations are 0.15 and 0.05. The empirical distributions $F_+(s)$ and $F_-(s)$ are shown in Fig. 1. A large gap between the distribution curves indicates a substantial spatial separation of the positive and negative score sets. The corresponding *DM* values along with other measures are shown in Table 1. As expected, $AUC = 1$ while all *DM* are less than 1 showing a greater sensitivity to a degree of the set separation. In the 2nd experiment ("intermediate" classifier) the means are set to 1 and 0 and standard deviations are 0.35 and 0.45. A separation

of the $F_+(s)$ and $F_-(s)$ curves shown in Fig. 2 is not very strong i.e. the score sets have some overlap.

TABLE 1

	Experiment 1	Experiment 2	Experiment 3
DM_1	0.84	0.38	0.17
DM_2	0.78	0.39	0.17
DM_3	0.71	0.36	0.15
$DM_{entropy}$	0.85	0.37	0.16
DM_{log}	0.83	0.39	0.17
AUC	1.00	0.88	0.67
AURC	0.88	0.73	0.61
AUPC	0.93	0.47	0.22

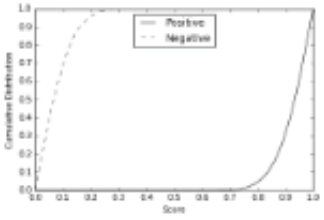

Fig. 1 Experiment 1.

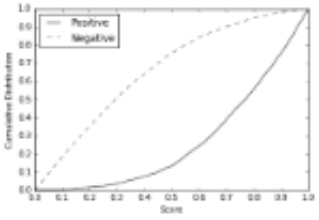

Fig. 2. Experiment 2.

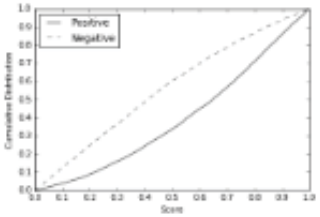

Fig 3. Experiment 3.

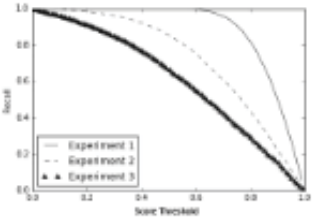

Fig. 4. Recall.

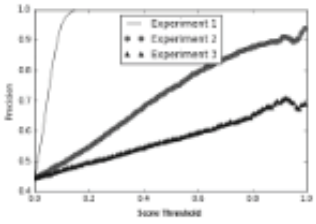

Fig. 5. Precision.

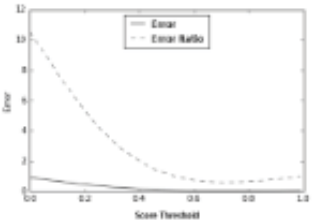

Fig. 6. Error Ratio.

In the 3rd experiment ("weak classifier") the same means as in the previous case are used. Standard deviations are set to 0.6 and 0.8. According to Fig. 3, there is a substantial overlap between the score sets. Fig. 4-5 show the corresponding *Recall* and *Precision* graphs. We observe some "oscillations" of two *Precision* curves near $t = 1$ due to ill-conditioning. The plots also illustrate advantages of *RUI* and *PUI* metrics. For the strong classifier, they are about 0.55 and 0.85 respectively, while for the weak classifier, the metrics are close to 0. *RUI* and *PUI* characterize spatial distributions of the score sets so both metrics should complement each other and *AUC*. Fig. 6 compares *Misclassification Error* and *Error Ratio* graphs in case of large class imbalance ($\beta = 0.095$). Both metrics attain minimum at the same score threshold (about 0.7). The *Misclassification Error* is very misleading because of very small values that may be wrongly interpreted as good results. The *Error Ratio* works very well.

5 Conclusion

Based on the empirical distributions (3) we develop the methodology to evaluate the performance of binary ML classifiers. Using this methodology, the paper introduces a family of *DM* measures (6) as well as re-visits other popular metrics. To illustrate the techniques we run three simulations to compare stronger and weaker classifiers. *DM* measures characterize a spatial separation of the positive and negative score sets and are more sensitive to that separation compared to *AUC*. In practice both metrics can complement each other. *Precision* (10) may exhibit an ill-posed behavior near point $t = 1$ and is also affected by the class imbalance. The other metrics, *RUI* (8), *AURC* (9), *PUI* (13) and *AUPC* (14) characterize spatial score set distributions and should be used along with *AUC*. In case of class imbalance, *Error Ratio* should be used instead of the *Error* metric. The preliminary results of our research look very promising, so we continue exploring the technique to apply it to real world problems.

6 Acknowledgment

We thank Garri Yakobson, Justin Smith and Scott Reed for reviewing an earlier version of the paper.

7 References

[1] M. Sokolova and G. Lapalme. "A Systematic analysis of performance measures for classification tasks"; Inf. Process. Manag., Vol. 45, 427-437, July 2009.

[2] J.D. Kelleher, B. M. Namee, and A. D'Arcy. "Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies". The MIT Press, 2015.

[3] A. Vickers and E. Elkin. "Decision curve analysis: a novel method for evaluating prediction models"; Med. Decis. Making, Vol. 26, 565-574, November-December 2006.

[4] A. N. Shiryaev. "Probability", 2nd ed. Springer-Verlag, 1995.

[5] D. Barber. "Bayesian Reasoning and Machine Learning". Cambridge University Press, 2012.

[6] V. N. Vapnik. "Statistical Learning Theory". John Wiley and Sons, 1998.