Binary Classification Metrics

A quick tour through binary classification metrics, laying out the different terminology used in different domains, and explaining (and motivating) some of the composite metrics.

Contents

1. Logical definitions 1
2. Basic metrics 1
3. Synonyms used in different domains 2
4. Composite metrics 3
   4.1. Accuracy 3
   4.2. Jaccard Index 3
   4.3. F-Score 3
   4.4. The Sørensen-Dice coefficient 5
   4.5. The Matthews Correlation Coefficient (MCC) 6
   4.6. Informedness and Markedness 7

1. Logical definitions

A thing is either in the set of interest or not. Label them "Yes" or "No".

A test tries to predict the truth. It asserts either a "Positive" or "Negative".

	Yes	**No**
Positive	True Positive	False Positive
Negative	False Negative	True Negative

Table 1: The four combinations of true label and predicted label

2. Basic metrics

The ratios to the truth (the columns):

$$\text{True Positive Rate} = P(\text{True Positive} \mid \text{Yes})$$
$$\text{False Negative Rate} = P(\text{False Negative} \mid \text{Yes})$$
$$\text{True Negative Rate} = P(\text{True Negative} \mid \text{No})$$
$$\text{False Positive Rate} = P(\text{False Positive} \mid \text{No})$$

The ratios to the prediction (the rows):

$$\text{Positive Predictive Value} = P(\text{True Positive} \mid \text{Positive})$$
$$\text{False Discovery Rate} = P(\text{False Positive} \mid \text{Positive})$$
$$\text{Negative Predictive Value} = P(\text{True Negative} \mid \text{Negative})$$
$$\text{False Omission Rate} = P(\text{False Negative} \mid \text{Negative})$$

Accuracy is just about whether the test is correct:

$$\text{Accuracy} = P(\text{True Positive} \vee \text{True Negative})$$

3. Synonyms used in different domains

	Statistics Terminology	**Machine Learning Terminology**	**Medical Diagnostics**	**Signal Detection Theory**
True Positive Rate (TPR)	Statistical Power $(1 - \beta)$	Recall	Sensitivity	Probability of Detection (P_d), Hit Rate
True Negative Rate (TNR)	Correct Rejection Rate $(1 - \alpha)$	Selectivity	Specificity	Correct Rejection Rate
Positive Predictive Value (PPV)	—	Precision	PPV	Precision
Negative Predictive Value (NPV)	—	NPV	NPV	—
False Positive Rate (FPR)	Type I Error Rate (α)	Fall-out	1 — Specificity	False Alarm Rate (P_{fa})
False Negative Rate (FNR)	Type II Error Rate (β)	Miss Rate	1 — Sensitivity	Miss Rate
False Discovery Rate (FDR)	FDR	1-Precision	1-PPV	—
False Omission Rate (FOR)	False Non-Discovery Rate (FNDR)	1-NPV	1-NPV	—

Table 2: Some of the equivalent terms used in other domains

4. Composite metrics

4.1. Accuracy

Note that Accuracy can be written as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

4.2. Jaccard Index

In a dataset with many more True Negatives than True Positives, the accuracy becomes saturated. A simplistic interpretation of the Jaccard Index is that it addresses this problem by simply removing the True Negatives from Accuracy to produce the metric:

$$\text{Jaccard Index} = \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}}$$

The Jaccard Index is also sometimes called the Jaccard Similarity Coefficient, which provides a more motivating derivation.

A similarity metric is a mirror image of a distance/dissimilarity metric. A similarity *coefficient* is such a metric based upon a ratio. Jaccard's formula is:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|}$$

The (Jaccard) similarity between the predicted set (**Positive**) and the actual set (**Yes**) is

$$\begin{aligned} J(\text{Positive}, \text{Yes}) &= \frac{|\text{Positive} \cap \text{Yes}|}{|\text{Positive} \cup \text{Yes}|} \\ &= \frac{|\text{True Positive}|}{|\text{True Positive} \cup \text{False Positive} \cup \text{False Negative}|} \\ &= \frac{\text{TP}}{\text{TP} + \text{FP} + \text{FN}} \end{aligned}$$

which is the same as before.

In passing, note that the Jaccard Distance = $1 - J$, satisfying the properties of distances (triangle inequality etc).

4.3. F-Score

To motivate, start by assuming we want a single measure sensitive to the two types of incorrect test result: False Positives and False Negatives. We need to normalise the contributions from each, i.e. make them relative to something, and a natural choice is to normalise using the number of True Positives.

- A test with high Precision (PPV) has relatively few False Positives (compared with True Positives).
- A test with high Recall (TPR) has relatively few False Negatives (compared with True Positives).

So Precision (P) and Recall (R) will be our normalised measures:

$$\begin{aligned} \text{High P} &\longrightarrow \text{Few FP} \\ \text{High R} &\longrightarrow \text{Few FN} \end{aligned}$$

Some kind of weighted average of P and R would create a parameterised score which can be used to choose which of FP and FN is more important.

Different averages (arithmetic, harmonic, geometric, ...) have different properties. To help us choose, consider these scenarios:

- High P, Zero R: A search engine that returns only 1 absolutely certain ($P \approx 1$) result out of a million relevant documents ($R \approx 0$) is useless.
- Zero P, High R: A search engine that returns the entire internet for every query catches all relevant documents ($R \approx 1$), is equally useless because it provides zero signal ($P \approx 0$).

So we would like the score to be 0 in the extreme cases of $P = 0$ and $R = 1$, or $P = 1$ and $R = 0$.

This rules out using the arithmetic mean. In a sense, the arithmetic mean is an "or" combination, but we want an "and" combination to make a reasonable score imply that both P and R are "good".

The geometric mean ($\sqrt{PR}$) has the right property, and is known as the G-Measure, but the harmonic mean is better for a number of reasons:

1. It is more punishing of imbalance because of the reciprocals $\frac{1}{P}$ and $\frac{1}{R}$. Or to see it another way, $\text{HM} \leq \text{GM} \leq \text{AM}$.
2. Unlike the GM, the HM has a simple expression of counts without square roots. This simplifies analysis and gradient calculations for optimisation.
3. In general:
   a) AM is good for additive quantities (P and R are not additive quantities);
   b) GM is good for multiplicative growth factors, compounding ratios, or metrics on vastly different scales (do not apply to P and R);
   c) HM is good for averaging rates or ratios where the numerator is fixed and the denominator is the thing of interest (P and R share a common numerator, TP).

So we use the weighted HM, which has the form:

$$F = \frac{w_p + w_r}{\frac{w_p}{P} + \frac{w_r}{R}}$$

Since we only need to parameterise the *relative* weighting, we can arbitrarily set $w_p = 1$ and then chose either $w_r > 1$ to make R contribute more to the score than P, or $w_r < 1$ to make P contribute more than R, or just $w_r = 1$ for equal contribution.

$$F_{w_r} = \frac{1 + w_r}{\frac{1}{P} + \frac{w_r}{R}}$$

This parameter w_r is adjusting the meaning of the metric F, which is either being used to assess a test, compare tests, or optimise a test algorithm. Optimising F_{w_r} with a high w_r will emphasise R at the expense of P, and probably result in higher R (fewer False Negatives) and lower P (more False Positives).

To help interpret w_r, consider when the marginal contribution from P is the same as from R, e.g. when a worsening P contributes the same to F as a worsening R of the same amount. This condition is satisfied when:

$$\frac{\partial F}{\partial P} = \frac{\partial F}{\partial R}$$

$$\frac{-F_{\hat{w}_r}^2}{1+\hat{w}_r} \times -\frac{1}{P^2} = \frac{-F_{\hat{w}_r}^2}{1+\hat{w}_r} \times -\frac{\hat{w}_r}{R^2}$$

$$\hat{w}_r = \left(\frac{R}{P}\right)^2$$

So if we chose a w_r and used F_{w_r} to optimise our test algorithm, it is possible that it reaches the point where marginal precision and marginal recall trade off equally. If that happens, then

$$\frac{R}{P} = \sqrt{w_r}$$

This suggests that replacing our w_r parameter with (say) β^2 makes the parameter slightly more interpretable, concluding our attempt to motivate the conventional form of the F_β-score:

$$F_\beta = \frac{1+\beta^2}{\frac{1}{P} + \frac{\beta^2}{R}} = \frac{(1+\beta^2)PR}{\beta^2 P + R}$$

- Setting $\beta = 1$ yields the standard F_1-score, treating Precision and Recall equally.
- Setting $\beta < 1$ places more emphasis on Precision, making it ideal for scenarios like spam detection where False Positives are costly.
- Setting $\beta > 1$ prioritizes Recall, making it suitable for high-stakes screening tasks (such as medical diagnostics or fraud detection), where missing a true positive (False Negative) carries severe consequences.

4.4. The Sørensen-Dice coefficient

In ecology, the F_1-score is called the Sørensen-Dice coefficient, independently developed by botanists Thorvald Sørensen and Lee Raymond Dice. In medicine, the F_1-score is called the Dice Coefficient or Dice Similarity Coefficient (DSC).

One derivation starts with two coverage fractions

$$R = \frac{|A \cap B|}{|A|}, \qquad P = \frac{|A \cap B|}{|B|}$$

The DSC can be defined as the harmonic mean of these two coverage fractions R and P,

$$\begin{aligned} D &= \frac{1}{\frac{1}{R} + \frac{1}{P}} \\ &= \frac{2\,|A \cap B|}{|A| + |B|} \end{aligned}$$

In the context of binary classification,

$$A = \text{Yes}, \qquad B = \text{Positive}$$

making R equal to Recall and P equal to Precision. Then DSC becomes

$$
\begin{aligned}
D &= \frac{2\ |\text{Yes} \cap \text{Positive}|}{|\text{Yes}| + |\text{Positive}|} \\
&= \frac{2\ \text{TP}}{(\text{TP} + \text{FN}) + (\text{TP} + \text{FP})} \\
&= \frac{2\ \text{TP}}{2\ \text{TP} + \text{FP} + \text{FN}}
\end{aligned}
$$

which is just another expression for F_1.

The DSC can also be seen as a variant of the Jaccard similarity which gives more credit (or weight) to the overlap. Start with the Jaccard, J,

$$
\begin{aligned}
J &= \frac{|A \cap B|}{|A \cup B|} \\
&= \frac{|A \cap B|}{|A - B| + |A \cap B| + |B - A|} \\
&\quad \text{... 'double weight' the } |A \cap B| \text{ in both numerator and denominator ...} \\
&\rightarrow \frac{2|A \cap B|}{|A - B| + 2|A \cap B| + |B - A|} \\
&= \frac{2\ |A \cap B|}{|A| + |B|}
\end{aligned}
$$

which is the DSC above.

Basic manipulation shows

$$
\begin{aligned}
D &= \frac{2J}{1 + J} \\
&\geq \frac{2J}{1 + 1}, \qquad \because J < 1 \\
&= J
\end{aligned}
$$

i.e. the DSC is never less than the Jaccard.

4.5. The Matthews Correlation Coefficient (MCC)

By design, apart from accuracy, none of the compound metrics above are a function of the True Negatives. They are also all built from the *harmonic-mean-of-fractions* family.

MCC takes a different approach. It treats the true label and the predicted label as two binary (0/1) numeric variables, and takes the Pearson correlation coefficient between them, returning a value between −1 and +1.

$$
\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}}
$$

Crucially, it is a function of all 4 quadrants of Table 1. Swapping labels does not alter the value.

As you would expect,

- +1 means perfect correlation between prediction and truth,
- 0 means the prediction is no better than random guessing,
- −1 means the prediction is systematically anti-correlated with the truth

Note the F_β/Dice/Jaccard metrics have no ability to indicate "no better than random guessing".

4.6. Informedness and Markedness

This is another approach which takes all 4 quadrants of Table 1 into account.

Informedness sums the two columns, True Positive Rate and True Negative Rate, and subtracts 1:

$$\begin{aligned}\text{Informedness} &= \text{TPR} + \text{TNR} - 1 \\ &= \text{TPR} - \text{FPR}, \qquad \because \text{TNR} = 1 - \text{FPR}\end{aligned}$$

By subtracting 1, an Informedness of 1 indicates a perfect test. Random guessing will result in $\text{TPR} = \text{FPR}$ and hence an Informedness of 0.

Markedness is like Informedness, but sums the two rows instead of columns:

$$\begin{aligned}\text{Markedness} &= \text{PPV} + \text{NPV} - 1 \\ &= \text{NPV} - \text{FDR}, \qquad \because \text{PPV} = 1 - \text{FDR}\end{aligned}$$

Informedness is deductive (given the truth, how well do we predict it), and Markedness is abductive (given a prediction, how well does it indicate the truth). Together they form a complementary pair. One asks "how informative is the ground truth about getting the right call?"; the other asks "how much does a given call actually mark the true state?".

As it happens. MCC is the geometric mean of Informedness and Markedness,

$$\text{MCC} = \sqrt{\text{Informedness} \times \text{Markedness}}$$