**Matthews correlation to evaluate a classifier**

An often underestimated problem in the data science community is the choice of the performance index to assess the goodness of a machine learning algorithm. In this post I will discuss pro’s and con’s of the main performance indices used to evaluate a **binary classifier** (footnote: same results holds for multi-class classifiers). Then I will discuss about an overlooked (according to the author of this post) performance index known in the literature as **Matthews correlation** and its connection with well-known performance indices.

Accuracy, Recall, Precision and F1 score

In the title of this section I cited the most commonly used performance indices. Given a binary **contingency matrix** of a binary classifier,

|  |  |  |
| --- | --- | --- |
|  | Actual positives | Actual negatives |
| Predicted positives | T(rue)P(ositives) | F(alse)P(ositives) |
| Predicted negatives | F(alse)N(egatives) | T(rue)N(egatives) |

Accuracy, Recall, Precision and F1 score are defined as:

Accuracy = (TP + TN) / N

Recall = TP / (TP+FN)

Precision = TP / (TP+FP)

F1 score = 1 / [(1/ Recall + 1/Precision)/2]

where N is the number of observation and thus is equal to the sum of all the entries of the contingency matrix (footnote: a detailed explanation of what a contingency matrix is and all those performance indices can be found in Wikipedia).

So far, so good. However, an experienced analyst would notice that none of the above indices properly summarise either all the information contained in the contingency matrix or equivalently the properties of the underlying binary classifier, as I am going to explain.

**Taking into account Bias and Prevalence**

In order to properly assess the performance of a classifier using the above indices it is necessary to define two new variables, namely the **Bias** (of the classifier) and the **Prevalence** (of the class of positives in the data):

Bias = (TP + FP) / N

Prevalence = (TP + FN) / N

The Bias measures the percentage of positives predicted by the classifier on the total number of predictions, while the Prevalence measures the percentage of positives on the total number of observations. In practice, the former is a measure of the bias of a classifier, while the latter is a measure of the bias in the data. Intuitively:

1. increasing the Bias can lead to: an increase in the Recall, e.g. keeping FP’s constant the number of TP’s increases with respect to number of FN’s when the Bias increases; a decrease in the Precision, e.g. keeping TP’s constant the number of FP’s increase when the Bias increases;
2. increasing the Prevalence can lead to: an increase in the Precision, e.g. keeping FN’s constant the number of TP’s increases with respect to number of FP’s when the Prevalence increases; a decrease in the Recall, e.g. keeping TP’s constant the number of FN’s increase when the Bias increases;

Furthermore, an high or low Bias can lead to an artificial increase of the Accuracy of a classifiers when the Prevalence is respectively low and high. Finally, the F1 score inherits the issues of Recall and Precision being their harmonic average.

Now we describe how to rigorously formalise the above intuitions introducing three new indices: **Markedness**, **Informedness** and **Matthews correlation**.

**Correcting for Bias and Prevalence**

Here we closely follow a paper from Powers (2007), which is full of good intuitions but lacks mathematical rigour, as I am going to shortly explain. Powers defines

Informedness = TP/(TP + FN) - FP/(FP + TN)

Markedness = TP/(TP + FP) - FN/(FN + TN)

and claims that they represent the corrected version of Recall and Precision respectively. Indeed, they can be equivalently expressed as

Informedness = Recall + Inverse Recall - 1

Markedness = Precision + Inverse Precision - 1

where here for convenience Inverse Recall and Inverse Precision are Recall and Precision for the same classifier but swapping the classes. So why the paper is not mathematically rigorous? The reason is because Powers defines Informedness and Markedness as “probabilities”, which cannot be possibly true since they can take value in the range [-1,1] in the binary case. I guess this lack of rigour is the one the reasons behind the lack of support obtained in the academic community by the **Matthews correlation**, previously introduced by Matthews ():

Matthews correlation = TP x TN – FP x FN / sqrt([bla bla bla])

which is connected to the geometric mean of Informedness and Markedness through the formula:

Matthews correlation = s x sqrt(Informedness x Markedness)

where s = sign(Informedness) = sign(Markedness).

In conclusion, we can see that Markedness, Informedness and Matthews correlation contain information about the contingency matrix in a non biased way with respect to Recall, Precision and Accuracy, see the next section for a more technical justification of this statement.

**Technicalities** (this section can be skipped at first reading)

Now we are equipped with three new performance indices which automatically accounts for Bias and Prevalence in Recall and Precision. Indeed, as shown in the Powers’ paper the following equalities hold:

Informedness = (Recall - Bias) / (1 - Prevalence)

Markedness = (Precision - Prevalence) / (1 - Bias)

Thanks to these equations we can now formalise the intuition about the effect of Bias and Prevalence on Recall, Precision and Accuracy, since

Recall = Informedness x (1 - Prevalence) + Bias

Precision = Markedness x (1 - Bias) + Prevalence

Accuracy = 2 x [Informedness x (1 - Prevalence) + Bias] x Prevalence – Bias – Prevalence + 1

Now it is easy to see that for a given level of Informedness (Markedness), inflating the Bias (Prevalence) inflates Recall (Precision) and deflates Precision (Recall).

For the Accuracy the we can see that for a given level of Informedness:

* Inflating the Bias can lead to:
  + Inflated Accuracy when Prevalence is higher than 0.5
  + Deflated Accuracy when Prevalence is lower than 0.5
  + No effect on Accuracy when Prevalence = 0.5
* Inflating the Prevalence deflates the Accuracy

And what about the F1 score? In the paper by Powers there is no explicit mention about how to fix it, since Matthews correlation is meant as a substitute of both Accuracy and F1 score. Anyway a possible solution is to correct the F1 score by taking the signed harmonic average of Informedness and Markedness.

Another interesting fact about the Matthews correlation is its relationship with the **Pearson correlation coefficient** (wiki ref). Indeed, it is easy to show that for a **1-0 binary problem** (positive class takes numeric value 1 and negative class takes numeric value 0 or viceversa for symmetry of Pearson correlation coefficient) the Matthews correlation can be computed as the Pearson correlation between the vector of actual values and the vector of predicted values.

**Conclusions**

…

**Python code**

….