POINTS OF SIGNIFICANCE

# Classification evaluation

It is important to understand both what a classification metric expresses and what it hides.

Last month we examined the use of logistic regression for classification, in which the class of a data point is predicted given training data[1]. This month, we look at how to evaluate classifier performance on a test set—data that were not used for training and for which the true classification is known. Classifiers are commonly evaluated using either a numeric metric, such as accuracy, or a graphical representation of performance, such as a receiver operating characteristic (ROC) curve. We will examine some common classifier metrics and discuss the pitfalls of relying on a single metric.

Metrics help us understand how a classifier performs; many are available, some with numerous tunable parameters. Understanding metrics is also critical for evaluating reports by others—if a study presents a single metric, one might question the performance of the classifier when evaluated using other metrics. To illustrate the process of choosing a metric, we will simulate a hypothetical diagnostic test. This test classifies a patient as having or not having a deadly disease on the basis of multiple clinical factors. In evaluating the classifier, we consider only the results of the test; neither the underlying mechanism of classification nor the underlying clinical factors are relevant.

Classification metrics are calculated from true positives (TPs), false positives (FPs), false negatives (FNs) and true negatives (TNs), all of which are tabulated in the so-called confusion matrix (**Fig. 1**). The relevance of each of these four quantities will depend on the purpose of the classifier and motivate the choice of metric. For a medical test that determines whether patients receive a treatment that is cheap, safe and effective, FPs would not be as important as FNs, which would represent patients who might suffer without adequate treatment. In contrast, if the treatment were an experimental drug, then a very conservative test with few FPs would be required to avoid testing the drug on unaffected individuals.

In **Figure 2** we show three classification scenarios for four different metrics: accuracy, sensitivity, precision and $F_1$. In each panel, all of the scenarios have the same value (0.8) of a given metric. Accuracy is the fraction of predictions that are true. Although this metric is easy to interpret, high accuracy does not necessarily characterize a good classifier. For instance, it tells us nothing about whether FNs or FPs are more common (**Fig. 2a**). If the disease is rare, predicting that all the subjects will be negative offers high accuracy but is not useful for diagnosis. A useful measure for understanding FNs is sensitivity (also called recall or the true positive rate), which is the proportion of known positives that are predicted correctly. However, neither TNs nor FPs affect this metric, and a classifier that simply predicts that all data points are positive has high sensitivity (**Fig. 2b**). Specificity, which measures the fraction of actual negatives that are correctly predicted, suffers from a similar weakness: not accounting for FNs or TPs. Both TPs and FPs are captured by precision (also called the positive predictive value), which is the proportion of predicted positives that are correct. However, precision captures neither TNs nor FNs (**Fig. 2c**). A very conservative test that predicts only one subject will have the disease—the case that is most certain—has a perfect precision score, even though it misses any other affected subjects with a less certain diagnosis.

Ideally a medical test should have very low numbers of both FNs and FPs. Individuals who do not have the disease should not be given unnecessary treatment or be burdened with the stress of a positive result, and those who do have the disease should not be given false optimism about being disease free. Several aggregate metrics have been proposed for classification evaluation that more completely summarize the confusion matrix. The most popular is the $F_\beta$ score, which uses the parameter $\beta$ to control the balance of recall and precision and is defined as $F_\beta = (1 + \beta^2)(\text{Precision} \times \text{Recall})/(\beta^2 \times \text{Precision} + \text{Recall})$. As $\beta$ decreases, precision is given greater weight. With $\beta = 1$, we have the commonly used $F_1$ score, which balances recall and precision equally and reduces to the simpler equation $2\text{TP}/(2\text{TP} + \text{FP} + \text{FN})$.

The $F_\beta$ score does not capture the full confusion matrix because it is based on the recall and precision, neither of which uses TNs, which might be important for tests of very prevalent diseases. One approach that can capture all the data in the confusion matrix is the Matthews correlation coefficient (MCC), which ranges from −1 (when the classification is always wrong) to 0 (when it is no better than random) to 1 (when it is always correct). It should be noted that in a comparison of the results of two classifiers, one

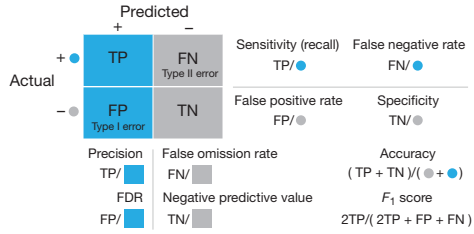


**Figure 1** | The confusion matrix shows the counts of true and false predictions obtained with known data. Blue and gray circles indicate cases known to be positive (TP + FN) and negative (FP + TN), respectively, and blue and gray backgrounds/squares depict cases predicted as positive (TP + FP) and negative (FN + TN), respectively. Equations for calculating each metric are encoded graphically in terms of the quantities in the confusion matrix. FDR, false discovery rate.

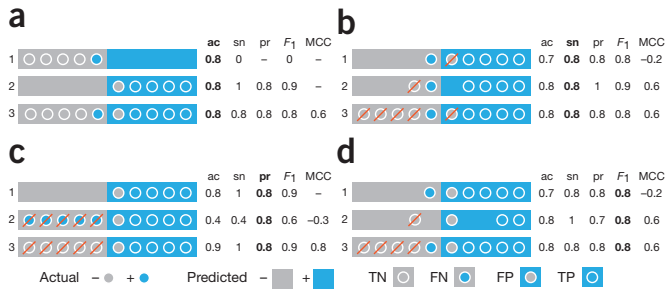


**Figure 2** | The same value of a metric can correspond to very different classifier performance. (**a**–**d**) Each panel shows three different classification scenarios with a table of corresponding values of accuracy (ac), sensitivity (sn), precision (pr), $F_1$ score ($F_1$) and Matthews correlation coefficient (MCC). Scenarios in a group have the same value (0.8) for the metric in bold in each table: (**a**) accuracy, (**b**) sensitivity (recall), (**c**) precision and (**d**) $F_1$ score. In each panel, those observations that do not contribute to the corresponding metric are struck through with a red line. The color-coding is the same as in **Figure 1**; for example, blue circles (cases known to be positive) on a gray background (predicted to be negative) are FNs.

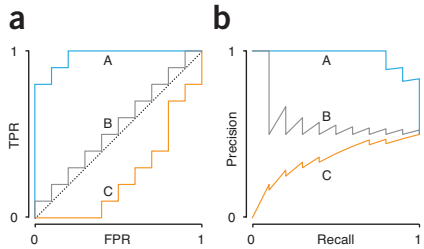


**Figure 3** | Graphical evaluation of classifiers. (**a**,**b**) Findings obtained with the (**a**) ROC, which plots the true positive rate (TPR) versus the false positive rate (FPR), and (**b**) PR curves. In both panels, curves depict classifiers that are (A) good, (B) similar to random classification and (C) worse than random. The expected performance of a random classifier is shown by the dotted line in **a**. The equivalent for the PR curve depends on the class balance and is not shown.

may have a higher $F_1$ score while the other has a higher MCC. No single metric can distinguish all the strengths and weaknesses of a classifier.

An important factor in interpreting classification results is class balance, which is the prevalence of a disease in the general population. Imbalance makes understanding FPs and FNs more important. For a rare disease affecting only 2 in 1,000 people, each FP has a much larger effect on the proportion of misdiagnoses than it would for a more prevalent disease that affects 200 in 1,000 people. We shall assume that the prevalence of the disease in the general population is reflected in the training and test data. If this is not the case, extra care is required to interpret the results.

Imagine a diagnostic test for a disease that gives a numeric score for a person having the disease. Instead of a simple positive or negative result, the score gives a level of certainty: individuals with a higher score are more likely to have the disease. In fact, almost all classifiers generate positive or negative predictions by applying a threshold to a score. As we discussed last month, a higher threshold will reduce the FP rate (in our example, this represents healthy individuals given unnecessary treatment), and a lower threshold will reduce the FN rate (diseased individuals who do not get treatment).

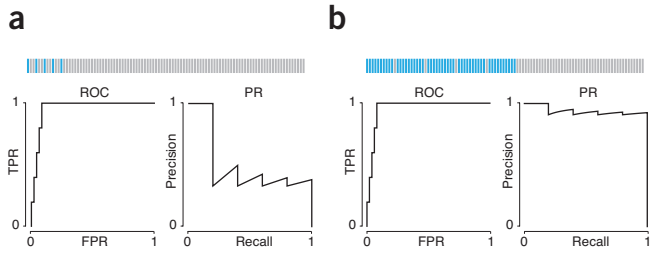


**Figure 4** | Graphical representation of classifier performance avoids setting an exact threshold on results but may be insensitive to important aspects of the data. (**a**,**b**) ROC and PR curves for two data sets with very different class balances: (**a**) 5% positive and (**b**) 50% positive observations. For each panel, observations are shown as vertical lines (top), of which 5% or 50% are positive (blue).

One might wish to evaluate the classifier without having to select a specific threshold. For this, consider a list of individuals with known disease status ordered by decreasing diagnostic score. This list can be visualized using the ROC curve (**Fig. 3a**). When creating an ROC curve, we start at the bottom left corner and at the top of our list of prediction scores. As we move down the list, if the data are known to be positive (an individual with the disease), the line moves up; otherwise it moves to the right. A good classifier should aim to reach as close to the top left corner as possible. An alternative visualization is the precision–recall (PR) curve (**Fig. 3b**). Its interpretation is slightly different, as the best classifier would be as close to the top right as possible, gaining the best trade-off of recall and precision. Unlike the ROC curve, the PR curve is not monotonic.

Class imbalance can cause ROC curves to be poor visualizations of classifier performance. For instance, if only 5 out of 100 individuals have the disease, then we would expect the five positive cases to have scores close to the top of our list. If our classifier generates scores that rank these 5 cases as uniformly distributed in the top 15, the ROC graph will look good (**Fig. 4a**). However, if we had used a threshold such that the top 15 were predicted to be true, 10 of them would be FPs, which is not reflected in the ROC curve. This poor performance is reflected in the PR curve, however. Compare this to a situation with 50 diseased individuals out of 100. A classifier that gives an equivalent ROC curve (**Fig. 4b**) will now have a favorable PR curve. For these reasons, PR curves are recommended for data sets with large class imbalances. Summary metrics of these two graphs are also used: the area under the curve (AUC) for the ROC curve and the area under the PR curve (AUPRC). Both of these metrics suffer from the same limitations as any other single metric.

Understanding the intended use of a classifier is the key to selecting appropriate metrics for evaluation. Using one metric—even an aggregate one like the $F_1$ score—is dangerous without proper inspection of the underlying results. Additionally, one should always be on the lookout for class imbalance, which is a confounding factor that can distort various metrics.

**COMPETING FINANCIAL INTERESTS**
The authors declare no competing financial interests.

**Jake Lever, Martin Krzywinski & Naomi Altman**

1. Lever, J., Krzywinski, M. & Altman, N. *Nat. Methods* **13**, 541–542 (2016).

**Corrected after print 16 September 2016.**

Jake Lever is a PhD candidate at Canada's Michael Smith Genome Sciences Centre. Martin Krzywinski is a staff scientist at Canada's Michael Smith Genome Sciences Centre. Naomi Altman is a Professor of Statistics at The Pennsylvania State University.

# Corrigendum: Classification evaluation

**Jake Lever, Martin Krzywinski & Naomi Altman**
*Nat. Methods* **13, 603–604 (2016); published online 28 July 2016; corrected after print 16 September 2016**

In the version of this article initially published, the expression defining the $F_\beta$ score was incorrect. The correct expression is $F_\beta = (1 + \beta^2)$ (Precision × Recall)/($\beta^2$× Precision + Recall).The error has been corrected in the HTML and PDF versions of the article.