POINTS OF SIGNIFICANCE

Graphical assessment of tests and classifiers

I do not think you can start with anything precise. You have to achieve such precision as you can, as you go along. — Bertrand Russell

Naomi Altman and Martin Krzywinski

To make decisions based on evidence, we need an algorithm to translate evidence into a decision—a decision function that takes data as input and outputs a decision. Appropriate choice of decision functions enhances reproducibility of results in testing, classification and diagnostics. This month we explore in greater depth how we can assess and choose such functions using receiver operating characteristic (ROC) and precision–recall (PR) curves[1].

We have already seen several kinds of decision functions: the *P*-value in statistical testing[2], logistic regression in binary classification[3] and diagnostic functions in screening[4]. Generally, the decision function f is continuous and uses as input such sample attributes as the sample mean, standard deviation and sample size. Its output is compared to a threshold f_0 to make a binary decision about the sample. For example, if f is the *P*-value and $f < f_0$, the decision is positive (for example, we reject the null at the type I error level $\alpha = f_0$) and otherwise the decision is negative. The direction of the inequality depends on the function; in logistic regression, $f > f_0$ results in a positive decision[3]. The function together with the threshold create a decision rule.

To assess the efficacy of the decision rule, we assume that there is a true state and that a perfect decision rule will always produce the correct decision for any sample. Practically, however, any rule will be imperfect and, when applied to a set of samples, will result in a mix of correct and incorrect decisions that can be tallied using a confusion matrix[1] (Table 1).

The confusion matrix can be used to define many different measures of goodness (for example, precision, accuracy, F_1 score[1]), and it's not entirely straightforward which of these to use to create a decision rule. While we always prefer decision rules that yield fewer errors (false positives and false negatives), the number of errors also depends on the proportion of positive (m^+) and negative (m^-) samples[4]. These considerations form the basis for ROC and PR curves.

Table 1 | A confusion matrix of positive (+) and negative (–) decision function outputs and truth state

Truth	$f(x) < f_0$	$f(x) \geq f_0$	Total
	+	-	
+	TP	FN	m^+
-	FP	TN	m^-
Total	n^+	n^-	

Rows correspond to positive and negative classes of samples with m^+ and m^- samples that are truly positive or negative, respectively. Columns correspond to positive and negative classes of decisions with n^+ and n^- decisions that are positive and negative, respectively. Individual decisions are classified as true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN). For some decision functions, the direction of the inequality is reversed.

Let's explore these curves using a simple decision function that predicts whether a cell is in a normal or diseased state using the levels of two enzymes, A and B. The level of enzyme A is normally distributed with mean 1 and s.d. 0.2 in normal cells and elevated to a mean of 1.2 in diseased cells (Fig. 1a). The level of enzyme B in is exponentially distributed mean = s.d. = 1 in normal cells and depressed to mean = s.d. = 0.29 in diseased cells. We will simulate 100 cells whose disease state is known (Fig. 1b) to train a logistic regression classifier[3] as our decision function and test it using a sample of 10,000 cells. We'll start with the scenario in which the classifier is sensitive only to levels of enzyme A and in which the normal and diseased classes are balanced ($m^+/m = 50\%$).

The ROC curve can be thought as a plot of success versus failure. Its y axis shows the true positive rate TPR = TP/m^+, which is the fraction of successful decisions in the class of diseased cells (that is, power). Its x axis is the false positive rate FPR = FP/m^-, which is the fraction of wrong decisions in the class of normal cells (that is, type I error). Given a decision function, the curve is constructed by plotting these quantities for each possible threshold value. For logistic regression, this range is $0 \leq f_0 \leq 1$, with larger values corresponding to a more conservative classifier (one that makes fewer positive decisions).

Let's walk along the ROC curve for our scenario (Fig. 2a). When $f_0 = 1$, the classifier never makes a positive decision: there are no false positives (no errors in the normal class, FPR = 0) and there are no true positives (no successes in the diseased class, TPR = 0). At the upper right end of the curve ($f_0 = 0$) the classifier never makes a negative decision: there are no true negatives (maximum errors in the normal, FPR = 1) and there are no false negatives (maximum successes in the diseased class, TPR = 1). At these threshold values, the performance of the classifier is neither useful nor interesting—what we care about is the shape of the curve between these points.

For example, at $f_0 = 0.61$, our classifier has the power to detect half of the diseased cells (TPR = 50%). Achieving this (or any nonzero) power comes at a cost of false negatives: we misclassify about FPR = 17% of normal cells as diseased. Our successes in the diseased class outnumber failures in the normal class by 3:1. We could achieve a higher power (90%) at a lower $f_0 = 0.34$, but now would need to accept an FPR = 62% and a success-to-failure ratio of 3:2.

If our classifier were perfect, it would correctly classify every diseased cell (TPR = 100%) and normal cell (FPR = 0%) (Fig. 2a, black point). If our classifier were random—one that makes a positive decision with probability f_0—it would have TPR = FPR and a ROC curve that is diagonal line (Fig. 2a, dashed line). If, for example, $f_0 = 0.5$ of our decisions are randomly positive (unbiased coin flip), we would incorrectly classify 50% of normal

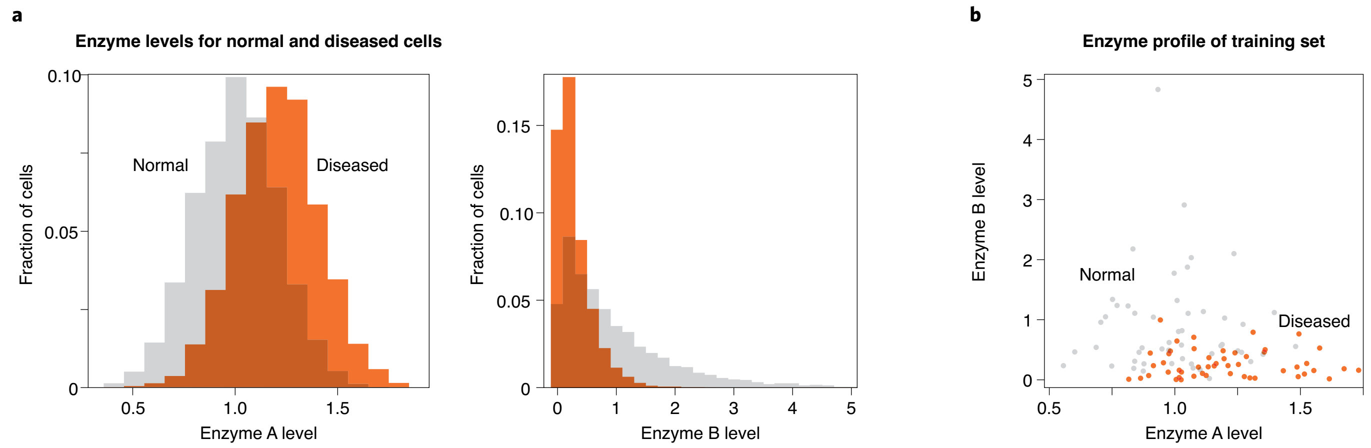

Fig. 1 | Enzyme level profiles for healthy and diseased cells. a, The level of enzyme A is normally distributed with means of 1.0 and 1.2 for normal and diseased cells, respectively, and s.d. of 0.2. The level of enzyme B is exponentially distributed with means of 1 and 0.29 in normal and diseased cells, respectively. The s.d. of the exponential distribution is the same as its mean. **b**, The enzyme profile of the training set of 50 normal and 50 diseased cells.

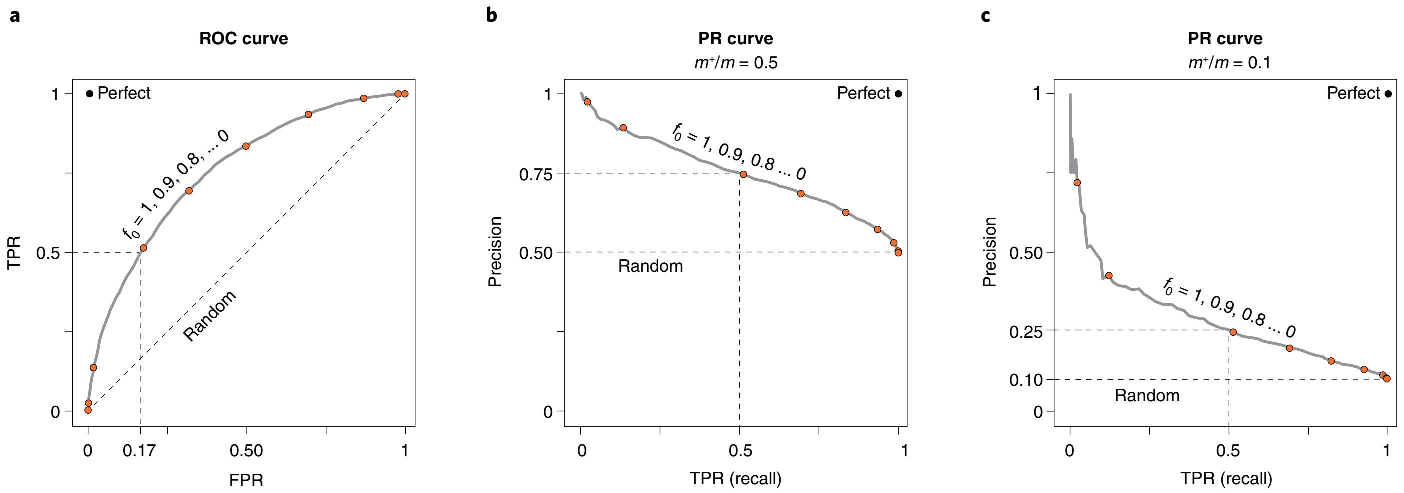

Fig. 2 | The ROC and PR curves for a logistic regression classifier of normal and diseased state using enzyme A level. a, The ROC curve shows the classifier's TPR as a function of FPR at each value of threshold f_0. Orange points correspond to steps in f_0 of 0.1. The classifier achieves a TPR (power) of 0.5 at $f_0 = 0.61$, where FPR = 0.17. For a perfect classifier, the curve goes through (FPR, TPR) = (0,1). For a random classifier, the curve is a diagonal line (TPR = FPR). **b**, The PR curve of the classifier for balanced normal and diseased classes ($m^+/m = 0.5$). For a TPR = 0.5, the classifier achieves a precision of 0.75. For a perfect classifier, the curve goes through (1,1). For a random classifier, the curve is a horizontal line at m^+/m. **c**, The PR curve for a highly imbalanced classes, where only m^+/m = 10% of the cells are diseased. For a TPR = 0.5, the classifier now achieves a precision of only 0.25.

cells (FPR = 50%) but also correctly classify 50% of the diseased cells (TPR = 50%). Any ROC curves below this diagonal line correspond to classifiers that are worse than random. Such classifiers (including the edge case of the classifier that is always wrong) can easily be made to be better than random by negating their decision.

Generally, we seek a ROC curve that comes closer to the perfect classifier (that is, further from the random classifier). Such a curve gives us a good chance of finding an f_0 with high TPR and low FPR. Remember that a point on a ROC curve corresponds to a decision rule (decision function and specific threshold) whereas a ROC curve corresponds to a decision function. Note also that the ROC curve is monotonically increasing with respect to both FPR and TPR because as we traverse the threshold range, false positives and true positives always increase (that is, we always make more positive decisions when we decrease the logistic regression threshold).

The ROC curve, however, does not tell the full story: neither FPR nor TPR depend on m^+/m because both quantities are relative to sums of rows in the confusion matrix. Our classifier will have the same ROC curve regardless whether it is applied to a common or rare disease. But we know that when m^+/m is small, even good classifiers can have unacceptably low precision (TP/(TP + FN)) because decision errors in the normal and larger class (FN) can greatly outnumber decision successes (TP). This is something that the ROC curve cannot tell us, and, if we suspect a class imbalance, it's useful to plot precision as a function of TPR. Because TPR is also known as recall, such a curve is called the precision–recall curve; it is a plot of success in the class of positive decisions as a function of success in the class of positive samples. Although it would be easier to compare the ROC and PR curves if both had

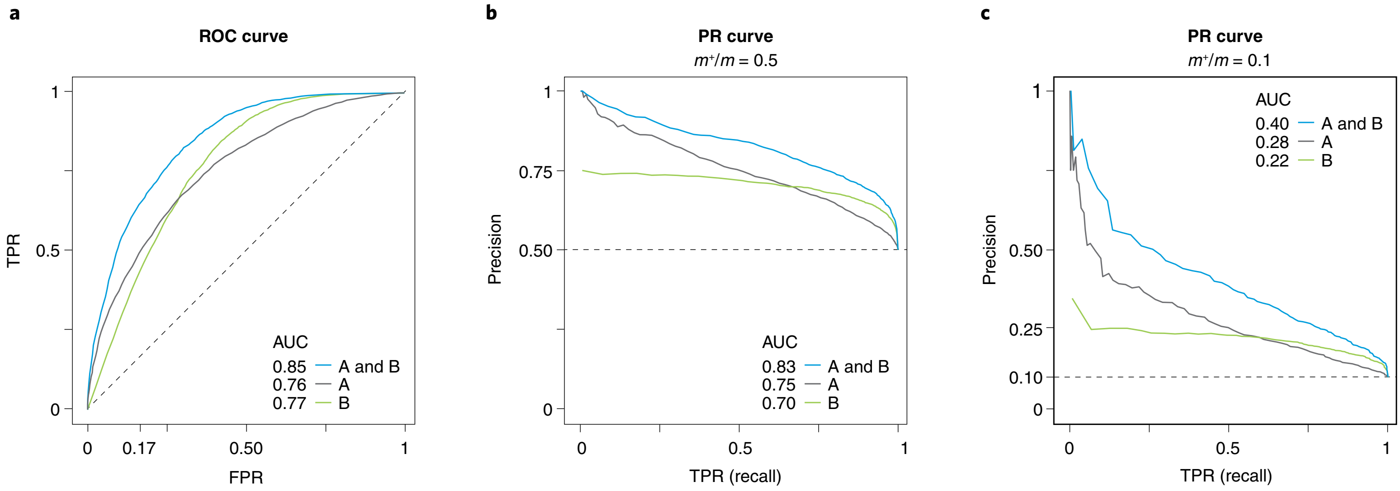

Fig. 3 | The ROC and PR curves for a logistic regression classifier of normal and diseased state using one or both enzyme levels. a, The classifier trained on both A and B levels (blue) performs better than one using enzyme A or B alone, as illustrated by a ROC curve that is above that of the other two scenarios and has the highest AUC. The classifier trained on enzyme A performs better than that trained on B for FPR < 0.25 but has a slightly lower AUC. **b**,**c**, The PR curves for classifiers in **a** for common ($m^+/m = 0.5$) and rare ($m^+/m = 0.1$) disease states.

TPR on the vertical axes, it is customary to have TPR on the horizontal axis in the PR curve.

Unlike those of the ROC curve, the endpoints of the PR curve depend on m+/m. When the classifier always makes a positive decision, there are no false negatives (TP = m^+) and no true negatives (FP = m^-) and precision reaches its minimum value of m^+/m. The other end of the curve will terminate at the most conservative threshold value for which TPR is minimum but precision is still defined (at least one positive decision is made). The curve does not include the point (TPR, precision) = (0,1) because it is an impossible combination.

When classes are balanced, the PR curve tells us that our classifier reaches a precision of 75% at a power of 50% (Fig. 2b). When the disease state is rare ($m^+/m = 0.1$), the PR curve drops substantially (Fig. 2c) and now we only have 25% precision at a power of 50%.

ROC and PR curves can be used to compare different decision functions because they are constructed from the confusion matrix and do not directly depend on either the value or the scale of the decision function. We illustrate how these curves can vary by comparing the performance (Fig. 3) of our classifier based on levels of enzyme A (black curve) to one trained on enzyme B (green curve) and on both enzymes (blue curve). In the case where ROC (or PR) curves for two decision functions cross and neither reaches a desired combination of FPR and TPR (or TPR and precision), the one with a larger area under the curve (AUC) would be selected. For example, the ROC curve for the classifier trained on enzyme A (AUC = 0.76) crosses that of the classifier trained on enzyme B (AUC = 0.77) and the latter has a marginally higher AUC (Fig. 3a). Performance spread is more apparent in their PR curves, where the AUC of the classifier trained on enzyme A is substantially higher (0.75 versus 0.70). Proportionately, this difference is larger when the disease state is rare, where we see AUC 0.28 versus 0.22 (Fig. 3c).

The ROC curve is useful when TP and TN are equally likely and the PR curve when classes are unbalanced. Both are handy graphical aids for comparing decision functions and choosing thresholds, and it's helpful to view their axes as showing successes and failures within a given class. ❐

Naomi Altman[1] and Martin Krzywinski[2] ✉
[1]*Department of Statistics, The Pennsylvania State University, State College, PA, USA.* [2]*Canada's Michael Smith Genome Sciences Centre, Vancouver, British Columbia, Canada.*
✉*e-mail: martink@bcgsc.ca*

Published online: 27 July 2021
https://doi.org/10.1038/s41592-021-01232-1

References

1. Lever, J., Krzywinski, M. & Altman, N. *Nat. Methods* **13**, 603–604 (2016).
2. Krzywinski, M. & Altman, N. *Nat. Methods* **11**, 215–216 (2014).
3. Lever, J., Krzywinski, M. & Altman, N. *Nat. Methods* **13**, 541–542 (2016).
4. Altman, N. & Krzywinski, M. *Nat. Methods* **18**, 224–225 (2021).

Competing interests

The authors declare no competing interests.