2. Measures for 2×2 confusion matrix

In classification problems with 2 classes (positive/negative), we can tally the observed (true) and predicted classes into a 2×2 table, also called a “confusion matrix”. (See also `https://en.wikipedia.org/wiki/Sensitivity_and_specificity`, where the confusion matrix is shown with rows for predicted classes and columns for true classes.)

		Predicted		
		Positive	Negative	Total
True	Positive	TP	FN	trueP
	Negative	FP	TN	trueN
	Total	predP	predN	n

Let P and N denote the true class, and P^* and N^* the predicted class. We can define some simple measures of performance as conditional probabilities:

- $\Pr(P^*|P) =$ TP/trueP : TPR, sensitivity, power, recall, 1−FNR, 1−Type II error rate
- $\Pr(N^*|N) =$ TN/trueN : TNR, specificity, 1−FPR, 1−Type I error rate
- $\Pr(P|P^*) =$ TP/predP : PPV, precision, 1−FDR
- $\Pr(N|N^*) =$ TN/predN : NPV

Some other simple measures of performance are functions of one of these four measures. For example, false positive rate is FPR=1−TNR; false discovery rate is FDR=1−PPV.

Some of these measures may not be meaningfully estimated if the data are not a random sample from the population. For example, in a case-control study, trueP and trueN are predetermined and trueP/n is often much higher than prevalence $\pi = \Pr(P)$. In such a study, PPV and NPV cannot be estimated as above. (They can be estimated if prevalence π is known: $PPV = \frac{\pi TRP}{\pi TPR + (1-\pi)FPR} = \frac{r}{r+1}$, where $r = \frac{\pi}{1-\pi}\frac{TPR}{FPR}$; $NPV = \frac{(1-\pi)TNR}{\pi FNR + (1-\pi)TNR} = \frac{s}{s+1}$, where $s = \frac{1-\pi}{\pi}\frac{TNR}{FNR}$.)

Ideally we would like all these measures to be high. Unfortunately, a high value in one measure does not guarantee a high value in any of the other three. We use sensitivity as an example. (1) Sensitivity and specificity often move on opposite directions; that is, a high $\Pr(P^*|P)$ may correspond to a low $\Pr(N^*|N)$. This can be shown in an ROC curve. (2) Sensitivity (recall) and PPV (precision) often move on opposite directions; that is, a high $\Pr(P^*|P)$ may correspond to a low $\Pr(P|P^*)$. This is called precision–recall tradeoff. (3) A high $\Pr(P^*|P)$ does not imply a high $\Pr(N|N^*)$. This is surprising because in pure logic, if A implies B, then not B must imply not A. Unfortunately, in probabilistic reasoning, there is no guarantee that if $\Pr(B|A)$ is high then $\Pr(notA|notB)$ is high. Thus, it is wrong to claim that a highly sensitive test is deemed effective at ruling out a disease when negative. Similarly, a high $\Pr(N^*|N)$ does not guarantee a high $\Pr(P|P^*)$, and thus it is wrong to claim that a highly specific test is effective at ruling in a disease when positive.

Prevalence has a large impact on PPV/NPV. A high prevalence often leads to a high PPV and a low prevalence leads to a high NPV. For example, if the positive class has a low prevalence $\pi = 0.01$, even a test with a high sensitivity TPR=0.95 and a high specificity TNR=0.95 has a low PPV=0.16 ($r = \frac{\pi}{1-\pi}\frac{TPR}{FPR} = \frac{0.01}{0.99}\frac{0.95}{0.05} = 0.19$ and $PPV = \frac{r}{r+1} = 0.16$); if the positive class has a high prevalence $\pi = 0.9$, a test with TPR=0.95 and TNR=0.95 has NPV=0.68 ($s = \frac{1-\pi}{\pi}\frac{TNR}{FNR} = \frac{0.1}{0.9}\frac{0.95}{0.05} = 2.11$, $NPV = \frac{s}{s+1} = 0.68$).

To balance some of these measures, composite measures of performance are defined:

- F-score: $F_1 = 2pr/(p + r)$, the harmonic average of precision and recall. A harmonic average is close to the smaller value. Its extension, $F_\beta = (1 + \beta^2)pr/(\beta^2 p + r)$, is the harmonic average of one portion of precision and β^2 portions of recall.

- G-score (Fowlkes–Mallows index): $G = \sqrt{pr}$, the geometric average of precision and recall. When $p \neq r$, $F_1 < G$.

- Youden's J (informedness): $J = \mathit{sens} + \mathit{spec} - 1$, which is $\mathit{sens} - (1 - \mathit{spec}) = y - x$ in the ROC curve. So a higher J corresponds to a larger distance from the diagonal line $y = x$ in the ROC curve. It is shown above that a classifier with a high sensitivity and a high specificity may still have a low PPV (e.g., when prevalence π is low).

- Markedness: $PPV + NPV - 1$. Similar to J, when prediction rate $\Pr(P^*)$ is low, a classifier with a high PPV and a high NPV may have a low sensitivity.

In some applications, there may not be a natural choice of positive and negative classes. For example, suppose we use demographic and socioeconomic variables to predict the party, either Republican or Democrat, a person will vote for in 2020. Suppose we designate voting Republican as being positive (scenario 1), and calculate various measures of performance as above. If we switch the designation and label voting Democrat as positive (scenario 2), sensitivity in scenario 1 becomes specificity in scenario 2 and specificity in scenario 1 becomes sensitivity in scenario 2. Similarly PPV and NPV are also swapped. In this situation, measures F_1 and G may not make sense because they are designed to focus on the performance over the positive class.

There are also measures of agreement between the observed and predicted classes. These measures are not defined through the conditional probabilities, but directly with the 2×2 table. They reflect how concentrated the counts are on the diagonal:

- Phi coefficient = Matthews correlation coefficient = $\sqrt{\chi^2/n}$ = Pearson's CC.

- Kappa, $\kappa = (c_o - c_e)/(n - c_e)$, where c_o = TP+TN and c_e is the expected count on the diagonal under independence between the observed and the predicted.

- Accuracy (TP+TN)$/n$ = 1−misclassification rate. It is the fraction of agreement. Under severe class imbalance, a high accuracy can be easily achieved with a classifier that assigns everything to the majority class. Accuracy $\geq \kappa$, where equality holds only when TP+TN $= n$.

3. ROC Curve

An ROC curve is a **profile** for a set $\{(a_i, b_i); i = 1, \ldots, n\}$, where a_i is binary and b_i is quantitative. For example, an ROC curve is often drawn for logistic regression, with a being the observed outcome and b the estimated probability. To draw an ROC curve, we first sort b, and then for every classification model resulting from dichotomizing b ($b \leq t$ and $b > t$ for any given t), we calculate the correponding sensitivity and specificity and plot sensitivity against $1-$ specificity (equivalently, power against type I error rate, or TPR against FPR). For n observations, there are at most $n-1$ distinct dichotomizations and thus at most $n-1$ points for the ROC curve. In addition, there are 2 extreme classifiers, all N and all P, which correspond to $(0, 0)$ and $(1, 1)$ on the ROC curve.

The AUC of an ROC curve is a summary measure of the relationship between a and b. It measures the level of concondance between a and b, because it is the same as the C-index (concordance index), which is defined as the fraction of randomly selected P–N pairs for which the direction of their b values is consistent with that of their a values (a tie is counted as a half). The ROC AUC is also related to Wilcoxon–Mann–Whitney test. When b is binary, the ROC curve only has three points, $(a_0, b_0) =$ (FP/trueN, TP/trueP), $(0, 0)$ and $(1, 1)$; its AUC is $\frac{1}{2}a_0b_0 + \frac{1}{2}(b_0 + 1)(1 - a_0) = \frac{1}{2}(b_0 - a_0 + 1)$, which is half of Youden's J; visually, AUC is monotonic to the distance of (a_0, b_0) from the line $y = x$, or equivalently, it is monotonic to $b_0 - a_0$.

ROC curves (and ROC AUCs) are useful for methods (such as logistic regression) that do not generate a classification model per se, but instead a quantitative measure (a score) per observation. `http://stats.stackexchange.com/questions/145566/` gives a good introduction of ROC curves and ROC AUCs.