Pre-read: measuring predictability

COS 597E/SOC 555, Princeton University, Fall 2020

We've encountered many ways to measure predictive accuracy. Our goal for Thursday will be to catalog these metrics, identify desirable properties of metrics, determine which metrics satisfy which properties, and use this to develop an intuition for how to pick a metric based on the problem and domain.

The first reading is on ROC analysis. The AUC-ROC metric (AUC for short) is of particular interest to us, but ROC analysis is broader and more useful than AUC or any other scalar metric can be.

The second reading explains the logarithmic scoring rule (log loss). It also shows that the logarithmic scoring rule belongs to an important class of metrics called "proper scoring rules" (although the reading doesn't use this term). Roughly speaking, a scoring rule for probabilistic classifiers is said to be *proper* if a classifier that "knows" the "true" probabilities of the outcomes optimizes its expected score by faithfully reporting the true probabilities. Note that the "naive strategy" in Section 1 is a variant of Mean Absolute Error, so this tells us that MAE is not a proper scoring rule.

The third and fourth readings are about R^2. The third reading describes nine (!) variants of R^2, their nuances, and pitfalls of application. The fourth reading is a one-sided but entertaining analysis of some of the limitations of R^2 (which R^2?).

As you do the readings, observe how the authors advocating for different metrics are motivated by different kinds of applications of prediction. AUC (and ROC analysis) is a good choice when the goal of prediction is to enable decision making (e.g. a bank might want to know how likely it is that a loan applicant will default on the loan if the application is approved). Which properties of AUC make it useful for this setting?

On the other hand, the second paper is motivated by the problem of belief elicitation: how to design incentives for (usually human) forecasters, such as Tetlock's experts, such that it is in their interest to divulge their true beliefs. Meanwhile R^2 comes from the data modeling culture, where the goal of prediction is to generate understanding (often the predictions themselves are not of interest).

Comparisons across these families of metrics are rare in the literature. For example, there are dozens of papers comparing proper scoring rules, but they never include (say) AUC. However, our course is all about intermingling of statistical subcultures and applying ideas outside the set of problems where they are usually applied. Understanding the properties of each metric at a granular level is the first step toward this.

Exercise. Identify the measures of predictive accuracy used in each of the readings so far (if you don't have time, pick a random subset of the readings). In each case, what justification did the authors provide for picking that measure? What other justifications, if any, can you think of? The breakout activity in class will build on the understanding you gain from this exercise.