Abstract

Forecasters should not be tested just by their calibration score (the average distance between each forecast and the average of its realizations), which can always be made arbitrarily small, but rather by their Brier score, which is the sum of the calibration and the refinement scores. The refinement score (the average variance of the realizations for each forecast) measures how good is the classification into bins with the same forecast, and thus attests to true expertise. We introduce the notion of "calibeating" a forecasting procedure b by an online procedure c: the Brier score of c is guaranteed to be, in the long run, lower than the refinement score of b. In other words, c calibeats b if c beats the Brier score of b by the calibration score of b. We show an easy way to calibeat, moreover by procedures that are calibrated (and thus cannot be calibeaten themselves), and extend this to simultaneous calibeating of mutiple procedures and to deterministic continuous calibration.