[MINITUTORIAL]

The Brier score and its decomposition

Stephan Kolassa

Sometimes we will forecast not a numerical value, but a class membership. How likely is it that someone will default on their credit, or have an accident… or that a drug will make it to the market? This exercise is one of *classification*, and probabilistic predictions are typically more useful than hard 0-1 predictions (<https://stats.stackexchange.com/a/312787/1352>). How do we assess the quality of such a probabilistic prediction?

The tool of choice is a *proper scoring rule*, which is simply a mapping that takes a probabilistic classification and the corresponding outcome and maps it to a number. We will typically average these scores over many prediction-outcome pairs to assess a model, similar to how we would evaluate a forecasting model using multiple forecast-actual pairs. Note, incidentally, that different proper scoring rules can also be used to assess predictive densities in numerical forecasting.

One very common proper scoring rule (there are others!) is the *Brier score*, which was introduced by Glenn W. Brier (1950) in weather forecasting. Suppose we have probabilistic predictions for events with corresponding outcomes , where if the target event occurred and if it didn’t. Then the Brier score is defined as

Thus, if the event occurred (), we get a contribution of , which is small if is large. And if the event did *not* occur, we have a contribution of , which is small if is small. We are thus aiming for a Brier score that is as small as possible: the score is said to be *negatively oriented*. (There is also the rarer opposite *positively* oriented convention, where we just put a minus sign in front of the sum and aim for *large* scores – it’s good to make sure everyone is using the same orientation.)

As forecasters, we note immediately that this is nothing else than the Mean Squared Error (MSE) of the variable which 0-1 codes the outcome. The key thing is that this score is *proper* in the sense that it incentivizes correct probabilistic predictions, just as the MSE incentivizes unbiased expectation forecasts. (Using the absolute value rather than squares would reward us for extreme predictions that are always 0 or 1., These extreme predictions lose all the nuance of true probabilities, which is why such a variation is not proper any more.)

If the predictions can only take on distinct values, e.g., if we only have predictions of , then we can decompose the Brier score into three components Reliability, Resolution and Uncertainty (<https://stats.stackexchange.com/q/631333/1352>):

where is the number of predictions where takes on the -th possible value (so in the example above, would be the number of predictions where ), is the average actual outcome for these predictions, and is the overall grand average of all outcomes. The components have human-readable interpretations:

* Reliability measures the loss in performance (i.e., in the Brier score) due to getting the true outcome proportions wrong, since it is zero precisely when for all .
* Resolution measures how well the forecast differentiates instances in terms of how far they are away from the population (“climatological”) probabilities. The forecast enters here because it determines which particular a given instance contributes to.
* Uncertainty measures the overall difficulty of the prediction task. The forecast does not enter here at all. Uncertainty is highest when the overall grand average probability is and zero if the event either never () or always occurs ().

This decomposition thus allows us to analyze a given Brier score more deeply and understand where we may be able to improve our probabilistic classification model. Below is an example where we have possible predictions with outcomes (among them 30 successes).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **k** | **1** | **2** | **3** | **4** | **5** |
| **Predicted probability** | 0 | 0.25 | 0.5 | 0.75 | 1 |
| **Successes** | 2 | 3 | 9 | 9 | 7 |
| **Failures** | 6 | 13 | 11 | 3 | 1 |
| **nk** | 8 | 16 | 20 | 12 | 8 |
| **ok** | 0.25 | 0.1875 | 0.45 | 0.75 | 0.875 |

Going through the math, we find a Brier score of 0.199, with a reliability of 0.012, a resolution of 0.061 and an uncertainty of 0.249, which we could compare with a competing forecast. We can also create a *calibration plot*, by plotting the observed bucketized outcomes against the probabilities – if our dots all lay on the dashed diagonal, we would have a perfect prediction:

