Ordered Categorical Scoring Rule

Summary

Multinomial questions are those with more than two answer options. These questions can be divided into two cases: one in which the order of the answer options does not matter (e.g., lists of candidates in an election) and one in which the order of answer options can matter (e.g., dates or values divided into multiple answer bins). Both types of multinomial questions can be scored using the traditional Brier scoring rules. In the second case, however, we may wish to assign partial credit for "near-misses" (i.e., incorrect but close answers). Below we describe the "ordered categorical scoring rule" through which we can assign such partial credit.

1. Examples of the first case, in which the order of the answer options does not matter:
    - Election questions with multiple candidates or parties ("Who will win the presidential election in Argentina?" and "Which party will the next Canadian Prime Minster come from?") or
    - Selection questions ("Which movie will win the Oscar for Best Picture?" and "Which toy will win the Innovative Toy of the Year Award for 2016?")

2. Examples of the second case, in which the order of the answer options can matter, because a range of possible values is divided into bins:
    - Dates divided into bins ("When will Iran next launch a ballistic missile?" with answer options of "Before 1 March 2016", "Between 1 March and 30 April 2016, inclusive" "Between 1 May and 30 June 2016, inclusive," and "Not before 1 July 2016.")
    - Values divided into bins, like exchange rates ("What will the end-of-day closing value for the dollar against the renminbi be on 1 January 2016?" with answer options of "Less than 6.30," "Between 6.30 and 6.35, inclusive," "More than 6.35 but less than 6.40," and "6.40 or more") and votes or seats ("How many seats will the Justice and Development Party win in Turkey's snap elections?" with answer options of "A majority," "A Plurality," and "Not a plurality.")

We could apply the traditional Brier scoring rule to both types of multinomial questions. In the second case, however, we might want to assign some partial credit to forecasters for getting closer to the true outcome.

Consider the generic four-outcome example where option "B" occurs and one forecaster had assigned the following forecasts

A: 0.25 B: 0.25 C: 0.50 D: 0

A second forecaster had assigned the following forecasts:

A: 0.25 B: 0.25 C: 0.30 D: 0.20

The first forecaster would receive a score of 0.875, which comes from

$$(0.25-0)^2+(0.25-1)^2+(0.50-0)^2+(0-0)^2=0.875$$

and the second forecaster would receive a better score of 0.755. One might argue that the first forecaster should receive a better score because she assigned more weight to answer C and less to answer D, and was "closer" to the eventual outcome of B than the second forecaster. The next section describes a method for dealing with "near-misses" and assigning partial credit.

Calculation

We assign partial credit using the ordered categorical scoring rule, which requires several more steps than the traditional Brier score:[1]

1. Take the original answer options and break them up into a set of binary pairs
2. Apply the scoring rule to each pair
3. Take the average across the binary pair scores

This is best illustrated with an example. Consider the first forecaster from the previous section. We would divide the four answer options of A-B-C-D into three binary pairs: A versus BCD, AB versus CD, and ABC versus D.

Using the ordered categorical scoring rule, we would have the following calculation:

- A vs BCD: The sum of forecasts for A is 0.25 and the sum of forecasts for BCD is 0.75. Because answer option B occurred, and outcome for BCD is 1 and the outcome for A is 0. We get the following score for this particular binary pair:

$$(0.25-0)^2+(0.75-1)^2=0.125$$

We can repeat the process for the other binary pairs.

- AB vs CD: $(0.5-1)^2+(0.5-0)^2=0.50$
- ABC vs D: $(1-1)^2+(0-0)^2=0$

Thus, this first forecaster would receive an ordered categorical score of 0.208, which is the average of 0.125, 0.50, and 0. Following the same process, the second forecaster would receive a worse score of 0.235.

[1] Jose, V., Nau, R. & Winkler, R. (2009). Sensitivity to distance and baseline distributions in forecast evaluation. *Management Science, 55(4),* 582-590.