Journal of Statistical Planning and Inference 143 (2013) 1781–1790

Contents lists available at SciVerse ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Contrasting probabilistic scoring rules

Reason L. Machete*

Department of Mathematics and Statistics, P.O. Box 220, Reading RG6 6AX, UK

ARTICLE INFO

Article history:
Received 24 May 2012
Received in revised form
18 March 2013
Accepted 22 May 2013
Available online 29 May 2013

Keywords:
Estimation
Forecast evaluation
Probabilistic forecasting
Utility function

ABSTRACT

There are several scoring rules that one can choose from in order to score probabilistic forecasting models or estimate model parameters. Whilst it is generally agreed that proper scoring rules are preferable, there is no clear criterion for preferring one proper scoring rule above another. This manuscript compares and contrasts some commonly used proper scoring rules and provides guidance on scoring rule selection. In particular, it is shown that the logarithmic scoring rule prefers erring with more uncertainty, the spherical scoring rule prefers erring with lower uncertainty, whereas the other scoring rules are indifferent to either option.

1. Introduction

Issuing probabilistic forecasts is meant to express uncertainty about the future evolution of some quantity of interest. Such forecasts arise in many applications such as macroeconomics, finance, weather and climate forecasting. There are several scoring rules that one can choose from in order to elicit probabilistic forecasts, rank competing forecasting models or estimate forecast distribution parameters. It is generally agreed that one should select scoring rules that encourage a forecaster to state his 'best' judgement of the distribution, the so called *proper* scoring rules (Friedman, 1983; Nau, 1985; Gneiting and Raftery, 2007), but which one to use is generally an open question. We shall take scoring rules to be loss functions that a forecaster wishes to minimise. Scoring rules that are minimised if and only if the issued forecasts coincide with the forecaster's best judgement are said to be *strictly proper* (Gneiting and Raftery, 2007; Brocker and Smith, 2007). We shall restrict our attention to strictly proper scoring rules.

Nonetheless using scoring rules to rank competing forecasting models poses a problem; scoring rules do not provide a universally acceptable ranking of performance. In estimation, different scoring rules will yield different parameter estimates (Gneiting and Raftery, 2007; Johnstone and Lin, 2011). Moreover, a forecaster's best judgement may depart from the ideal; the ideal is a distribution that nature or the data generating process would give (Gneiting et al., 2007). Although strictly proper scoring rules encourage experts to issue their best judgements, such judgements may yet differ from each other and the ideal. Which scoring rule should one use to choose between two experts? Savage (1971) made the instructive statement that "any criteria for distinguishing among scoring rules must arise out of departures of actual subjects from the ideal." There have been some efforts to contrast scoring rules, but none seem to have followed this insight.

Bickel (2007) made empirical comparisons of the quadratic, spherical and logarithmic scoring rules and found them to yield different rankings of competing forecasts but failed to see why. Considering a concave nonlinear utility function that

* Tel.: +44 26774905669.
E-mail addresses: Reason.Machete@mopipi.ub.bw, r.l.machete@lse.ac.uk

0378-3758/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jspi.2013.05.012

explicitly depends on the scoring rule, he also found the logarithmic scoring rule to yield the least departures from honest opinions at maximal utility, a point he claimed favours it as a rule of choice. But a utility function need not be exponential nor explicitly depend on the scoring rule. Jose et al. (2008) considered weighted scoring rules and showed that they correspond to different utility functions. A limiting feature of the utility functions considered is that they are defined on bounded intervals; there are many applications in which the variable of interest is unbounded. Their motivation for weighted scoring rules is based on betting arguments, but it is not clear what the betting strategies (if any) are. Recently, Boero et al. (2011) empirically compared the Quadratic Probability Score (QPS), Ranked Probability Score (RPS) and the logarithmic scoring rule on UK inflation forecasts by the Monetary Policy Committee and the Survey of External Forecasters (SEF). They found the scoring rules to rank the two sets of distributions similarly. Upon ranking individual forecasters from the SEF, they found the RPS to have better discriminatory power than the QPS, a feature they attributed to the RPS's sensitivity to distance. Despite the foregoing efforts, there is lacking a theoretical assessment of what the preferences of the commonly used scoring rules are with respect to the ideal.

This paper contrasts how different scoring rules would rank competing forecasts of specified departures from ideal forecasts and provides guidance on scoring rule selection. It focuses upon those scoring rules that are commonly used in the forecasting literature, including econometrics and meteorology. More specifically, we contrast the relative information content of forecasts preferred by different scoring rules. Implications of the results on decision making are then suggested, noting that it may be desirable to be more or less uncertain when communicating probabilistic forecasts. We realise that an appropriate utility function may be unknown (Bickel, 2007) and expected utility theory may not even be appropriate (Kahneman and Tversky, 1979).

In Section 2, we consider three scoring rules of categorical forecasts, which then inspires our study of density forecasts in Section 3, where we consider four scoring rules. We conclude with a discussion of the results in Section 4.

2. Categorical forecasts

In this section, we consider the scoring of categorical forecasts. The scoring rules considered are Brier score (Brier, 1950), the logarithmic scoring rule and the spherical scoring rule (Friedman, 1983). In order to aid intuition in the next section, here we focus on the binary case. Another commonly used scoring rule for categorical forecasts is the Ranked Probability Score (RPS) (Epstein, 1969). In the binary case, the RPS score reduces to the Brier score.

It will be useful to be aware of the following basics. Given any vectors $\boldsymbol{f},\boldsymbol{g}\in\Re^m$, the *inner product* between the two vectors is

$$\langle \boldsymbol{f},\boldsymbol{g}\rangle = \sum_{i=1}^{m} f_i g_i,$$

from which the L_2-norm is defined by $\|\boldsymbol{f}\|_2 = \langle \boldsymbol{f},\boldsymbol{f}\rangle^{1/2}$.

2.1. The Brier score

Consider a probabilistic forecast $\{f_i\}_{i=1}^m$ of m categorical events. Suppose the true distribution is $\{p_i\}_{i=1}^m$. If the actual outcome is the jth category, the Brier score is given by (Brier, 1950)

$$BS(\boldsymbol{f},j) = \frac{1}{m}\sum_{i=1}^{m}(f_i-\delta_{ij})^2,$$

where $\delta_{ij}=0$ if $i\neq j$ and $\delta_{ij}=1$ if $i=j$. If follows that if we expand out the bracket we get

$$BS(\boldsymbol{f},j) = \frac{1}{m}\left(\sum_{i=1}^{m} f_i^2 - 2f_j + 1\right).$$

The expected Brier score is then given by

$$\mathbb{E}[BS(\boldsymbol{f},J)] = \frac{1}{m}\left\{\|\boldsymbol{\gamma}\|_2^2 + \sum_{i=1}^{m} p_i(1-p_i)\right\},$$

where $\boldsymbol{\gamma}$ is a vector with components $\gamma_i = f_i - p_i$ for all $i=1,\ldots,m$. It is evident from the last expression on the right hand side that the Brier score is effective with respect to the metric $d_2(\boldsymbol{f},\boldsymbol{g}) = \|\boldsymbol{f}-\boldsymbol{g}\|_2$. When $m=2$, we can put $f_1 = p+\gamma$, $p_1=p$ and $p_2=q$ and obtain

$$\mathbb{E}[BS(\boldsymbol{f},J)] = \gamma^2 + pq.$$

It follows that $\pm\gamma$ will yield the same Brier score. This means the Brier score does not discriminate between over-estimating and under-estimating the probabilities with the same amount. Furthermore, for any two forecasts $\boldsymbol{f}_i = (p+\gamma_i, q-\gamma_i)$, $i=1,2$, with $|\gamma_1|<|\gamma_2|$, the Brier score would prefer the forecast corresponding to γ_1.