A Distribution-oriented Probabilistic Verification Score for Forecasts

Chien Lu and Chuhsing Kate Hsiao

Department of Public Health

National Taiwan University

Abstract

In this research, a new score is proposed to evaluate the performance of binary probabilistic forecasts based on statistical predictive models. This new score utilizes the Kullback-Leibler divergence between the "original" probability density function (PDF) generated the forecast value and a "modified" PDF. The "original" PDF is indeed the distribution used to provide predictions, and is constructed on the basis of a certain statistical model; while the "modified" PDF is constructed with extra information of new observations. This new score can be decomposed into two components: one is the assessment for the forecast accuracy and the other is for systematic precision. Furthermore, it can be shown that the first part about precision can take into account the uncertainty inherited in the statistical model considered for forecast, and the second part about accuracy is equivalent to the divergence score.

Key word: Forecast verification, Brier score, Ignorance score, Kullback-Leibler divergence

1. Introduction

Many statistical models have been proposed in recent years for prediction of binary weather events, see, for example, the review by Casati et al. (2008).The verification of such forecasts generated from statistical models, however, is often limited to single point verification like Brier score (Brier 1950). Alternative verification scores based on information theory include the ignorance score (Good 1952; Roulston and Smith 2002) and the divergence score (Weijs et al. 2010). The divergence score is simply the Kullback-Leibler divergence (Kullback and Leibler 1951) from the perspective of single forecast verification; while the ignorance score is the negative logarithm of the forecast value measuring the information deficit. Both were advocated over Brier score (Bröcker and Smith 2007) because they are shown to be proper and they accommodate a penalty term in evaluations. Nevertheless, these verification scores fail to account for the uncertainty inherited in the statistical models used to generate weather forecasts. An appropriate verification score for any prediction should include not only the point estimate but also the uncertainty of that estimate. In other words, a high level uncertainty implies a low level confidence in the forecast, and a good verification score should be able to reflect this characteristic.

This article proposes a probabilistic verification score to evaluate the forecasts generated from statistical prediction models. This verification score is distribution oriented, and can take into account the prediction variation. The proposed score measures the "distance" between two probability density functions (PDFs), one is the "original" PDF for forecast and the other is a "modified" (or called updated) PDF. The "original" PDF is constructed based on a sampling distribution; while the "modified" PDF is constructed with extra information of the new observation through conditional probability derivation.

2. Score determination

Considering a binary event $Y^* \in \{0,1\}$ following a Bernoulli distribution with an unknown parameter p. Based on historical data $\widetilde{y}$ and a statistical forecast model $\pi(p \mid \widetilde{y})$, the estimate of p, denoted as $\hat{p}$, can be derived based on standard statistical tools. Take the simplest example, it can be the mean of all previous data, representing the estimated probability of Y^* being 1. In general, the forecast y^* for Y^* follows a distribution

$$Y^* \sim f(Y^* = y^* \mid \hat{p}) = \hat{p}^{y^*}(1-\hat{p})^{1\text{-}y^*}$$

derived from $\pi(p \mid \widetilde{y})$ and the Bernoulli density of $\widetilde{y}$.

We defined the "origin distribution" $\pi(p \mid \widetilde{y})$ containing information from previous observations $\widetilde{y}$ and prior knowledge in p, and utilize it to provide an estimate $\hat{p}$. The "modified distribution" $\pi(p \mid \widetilde{y}, y^*)$ is the distribution with extra information from the new observation y^*.The distance

$$d\{\pi(p \mid \widetilde{y}), \pi(p \mid \widetilde{y}, y^*)\}$$

between these two distributions can represent the magnitude of the influence of the new observation on the inference of the parameter p. Therefore, it can be used as a verification criterion.

Based on the above equations, we can calculate the probability density function of p conditioning on the existing outcome y^* with Bayes' rule:

$$\pi(p \mid \widetilde{y}, y^*) = \frac{\pi(p \mid \widetilde{y}) \times f(y^* \mid p)}{\int \pi(p \mid \widetilde{y}) \times f(y^* \mid p)dp}$$

3. Interpretation of the new score

There are many published methods measuring the difference between two probability density functions. In this article, we use Kullback-Leibler divergence for calculation.

For two continuous random variables *G* and *F* with density functions $g(x)$ and $f(x)$, respectively, the Kullback-Leibler divergence from *G* to *F* is defined as:

$$D_{KL}(G \mid F) \equiv \int g(x)\frac{g(x)}{f(x)}dx$$

In general, *G* represents the "real" distribution and *F* represents another model or a prediction. Here we calculate the distance from the original to the modified distribution as the verification score:

$$d\{\pi(p \mid \widetilde{y}), \pi(p \mid \widetilde{y}, y^*)\} \equiv \int \pi(p \mid \widetilde{y}, y^*)\log\frac{\pi(p \mid \widetilde{y}, y^*)}{\pi(p \mid \widetilde{y})}dp$$

After simple algebra, this score becomes

$$\frac{E[\hat{p}\log\hat{p}]}{E[\hat{p}]} - \log E[\hat{p}]$$

when $y^* = 1$, and

$$\frac{E[\hat{q}\log\hat{q}]}{E[\hat{q}]} - \log E[\hat{q}] \quad \text{where} \quad \hat{q} = (1-\hat{p})$$

when $y^* = 0$.

For the first term in each case,

$$\frac{E[\hat{p}\log\hat{p}]}{E[\hat{p}]}$$

when $y^* = 1$, and

$$\frac{E[\hat{q}\log\hat{q}]}{E[\hat{q}]} \quad \text{where} \quad \hat{q} = (1-\hat{p})$$

when $y^* = 0$. This term can be interpreted as the adjusted uncertainty. The numerators are entropy-like functions that represent the uncertainty of the prediction. The denominator is the point estimate. If the point estimates is close to $y*$, the denominator becomes large, and the function decreases. In other words, the first part represents the precision of the statistical model.

As for the second part, it is

$$-\log E[\hat{p}]$$

when $y*=1$, and

$$-\log E[\hat{q}] \quad \text{where} \quad \hat{q}=(1-\hat{p})$$

when $y*=0$. This is the same as the divergence score proposed by Weijs et al (2010). As mentioned in Section 1, the divergence score is equal to the ignorance score but different in the logarithmic base. In addition, the divergence score is a verification criterion of the single value forecast and can be separated according to the classic reliability-resolution-uncertainty decomposition. It represents the quality derived from the point estimate of the statistical predictive model. The details of calculation will be provided in appendix.

Appendix

Details of score calculation

The "modified" PDF by Bayes' Rrle is

$$\pi(p \mid \tilde{y}, y*) = \frac{\pi(p \mid \tilde{y}) \times f(y* \mid p)}{\int \pi(p \mid \tilde{y}) \times f(y* \mid p)dp}$$
$$= \frac{\pi(p \mid \tilde{y}) \times f(y* \mid p)}{f(\tilde{y}, y*)}$$

The Kullbak-Leibler Divergence (D_{KL}) is then

$$\begin{aligned}
D_{KL} &= \int \pi(p \mid \tilde{y}, y^*) \log \frac{\pi(p \mid \tilde{y}, y^*)}{\pi(p \mid \tilde{y})} dp \\
&= \int \pi(p \mid \tilde{y}, y^*) \log \frac{\pi(p \mid \tilde{y}) \times f(y^* \mid p)}{\pi(p \mid \tilde{y}) \times f(\tilde{y}, y^*)} dp \\
&= \int \pi(p \mid \tilde{y}, y^*) \log \frac{f(y^* \mid p)}{f(\tilde{y}, y^*)} dp \\
&= \int \pi(p \mid \tilde{y}, y^*) \log f(y^* \mid p)dp - \log f(\tilde{y}, y^*) \int \pi(p \mid \tilde{y}, y^*)dp \\
&= \int \pi(p \mid \tilde{y}, y^*) \log f(y^* \mid p)dp - \log f(\tilde{y}, y^*)
\end{aligned}$$

$$\begin{aligned}
\int \pi(p \mid \tilde{y}, y^*) \log f(y^* \mid p)dp &= \int \left(\frac{\pi(p \mid \tilde{y}) \times f(y^* \mid p)}{f(\tilde{y}, y^*)} \right) \log p^{y^*}(1-p)^{y^*} dp \\
&= \frac{\int \pi(p \mid \tilde{y}) \times p^{y^*}(1-p)^{y^*} \log p^{y^*}(1-p)^{y^*} dp}{f(\tilde{y}, y^*)} \\
&= \frac{E\left[p^{y^*}(1-p)^{y^*} \log p^{y^*}(1-p)^{y^*}\right]}{E\left[p^{y^*}(1-p)^{y^*}\right]} \\
&= \frac{E[\hat{p} \log \hat{p}]}{E[\hat{p}]}, \text{when } y^* = 1 \\
&= \frac{E[\hat{q} \log \hat{q}]}{E[\hat{q}]}, \text{when } y^* = 1, \hat{q} = (1-\hat{p})
\end{aligned}$$

$$\begin{aligned}
\log f(\tilde{y}, y*) &= \log\left(\int \pi(p \mid \tilde{y}) \times f(y* \mid p)dp \right) \\
&= \log\left(\int \pi(p \mid \tilde{y}) \times \left(p^{y*}(1-p)^{1-y*} \right) dp \right) \\
&= \log E\left[\hat{p}^{y*}(1-\hat{p})^{1-y*} \right] \\
&= \log E[\hat{p}], \text{when } y* = 1 \\
&= \log E[\hat{q}], \text{when } y* = 0, \hat{q} = (1-\hat{p})
\end{aligned}$$

Combining two equations above, we have

$$\begin{aligned}
D_{KL} &= -\log E[\hat{p}] + \frac{E[\hat{p} \log \hat{p}]}{E[\hat{p}]}, \text{when } y^* = 1 \\
&\quad -\log E[\hat{q}] + \frac{E[\hat{q} \log \hat{q}]}{E[\hat{q}]}, \text{when } y^* = 0, \hat{q} = (1-\hat{p})
\end{aligned}$$

Reference

Brier, G. W., 1950: "Verification of forecasts expressed in terms of probability", Mon. Wea. Rev., 78, 1–3.

Bröcker J, Smith L.A., 2007: "Scoring probabilistic forecasts: on the importance of being proper", Weather and Forecasting 22(2): 382–388.

Casati, B., Wilson, L.J., Stephenson, D.B., 2008: "Forecast verification: current status and future directions", Meteorological Applications 15 (1), 3–18.

Good IJ., 1952: "Rational decisions", Journal of the Royal Statistical Society Series B-Methodological 14: 107–114.

Kullback S. Leibler, R.A., 1951: "On Information and Sufficiency", Annals of Mathematical statistics 22 (1): 79–86.

Roulston MS, Smith L.A., 2002: "Evaluating probabilistic forecasts using information theory", Monthly Weather Review 130: 1653–1660.

Weijs, S., Van Nooijen, R., and Van de Giesen, N., 2010: "Kullback–Leibler divergence as a forecast skill score with classic reliability-resolution-uncertainty decomposition", Mon. Weather Rev., 138, 3387–3399.

Acknowledgement

Central Weather Bureau has provided the author an opportunity for internship and it motivates this research of forecast verification. This work was sponsored by Department of Public Health, National Taiwan University, Taipei, Taiwan.