Notes on Brier score consistency

James D. Malley

Original 9 May 2011 Revised 3 February 2014

Suppose given data (y_i, x_i), $i = 1,\ldots,n$, with binary outcomes $y_i = 1,0$, and data vectors x_i. No assumptions are made for the distribution X from which the x are drawn. Define the Brier score by

$$Brier\ score = (1/n)\Sigma(y_i - \hat{p}_i)^2$$

for $\hat{p}_i$ an estimate of the true probability $p(x_i)$ evaluated at the data vector x_i :

$$p_i = p(x_i) = \Pr(y_i = 1 \mid x_i) = E(Y_i \mid X = x_i).$$

If the estimator $\hat{p}_i$ is assumed to be based on all the data, and is evaluated at x_i in the data set, then it is a resubstitution estimate. It is likely that this double use of the data will lead to an optimistic—too low—value for the Brier score; More on this below.

It is known that the Brier score is a *proper score* as its expectation is minimized at the true (but unknown) probability p_i; See [1]. Hence

$$E[(1/n)\Sigma(y_i - p_i)^2] \le E[(1/n)\Sigma(y_i - \hat{p}_i)^2] = E[Brier\ score]. \qquad (1)$$

In the other direction, consider the decomposition

$$y_i - \hat{p}_i = y_i - p_i + p_i - \hat{p}_i,$$

so that

$$(y_i - \hat{p}_i)^2 \le (y_i - p_i)^2 + (p_i - \hat{p}_i)^2.$$

Introduce notation as in [2] for the regression estimate $m_n(x_i)$ based on all the data, and let $m(x_i)$ be the true regression function evaluated at x_i. Define the *resubstitution squared-error* as

$$\delta(n,i) = E(p_i - \hat{p}_i)^2 = E(m(x_i) - m_n(x_i))^2. \tag{2}$$

Then:

$$\begin{aligned} E[(1/n)\Sigma(y_i - \hat{p}_i)^2] &\le E[(1/n)\Sigma(y_i - p_i)^2] + E[(1/n)\Sigma((m(x_i) - m_n(x_i))^2] \\ &= E[(1/n)\Sigma(y_i - p_i)^2] + \delta(n,i)/n. \end{aligned} \tag{3}$$

In view of (1) and (2), and *if* the resubstitution estimate $\delta(n,i)$ is consistent with a rate at least $o(n)$, it would follow that the Brier score is also consistent at the same rate. However, it is not known if the resubstitution estimate is consistent, even if the regression estimate is so: Does $E(m(x) - m_n(x))^2 \to 0$ as $n \to \infty$ imply $E(p_i - \hat{p}_i)^2 \to 0$?

Consider instead a Brier score based on training and test data, where the estimator $\hat{p}_i$ is derived from the training data but is evaluated at a test point x_i not in the data. Then, a more conclusive result is possible.

Thus in (3) replace the term $E[(1/n)\Sigma((m(x_i) - m_n(x_i))^2]$ with the same figure of merit but where the test point x_i is not in the training data. Under regression consistency

$$\begin{aligned} E[(1/n)\Sigma((m(x_i) - m_n(x_i))^2] &= (1/n)\Sigma E[(m(x_i) - m_n(x_i))^2 \\ &= (1/n)\Sigma E[(m(x) - m_n(x))^2 = E[(m(x) - m_n(x))^2 \to 0, \end{aligned} \tag{4}$$

as $n \to \infty$. Consequently the Brier score, evaluated on test data, is also consistent.

References.

[1] Tilmann Gneiting, Adrian Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378.

[2] László Györfi, Michael Kohler, Adam Krzyżak, Harro Walk (2002). **A Distribution-Free Theory of Nonparametric Regression**. Springer.