Chap.10 Forecast verification [Book, Sect. 8.5]

After a forecast model has been built, need to evaluate the quality of its forecasts, a process known as *forecast verification* or forecast evaluation (Jolliffe and Stephenson, 2003).

10.1 Binary forecasts

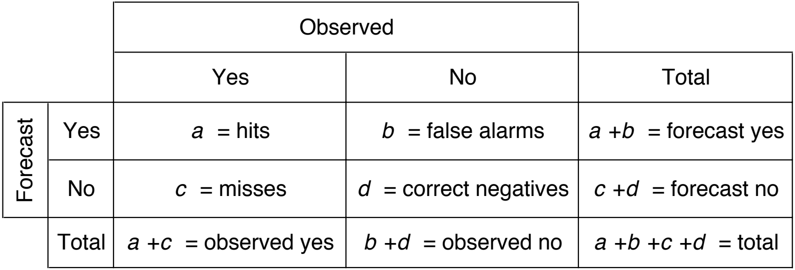
Start with forecasts for 2 classes or categories, where class 1 is for an event (e.g. tornado) and class 2 for a non-event. Model forecasts and observed data can be compared and arranged in a 2×2 contingency table .

The number of events forecasted and indeed observed are called "hits" and are place in entry *a* of the table. E.g. forecasts for tornados which turned out to be correct.

Entry b is the number of false alarms, e.g. tornados forecasted but never materialized.

Entry c is the number of misses, e.g. tornados appeared in spite of non-tornado forecasts.

Entry d is the number of correct negatives, i.e. non-tornado forecasts which turned out to be correct.



		Observed		
		Yes	No	Total
Forecast	Yes	a = hits	b = false alarms	$a+b$ = forecast yes
	No	c = misses	d = correct negatives	$c+d$ = forecast no
	Total	$a+c$ = observed yes	$b+d$ = observed no	$a+b+c+d$ = total

Figure : A 2 × 2 contingency table used in the forecast verification of a 2-class problem. The number of forecasted “yes” and “no”, and the number of observed “yes” and “no” are the entries in the table. Marginal totals are also listed, e.g. the top row sums to $a + b$, the total number of tornados forecasted, whereas the first column sums to $a + c$, the total number of tornados observed. Finally, the total number of cases N is given by $N = a + b + c + d$.

The simplest measure of accuracy of binary forecasts is the fraction correct (FC) [or hit rate (obsolete)], i.e. the number of correct forecasts divided by the total number of forecasts,

$$\mathrm{FC} = \frac{a+d}{N} = \frac{a+d}{a+b+c+d} \ , \tag{1}$$

where FC ranges between 0 and 1, with 1 being the perfect score.

Unfortunately, this measure becomes very misleading if the number of non-events vastly outnumber the number of events. E.g if $d \gg a, b, c$, then (1) yields FC ≈ 1.

E.g., in Marzban and Stumpf (1996), one NN tornado forecast model has $a = 41$, $b = 31$, $c = 39$ and $d = 1002$, since the vast majority of days has no tornado forecasted and none observed. The overwhelming size of d lifts FC to a lofty value of 0.937.

In such situations, where the non-events vastly outnumber the events, including d in the score is rather misleading. Dropping d in both the numerator and the denominator in (1) gives the threat score (TS) or critical success index (CSI) ,

$$\mathrm{TS} = \mathrm{CSI} = \frac{a}{a+b+c} \ , \tag{2}$$

which is a much better measure of forecast accuracy than FC in such situations. The worse TS is 0 and the best TS is 1.

Q1: What is the threat score for the tornado forecast model?

To see what fraction of the observed events ("yes") were correctly forecasted, we compute the probability of detection (POD) (or hit rate)

$$\mathrm{POD} = \frac{\mathrm{hits}}{\mathrm{hits} + \mathrm{misses}} = \frac{a}{a+c} \ , \tag{3}$$

with the worst POD score being 0 and the best score being 1.

Easy to increase POD if we simply issue many more forecasts of events ("yes"), despite most of them being false alarms. Need to know if there is forecast bias or frequency bias :

$$B = \frac{\text{total ``yes'' forecasted}}{\text{total ``yes'' observed}} = \frac{a+b}{a+c} \ . \tag{4}$$

B would raise concern that the model is forecasting far too many events compared to the number of observed events.

To see what fraction of the forecasted events ("yes") never materialized, we compute the false alarm ratio (FAR)

$$\mathrm{FAR} = \frac{\text{false alarms}}{\text{hits} + \text{false alarms}} = \frac{b}{a+b} \ , \tag{5}$$

with the worst FAR score being 1 and the best score being 0.

Don't confuse the false alarm ratio (FAR) with the false alarm rate (F), also known as the probability of false detection (POFD) . F measures the fraction of the observed "no" events which were incorrectly forecasted as "yes", i.e.

$$F = \mathrm{POFD} = \frac{\text{false alarms}}{\text{false alarms} + \text{correct negatives}} = \frac{b}{b+d} \ , \tag{6}$$

with the worst F score being 1 and the best score being 0.

While F is not as commonly given as FAR and POD, it is one of the axes in the relative operating characteristic (ROC) diagram, used widely in probabilistic forecasts (Marzban, 2004; Kharin and Zwiers, 2003).

In an ROC diagram, F is the abscissa and POD, the ordinate. Although our model may be issuing probabilistic forecasts in a 2-class problem, we are actually free to choose the decision threshold used in the classification, i.e. instead of using a posterior probability of 0.5 as the threshold for deciding whether to issue a "yes" forecast, we may want to use 0.7 as the threshold if we want fewer false alarms (i.e. lower F) (at the expense of a lower POD), or 0.3 if we want to increase our POD (at the expense of increasing F as well). The result of varying the threshold generates a curve in the ROC diagram.

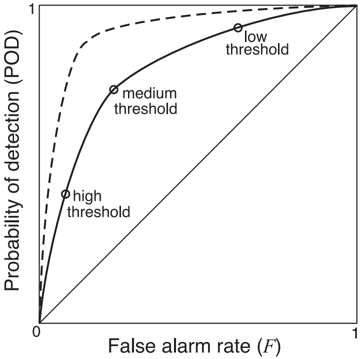


Figure : A relative operating characteristic (ROC) diagram illustrating the trade-off between the false alarm rate (F) and the probability of detection (POD) as the classification decision threshold is varied for a given model (solid curve). Dashed curve shows the ROC of a better model while diagonal line (POD $= F$) means a model with zero skill.

The choice of the threshold hinges on the cost associated with missing an event and that with issuing a false alarm. For instance, if we miss forecasting a powerful hurricane hitting a vulnerable coastal city, the cost may be far higher than that from issuing a false alarm, so we would want a low threshold value to increase the POD.

10.2 Skill scores [Book, Sect. 8.5.1]

Various *skill scores* have been designed to measure the relative accuracy of a set of forecasts, with respect to a set of *reference* or control forecasts.

Choices for the reference forecasts include (i) persistence, (ii) climatology, (iii) random forecasts and (iv) forecasts from a standard model.

(i) Persistence forecasts simply persists the anomalies to the future (e.g. tomorrow's weather is forecasted to be the same as today's weather).
(ii) Climatological forecasts simply issue the climatological mean value in the forecast.
(iii) In random forecasts, events are forecasted randomly but in agreement with the *forecasted* frequency of such events. E.g., if tornados are *forecasted* only 2% of the time in your model, then random forecasts also only forecast tornados 2% of the time.
(iv) Finally the reference model can be a standard model, as the researcher is trying to show that her new model is better.

For a particular measure of accuracy A, the skill score (SS) is defined generically by

$$\mathrm{SS} = \frac{A - A_{\mathrm{ref}}}{A_{\mathrm{perfect}} - A_{\mathrm{ref}}} \,, \tag{7}$$

where A_{perfect} is the value of A for a set of perfect forecasts, and A_{ref} is the value of A computed over the set of reference forecasts.

Note that if we define $A' = -A$, then SS is unchanged if computed using A' instead of A. This shows that SS is unaffected by whether A is positively or negatively oriented (i.e. whether better accuracy is indicated by a higher or lower value of A).

The Heidke skill score (HSS) (Heidke, 1926) is the skill score (7) using the fraction correct (FC) for A and random forecasts as the reference, i.e.

$$\mathrm{HSS} = \frac{\mathrm{FC} - \mathrm{FC}_{\text{random}}}{\mathrm{FC}_{\text{perfect}} - \mathrm{FC}_{\text{random}}} \ . \tag{8}$$

Hence, if the forecasts are perfect, HSS $= 1$; if they are only as good as random forecasts, HSS $= 0$; and if they are worse than random forecasts, HSS is negative. From (1), FC can be interpretted as the fraction of hits (a/N) plus the fraction of correct negatives (d/N).

For $\mathrm{FC}_{\mathrm{random}}$ obtained from random forecasts, the fraction of hits is the product of two probabilities, P("yes" forecasted) and P("yes" observed), i.e. $(a+b)/N$ and $(a+c)/N$, respectively.
Similarly, the fraction of correct negatives from random forecasts is the product of P("no" forecasted) and P("no" observed), i.e. $(c+d)/N$ and $(b+d)/N$.

With $\mathrm{FC}_{\mathrm{random}}$ being the fraction of hits plus the fraction of correct negatives, we have

$$\mathrm{FC}_{\mathrm{random}} = \left(\frac{a+b}{N}\right)\left(\frac{a+c}{N}\right) + \left(\frac{c+d}{N}\right)\left(\frac{b+d}{N}\right). \qquad (9)$$

Substituting this into (8) and invoking (1) and $\mathrm{FC}_{\mathrm{perfect}} = 1$, HSS is then given by

$$\mathrm{HSS} = \frac{(a+d)/N - [(a+b)(a+c)+(b+d)(c+d)]/N^2}{1-[(a+b)(a+c)+(b+d)(c+d)]/N^2} \,, \tag{10}$$

which can be simplified to

$$\mathrm{HSS} = \frac{2(ad-bc)}{(a+c)(c+d)+(a+b)(b+d)} \,. \tag{11}$$

The Peirce skill score (PSS) (Peirce, 1884), also called the *Hansen and Kuipers' score*, or the *true skill statistics* (TSS) is similar to the HSS, except that the reference used in the denominator of (8) is unbiased, i.e. $P($"yes" forecasted$)$ is set to equal $P($"yes" observed$)$, and P("no" forecasted) to P("no" observed) for $\mathrm{FC}_{\mathrm{random}}$ in the denominator of (8), whence

$$\mathrm{FC}_{\mathrm{random}} = \left(\frac{a+c}{N}\right)^2 + \left(\frac{b+d}{N}\right)^2 . \tag{12}$$

The PSS is computed from

$$\mathrm{PSS} = \frac{(a+d)/N - [(a+b)(a+c)+(b+d)(c+d)]/N^2}{1-[(a+c)^2+(b+d)^2]/N^2} , \tag{13}$$

which simplies to

$$\mathrm{PSS} = \frac{ad-bc}{(a+c)(b+d)} . \tag{14}$$

PSS can also be expressed as

$$\mathrm{PSS} = \frac{a}{a+c} - \frac{b}{b+d} = \mathrm{POD} - F, \tag{15}$$

upon invoking (3) and (6).

Again, if the forecasts are perfect, PSS $= 1$; if they are only as good as random forecasts, PSS $= 0$; and if they are worse than random forecasts, PSS is negative.

Q2: For *continuous* variables, the root mean squared error (RMSE) between forecasted and observed values is commonly used to evaluate a forecast model. If model 1 has RMSE of 0.0395, model 2 has RMSE of 0.0374 and the RMSE of a standard model is 0.0389. What are the RMSE skill scores for model 1 and model 2 with the standard model as reference?

Mean absolute error (MAE) is actually a better measure of the average error than RMSE (Willmott and Matsuura, 2005), so MAE SS is better than RMSE SS.

10.3 Multiple classes [Book, sect. 8.5.2]

Next consider the forecast verification problem with c classes, where c is an integer > 2. The contingency table is then a $c \times c$ matrix, with the ith diagonal element giving the number of correct forecasts for class C_i.

Fraction correct (FC) in (1) generalizes easily, as FC is simply the sum of all the diagonal elements divided by the sum of all elements of the matrix.

Other measures such as POD, FAR, etc. do not generalize naturally to higher dimensions. Instead the way to use them is to collapse the $c \times c$ matrix to a 2×2 matrix.

E.g., if the forecast classes are “cold”, “normal” and “warm”, we can put “normal” and “warm” together to form the class of “non-cold” events. Then we are back to two classes, namely “cold” and “non-cold”, and measures such as POD, FAR, etc. can be easily applied. Similarly, we can collapse to only “warm” and “non-warm” events, or to “normal” and “non-normal” events.

For multi-classes, HSS in (8) and (10) generalizes to

$$\mathrm{HSS} = \frac{\sum_{i=1}^{c} P(f_i, o_i) - \sum_{i=1}^{c} P(f_i)P(o_i)}{1 - \sum_{i=1}^{c} P(f_i)P(o_i)} \ , \tag{16}$$

where f_i denotes class C_i forecasted, o_i denotes C_i observed, $P(f_i, o_i)$ the joint probability distribution of forecasts and observations, $P(f_i)$ the marginal distribution of forecasts and $P(o_i)$ the marginal distribution of observations. It is easy to see that (16) reduces to (10) when there are only 2 classes.

PSS in (13) also generalizes to

$$\mathrm{PSS} = \frac{\sum_{i=1}^{c} P(f_i, o_i) - \sum_{i=1}^{c} P(f_i)P(o_i)}{1 - \sum_{i=1}^{c}[P(o_i)]^2} \ . \tag{17}$$

10.4 Probabilistic forecasts [Book, Sect. 8.5.3]

In probabilistic forecasts, one can issue forecasts for binary events based on the posterior probability, then compute skill scores for the classification forecasts. Alternatively, one can apply skill scores directly to the probabilistic forecasts.

The most widely used score for the probabilistic forecasts of an event is the Brier score (BS) (Brier, 1950). Formally, this score resembles the MSE, i.e.

$$\mathrm{BS} = \frac{1}{N}\sum_{n=1}^{N}(f_n - o_n)^2, \tag{18}$$

where there is a total of N pairs of forecasts f_n and observations o_n. While f_n is a continuous variable within $[0, 1]$, o_n is a binary variable, being 1 if the event occurred and 0 if it did not occur.

BS is negatively oriented, i.e. the lower the better. Since $|f_n - o_n|$ is bounded between 0 and 1 for each n, BS is also bounded between 0 and 1, with 0 being the perfect score.

From (7), the Brier skill score (BSS) is then

$$\mathrm{BSS} = \frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{ref}}}{0 - \mathrm{BS}_{\mathrm{ref}}} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}}} \ , \tag{19}$$

where the reference forecasts are often taken to be random forecasts based on climatological probabilities.

Unlike BS, BSS is positively oriented, with 1 being the perfect score and 0 meaning no skill relative to the reference forecasts.

For BS and BSS, the observed variable is a binary variable. If the observed variable is a continuous variable, there are also probabilistic forecast scores described in Book, Sect.9.4.

References:

Brier, W. G. (1950). Verification of forecasts expressed in terms of probabilities. *Monthly Weather Review*, 78:1–3.

Heidke, P. (1926). Berechnung des Erfolges und der Güte der Windstärkevorhersagen in Sturmwarnungsdienst. *Geografiska Annaler*, 8:310–349.

Jolliffe, I. T. and Stephenson, D. B., editors (2003). *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley, Chichester.

Kharin, V. V. and Zwiers, F. W. (2003). On the ROC score of probability forecasts. *Journal of Climate*, 16(24):4145–4150.

Marzban, C. (2004). The ROC curve and the area under it as performance measures. *Weather and Forecasting*, 19(6):1106–1114.

Marzban, C. and Stumpf, G. J. (1996). A neural network for tornado prediction based on doppler radar-derived attributes. *Journal of Applied Meteorology*, 35(5):617–626.

Peirce, C. S. (1884). The numerical measure of the success of predictions. *Science*, 4:453–454.

Willmott, C. J. and Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Clim. Res.*, 30:79–82.