HOW TO INTERPRET PROBABILISTIC FORECASTS (IN PARTICULAR FOR WEATHER AND CLIMATE)

Jochen Bröcker
Ackn: S. Siegert, H. Kantz

MPI für Physik komplexer Systeme, Dresden, Germany (on leave)

CATS

Outline

Probability forecasts in weather and climate

Historical references after Murphy (1998)

The probability of rain was much smaller than other times.

The probability of a fair day to that of a wet one is as ten to one.

Dalton, J. (1793)

A knowledge of the degree of certainty with which an event may be expected, increases the value of information.

Nichols, W.S. (1890)

The Signal Corps Meteorological Service (. . .) in WWI issued forecasts that included a statement as to the probable accuracy of the forecast, (. . .) expressed in terms of the odds in favour of the forecast. (. . .) the inclusion of this information made it possible "to make the forecast absolutely definite and such qualifications as 'probable' or 'possibly' have never been used"

Murphy (1998)

Due to the inherent uncertainties, any forecast (in particular for weather and climate) should include a statement as to its probable accuracy. One option are probabilities.

Scoring rules

Problem: Evaluation

How to compare forecasts with observations, as these are two unlike objects?

- Observations $\eta \in \{1 \dots K\}$
- Forecasts $q = (q^{(1)} \dots q^{(K)})$, with $q^{(k)} \geq 0$, $\sum_k q^{(k)} = 1$.

A *Scoring Rule* $S(q, \eta)$ assigns "points" to q based on the observation η.
Convention: A smaller score indicates a better forecast.

Proper Scoring rules

Scoring rules have to be *proper* in order to avoid inconsistencies (Brown, 1970; Bröcker and Smith, 2007).

$$\sum_k S(q,k)p^{(k)} \geq \sum_k S(p,k)p^{(k)}$$

Roughly speaking: If you expect the event $\eta = k$ to occur with probability $p^{(k)}$ for $k = 1 \ldots k$, then you also expect the forecast p to score better than any other forecast q.

Aside: Defining personal probabilities through scoring rules

- Suppose a subject is a HOMO OECONOMICUS
- You promise the subject a reward $R(q,k)$, shoul it turn out that $\eta = k$.
- The subject has to chose q

If $-R$ is a proper scoring rule, then q can be interpreted as a *subjective probability* (Savage, 1971).

Probabilistic forecasting *systems*

Various probabilistic forecasts are issued on a daily basis. Such a probabilistic forecasting systems should be evaluated on the basis of many forecast–verification *pairs*.

If $T := \{(q_n, \eta_n); n = 1 \ldots N\}$ is an *archive* of forecast–verification pairs, we define the *empirical score* as

$$\frac{1}{N} \sum_n S(q_n, \eta_n)$$

i.e. the empirical average of the individual scores.

Reliability and resolution

Alternative concepts of forecast quality exist, such as *Reliability* and *Resolution*.
To define these, we assume that q and η are random variables on some probability space.
The archive $T = \{(q_n, \eta_n)\}$ is a series of independent realisations of (q, η).

Reliability

Reliability means: forecast probabilities should agree with actually observed relative frequencies.

> *There should be rain on 20% of those days where the forecast for rain was* 0.2*.*

More rigorous definition:

$$\mathbb{P}(\eta = k | q) = q^{(k)}, \qquad k = 1 \ldots K$$

Resolution

Resolution means: different forecasts are indicative of different events.

> *Days with probability of rain $p < .5$ and days with probability of rain $p \geq .5$ should have different fractions of rainy days.*

More rigorous definition: A forecast has resolution if η and q are *not* independent. Equivalently, no resolution means

$$\mathbb{P}(\eta = k|q) = \mathbb{P}(\eta = k), \qquad k = 1 \ldots K$$

Proper scoring rules and reliability/resolution

Proper scoring rules quantify both reliability and resolution (Murphy and Winkler, 1987; Bröcker, 2009)

- $\mathrm{s}(q, p) := \sum_k S(q, k) p^{(k)}$
- $\mathrm{e}(p) := \mathrm{s}(p, p)$
- $\mathrm{d}(q, p) := \mathrm{s}(q, p) - \mathrm{e}(p)$
- Proper score means $\mathrm{s}(q, p) \geq \mathrm{e}(p)$ or $\mathrm{d}(q, p) \geq 0$

Proper scoring rules decomposition

- $\pi^{(k)}(q) = \mathbb{P}(\eta = k|q)$
- $\overline{\pi}^{(k)} = \mathbb{P}(\eta = k)$
- Reliability: $\pi^{(k)}(q) = q^{(k)}$
- No resolution: $\pi^{(k)}(q) = \overline{\pi}^{(k)}$

Proper scoring rules decomposition

$$\mathbb{E}\left(S(q,\eta)\right) = \text{UNC} + \text{REL} - \text{RES}$$

where

- UNC $= \mathrm{e}(\overline{\pi})$ quantifies the uncertainty in η (no q–dependence)
- REL $= \mathrm{d}(q, \pi(q))$ quantifies the *UNreliability* of q.
- RES $= \mathrm{d}(\overline{\pi}, \pi(q))$ quantifies the *resolution* of q

Part II

How to misinterprete probability forecasts

Probabilities in weather

Question Will there be rain tomorrow?

Forecast Probability p that tomorrow there will be rain.

Interpretation of p

The probability expresses a degree of belief of the forecaster. If probability forecasts are issued over long periods, they might be compared with observed frequencies $\Rightarrow$ reliability.

Probabilities in climate

Question Will hurricanes become more frequently in the future?

Forecast Probabilities $p^{(0)} \dots p^{(K)}$ that there will be $0 \dots K$ hurricanes per year in the North Atlantic.

Interpretation of p

This is a *very different* kind of probability! The verifying event is *not* "k hurricanes in a given year". The verifying event is "The fraction of years with k hurricanes is about $p^{(k)}$".

Roll the dice

Roll the dice

An example
single shot

Let $\eta \in \{0, 1\}$. Your reward is $R = a + b\eta$. Forecaster: “The probability of $\eta = 1$ is q”.
Then

$$\mathbb{E}R = a + bq$$
$$\mathrm{Var}R = b^2 q(1 - q)$$

An example
multiple shots

Let $\eta_n \in \{0, 1\}, n = 1 \dots N$. Your reward on average per n is $R = a + b\frac{1}{N}\sum_n \eta_n$. Forecaster says: “The probability of $\eta = 1$ is q”, but he probably *means*: “The long–term frequency of $\eta_n = 1$ is about q”.
We make this precise: $q \in [0, 1]$ is a random variable, and given q, the η_n are iid with expectation q.
Then

$$\mathbb{E}R = a + b\overline{q}$$

$$\mathrm{Var}R = b^2 \left\{ \frac{\overline{q}(1 - \overline{q})}{N} + (1 - \frac{1}{N})\mathrm{Var}q \right\}$$

Second order probabilities?

A rigorous way to deal with these problems would be to introduce *second order probabilities*.
Downside: We would end up with statements like

> *In the future, the probability of getting more than six tropical cyclones per season in the North Atlantic is larger than 30% with probability 1/2.*

How can we make such statements more detailed without making them more awkward?
Sometimes, it is possible to express second order probabilities using non–probabilistic odds (J.B., Kevin Judd).

- Probability forecasts have a well defined meaning and can be evaluated objectively using scoring rules.
- Scoring rules quantify a combination of reliability and resolution, two notions of forecast value for which the case can be made independently.
- Probability forecasts must not be confused with forecasts for future observed frequencies (as in climate forecasts). Providing reliable uncertainty information for the latter maybe requires introduction of second order probabilities.
- As second order probabilities are cumbersome to deal with, it would be nice to have “something simpler” to give decision support.

References I

Jochen Bröcker. Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512 – 1519, 2009.

Jochen Bröcker and Leonard A. Smith. Scoring probabilistic forecasts: The importance of being proper. *Weather and Forecasting*, 22(2):382–388, 2007.

Thomas A. Brown. Probabilistic forecasts and reproducing scoring systems. Technical Report RM–6299–ARPA, RAND Corporation, Santa Monica, CA, June 1970.

Allan H. Murphy. The early history of probability forecasts: some extensions and clarifications. *Weather and Forecasting*, 13:5–15, 1998.

References II

Allan H. Murphy and Robert L. Winkler. A general framework for forecast verification. *Monthly Weather Review*, 115: 1330–1338, 1987.

Leonard J. Savage. Elicitation of personal probabilities and expectation. *Journal of the American Statistical Association*, 66(336):783–801, 1971.