Measuring the accuracy of probabilistic forecasts with scoring rules

Evgeni Ovcharov

Kiten, Bulgaria
June 28, 2017

Introduction

- Let Y be the quantity to be forecast, commonly referred to as the observation, predictant, verification, or target.
- Y is considered a random variable taking values in some space $\mathcal{Y} \subset \mathbb{R}^n$.
- A probabilistic forecast for Y tries to predict the probability distribution of Y given some information known before and at the time of making the forecast.
- If I is the information available to the forecaster (often in the form of observations of random variables), his model predicts the conditional distributions of $Y|I$.
- Sometimes a forecaster may be asked to report the whole predictive distribution of Y
  - such a forecast is called distributional, or probabilistic,

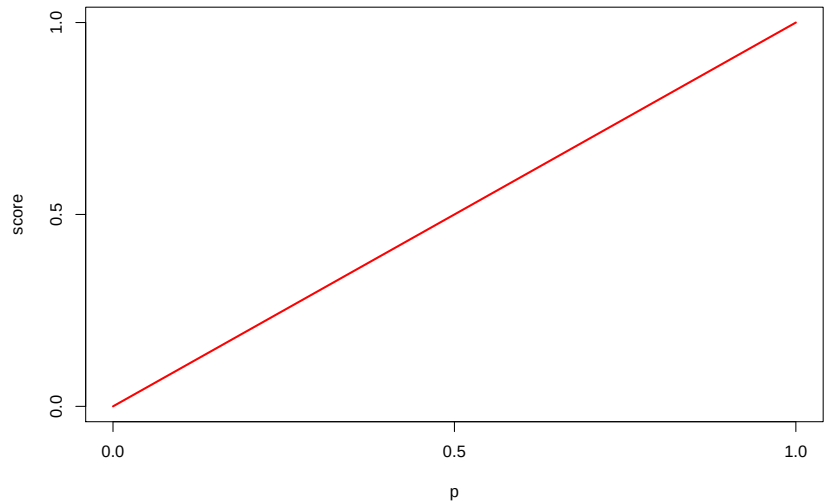
  or he may be asked to report only a specific parameter or functional of the predictive distribution such the mean, the median, the standard deviation, etc.,
  - the latter type of forecast is called a point-forecast.

The meteorologists' story

- Ivan oversees the weather stations in Bulgaria.
- He wants to validate the quality of precipitation forecasts of each weather station.
- He instructed each weather station to forecast the likelihood of rain $p \in [0, 1]$ and amount of rainfall (in mm) $r \in [0, \infty)$.
- He decided to make the salaries of each stationmaster depend on the accuracy of their reports.
- To quantify the accuracy, he settled on the simple day score: $50p$ (in €) if it rains, and $50(1 - p)$ (in €) if it does not.

The meteorologists' story

Ivan implements the linear score. When p is the forecasted probability of the event that materializes, the linear score has the following graph:



The meteorologists' story

- The linear score seems quite a reasonable choice to measure the accuracy of probabilistic forecasts.
- Ivan also settled to give a bonus of 30 € if the predicted rainfall r is within 1 mm of being correct.
- The full score which Ivan implemented is given by

$$50 \cdot (p\mathbb{1}_{\{\text{rain}\}} + (1-p)\mathbb{1}_{\{\text{no rain}\}}) + 30 \cdot \mathbb{1}_{\{|r-r^*|<1\}} \tag{1}$$

- At first the plan worked phenomenally, but at payday some of the stationmasters were enraged by the salaries they got and demanded that Ivan reveals the formula by which he calculated their accuracy. He obliged, quelling the upset.
- Unfortunately for Ivan, the next month did not bode as well. After receiving numerous complaints by citizens about the weather reporting, he took a look at the forecast history again. The forecasts were substantially worse than before and the records contained some peculiar patterns:

The meteorologists' story

- nearly all probabilistic forecasts were either 0 or 1 and about a third of all "100% chance of rain" forecasts were accompanied with a projection of 0 mm of rainfall.
- After discussing with colleagues, who were equally confused, Ivan decided it must be his scoring mechanism. He did a little research and discovered that a more common way to evaluate rainfall forecasts was to use the relative error,

$$\mathrm{RE}(r, r^*) = \left| \frac{r - r^*}{r} \right|.$$

- Satisfied, he announced the change of the rainfall bonus

$$30 \cdot (1 - \mathrm{RE}(r, r^*)),$$

thus rewarding smaller relative errors.

- To his relief, the rainfall forecasts improved, but after a few weeks he started getting even more preposterous forecasts: stations were reporting 0% chance of rain with 50 mm rainfall.

The meteorologists' story

- Exasperated, Ivan was determined to solve the mystery of these bogus forecasts. Comparing the data from each of the three months, he finally figured out what was going on.
- Starting with probabilities, he put himself in the shoes of the stationmasters and imagined that the chance of rain was 20%.
- On average, he would be paid (if he reported his true belief)

$$50 \cdot 0.2^2 + 50 \cdot 0.8^2 = 50 \cdot 0.68 = 34\,€.$$

- On the other hand, if he reported 0% chance of rain, he would be paid

$$(50 \cdot 0)0.2 + (50 \cdot 1)0.8 = 40\,€.$$

- Notice that in the above argument it is immaterial whether $p = 0.2$ is the true probability of rain or not as each stationmaster's decision whether to be honest or not is based on his subjective probability of rain.

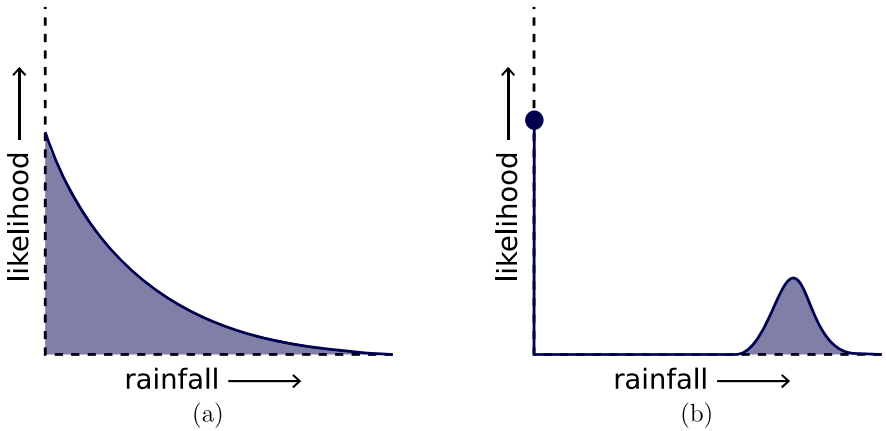
The meteorologists' story

- Ivan realized that the linear score does not incentivize the stationmasters to report honestly their true beliefs about the probability of rain.
- Similar thinking told him that the zero-one function, $\mathbb{1}_{\{|r-r^*|<1\}}(r, r^*)$, is optimized at the mode of the predictive distribution.
- Indeed, one may easily compute that

$$\mathbb{E}\mathbb{1}_{\{|r-R^*|<1\}}(r, R^*) \approx \mathbb{P}(R^* \in (r-1, r+1)),$$

  where R^* is the random variable giving the quantity of rainfall.
- It can be checked that the relative error is optimized at the median of the random variable with probability density proportional to $yp(y)$ if the predictive distribution has density $p(y)$.

The meteorologists' story



- Diagram (a) corresponds to a case where the meteorologists are highly uncertain of the prospects of rain and we have mode$(R^*) = 0$, $\mathbb{P}(R^* > 0) > 0.5$. Diagram (b) corresponds to a case of unlikely storm with heavy rainfall and $\mathbb{P}(R^* = 0) > 0.5$.
- The counterintuitive forecast 100% chance of rain with 0 mm rainfall occurred when the meteorologists thought that the probability of rain was above 0.5 but the mode of R^* was 0.
- The counterintuitive forecast 0% chance of rain with 50 mm rainfall occurred when the meteorologists thought that there was a small probability of a heavy storm.

The meteorologists' story

- Can Ivan choose a scoring rule that incentivizes the meteorologists to report honestly their true beliefs about the probability of rain?
- Well, he remembered a property of the mean value known by Carl Friedrich Gauss:

$$\mathbb{E}Y = \arg\min_{x\in\mathbb{R}} \mathbb{E}(x - Y)^2. \tag{2}$$

- Let $Y = \mathbb{1}_{\{\text{rain}\}}$ and let I be the information set of a forecaster when making the forecast p. Suppose that $\mathbb{E}(Y|I) = q_I$.
- We have

$$\begin{aligned}\mathbb{E}((p - Y)^2|I) &= p^2 - 2p\mathbb{E}(Y|I) + \mathbb{E}(Y^2|I)\\ &= (p - q_I)^2 + q_I - q_I^2\\ &= (p - q_I)^2 + \text{Var}(Y|I),\end{aligned} \tag{3}$$

which shows (2) in a special context as $p = q_I$ is the minimizer of (3).

The meteorologists' story

- Hence, to pay for the probability forecasts, Ivan chose the score

$$\begin{aligned} S(p, y) &= 100 - 50(p - y)^2 \\ &= 100 - 50(p - \mathbb{1}_{\{\text{rain}\}})^2. \end{aligned}$$

- The expected score of a forecaster with information set I is given by

$$\mathbb{E}(S(p, Y)|I) = 100 - 50\,\mathrm{Var}(Y|I) - 50(p - q_I)^2. \tag{4}$$

- The term $(p - q_I)^2$ measures the calibration of the forecast. When $p = q_I$ the forecast is called calibrated.
- In other words, a forecast predicting rain with probability p is calibrated if the observed relative frequency of rainy days is equal to p.
- The term

$$\mathrm{Var}(Y|I) = \mathbb{E}((Y - \mathbb{E}(Y|I))^2|I)$$

measures the residual variance if we use $\mathbb{E}(Y|I) = q_I$ to predict Y.

The meteorologists' story

- If the forecaster has a sufficiently long archive of predictions and subsequent observations, he may calibrate his forecasts.
- To that end, for every I, he must compute the corresponding relative frequency q_I of rainy days from that archive and report $p = q_I$.
- Suppose we have two calibrated forecasts, $q_1 = 0.7$ and $q_2 = 0.9$, based on information sets I_1 and I_2, respectively. Since the residual variance of the first forecast is bigger than that of the second forecast,

$$\mathrm{Var}(Y|I_1) = q_1(1 - q_1) = 0.21,$$
$$\mathrm{Var}(Y|I_2) = q_2(1 - q_2) = 0.09,$$

the forecast $q_2 = 0.9$ is more informative than the forecast $q_1 = 0.7$.

The meteorologists' story

- The scoring rule

$$S(p, y) = 100 - 50(p - y)^2$$

  is called proper because it assigns the optimal expected score to the true, calibrated forecast relative to any information set.
- The expected score of every proper scoring rule (PSR) combines two measures for predictive performance - the degree of calibration of the forecast and the residual variance of the predicted variable around the forecast - and gives their sum as a measure of total accuracy.
- We recall that the decomposition for S into entropic term and calibration term is given by

$$\mathbb{E}(S(p, Y)|I) = 100 - 50\,\mathrm{Var}(Y|I) - 50(p - q_I)^2.$$

The meteorologists' story

- The final score on which Ivan settled to pay the meteorologists was

$$100 - 50(p - \mathbb{1}_{\{\text{rain}\}})^2 - 30(r - r^*)^2.$$

- The bonus term $-30(r - r^*)^2$ incentivizes the meteorologists to report the mean value of their predictive distribution for R^*.
- Ivan could have used the zero-one score or the relative error instead of the quadratic error above. In this case he would have been incentivizing the meteorologists to report the mode or the skewed median of their predictive distribution for R^*, respectively.
- What would prevent the occurrence of counterintuitive forecasts as before is the fact that Ivan now uses a proper scoring rule to evaluate the probabilistic forecasts.

Formal definitions

- Let Y be a random variable taking values in $\mathcal{Y} \subset \mathbb{R}^n$ and let $\mathcal{P}$ be a convex class of probability distributions on $\mathcal{Y}$.
- A scoring rule is a mapping $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$.
- For a prediction $P \in \mathcal{P}$ and a subsequent realization $y \in \mathcal{Y}$ of Y, S gives a score $S(P, y)$, which is a negative or positive incentive, depending on the orientation of S.
- The expected score of a forecast $P \in \mathcal{P}$ under the true distribution $Q \in \mathcal{P}$ of Y is given by

$$S(P, Q) := \mathbb{E}_{Y \sim Q} S(P, y).$$

- A (negatively oriented) scoring rule S that minimizes its expected score,

$$S(Q, Q) = \min_{P \in \mathcal{P}} S(P, Q), \tag{5}$$

at the true distribution $Q \in \mathcal{P}$ is called proper. If the true distribution is always a unique minimizer, S is called strictly proper.

Formal definitions

- The function $\Phi : \mathcal{P} \to \mathbb{R}$ given by $\Phi(P) = S(P, P)$, for every $P \in \mathcal{P}$, is the entropy function associated with a strictly proper scoring rule S.
- The entropy Φ is a concave function as a pointwise minimum of linear functionals.
- In fact, Φ is strictly concave if and only if S is strictly proper.
- Every strictly proper scoring rule S defines a divergence measure $D : \mathcal{P} \times \mathcal{P} \to \mathbb{R}$ given by

$$D(P, Q) = S(P, Q) - S(Q, Q), \quad P, Q \in \mathcal{P}. \tag{6}$$

- $D(P, Q)$ is a positive definite function, that is, $D(P, Q) \geq 0$ for every $P, Q \in \mathcal{P}$, and $D(P, Q) = 0$ if and only if $P = Q$.
- $D(P, Q)$ being a divergence (positive definite) is equivalent to S being strictly proper.

Characterization of proper scoring rules

Notice that $S(P, y)$ may be considered as a linear functional on the vector space span $\mathcal{P}$ acting as follows:

$$Q \to S(P, Q) = \int S(P, y) dQ(y),$$

for every $Q \in \operatorname{span} \mathcal{P}$.

Theorem (Characterization)

Let $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$ be a scoring rule and consider the function $\Phi : \mathcal{P} \to \mathbb{R}$ given by $\Phi(P) = S(P, P)$. Extend Φ to cone $\mathcal{P}$ *as a 1-homogeneous functional,*

$$\Phi(\lambda \mathcal{P}) = \lambda \Phi(P), \qquad \text{for every } P \in \mathcal{P} \text{ and every } \lambda > 0.$$

Then S is (strictly) proper if and only if S is a (strict) supergradient of Φ on cone $\mathcal{P}$.

Score divergences

- Every score divergence, that is, the divergence induced by a PSR is a special case of a Bregman divergence.
- Bregman divergences have been introduced by Lev M. Bregman in 1967 in connection to convex optimization as generalizations to the Euclidean metric.

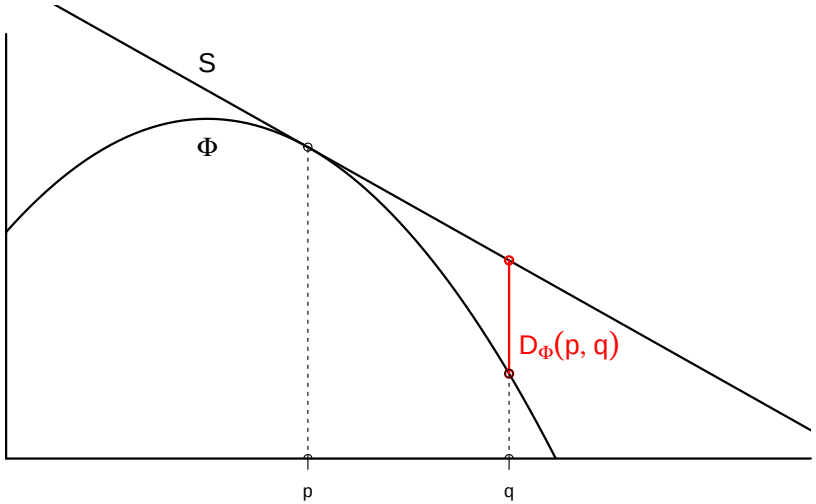


Figure: The Bregman divergence is the vertical distance between a supergradient of a concave functional and the concave functional.

Examples of proper scoring rules - the logarithmic score

- One of the best known PSRs is the logarithmic score,

$$S(p, y) = -\ln p(y),$$

where p is the probability mass function or probability density of the distribution P of the random variable Y.

- The associated entropy,

$$\Phi(p) = -\int p(y) \ln p(y) dy,$$

is the celebrated Boltzmann-Shannon entropy.

- The associated divergence is the celebrated Kullback-Leibler divergence

$$D(p, q) = \int q(y) \ln \frac{q(y)}{p(y)} dy.$$

Examples of proper scoring rules - the logarithmic score

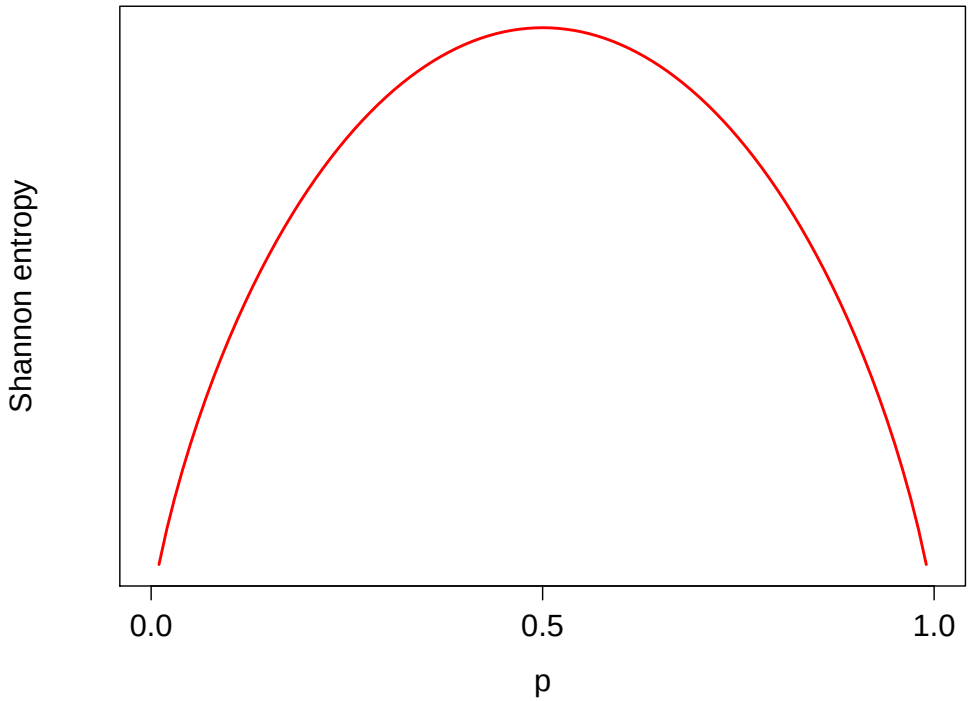


Figure: Shannon entropy for a binary variable.

Examples of proper scoring rules - the Brier score

- Let Y be a categorical random variable with probability mass function $p = (p_1, \ldots, p_n)$. We identify Y with the vector variable $Y = (Y_1, \ldots, Y_n)$, where Y_i is the characteristic function of the i-th event.
- The Brier score is defined as

$$S(p, y) = \sum (p_i - y_i)^2.$$

- We have

$$\begin{aligned}\Phi(p) &= \mathbb{E}_{Y \sim p} \sum (p_i - y_i)^2 = \mathbb{E}_{Y \sim p} \sum (p_i^2 - 2 p_i y_i + y_i^2) \\ &= \sum (p_i^2 - 2 p_i^2 + p_i) = -\sum p_i^2 = 1 - \|p\|_2^2 .\end{aligned}$$

- Similarly, we have that

$$D(p, q) = \|p - q\|_2^2 .$$

Predicting asset prices

- We show a simulation study of a highly volatile daily asset value, y_t.
- The data generating process is such that the y_t is a realization of the random variable

$$Y_t = Z_t^2,$$

where Z_t follows a conditionally heteroscedastic Gaussian time series model

$$Z_t \sim N(0, \sigma_t^2), \quad \text{where } \sigma_t^2 = 0.20 Z_{t-1} + 0.75\sigma_{t-1}^2 + 0.05.$$

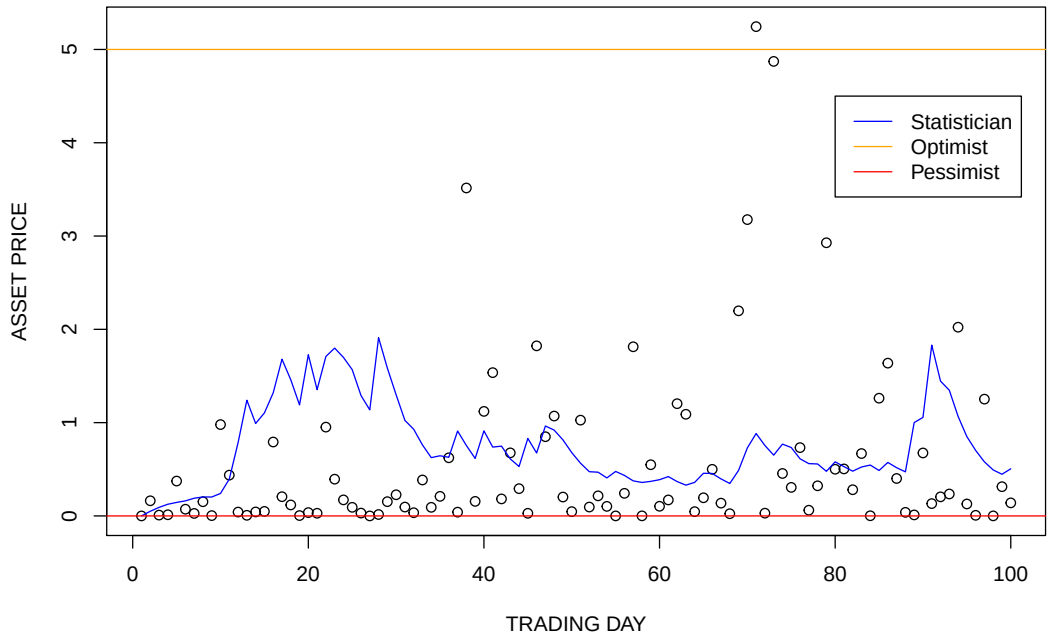
- We consider three forecasters, each of whom issues a one-day ahead point forecast for the asset value.
- The statistician has knowledge of the data-generating process and thus predicts the true conditional mean,

$$x_t = \mathbb{E}(Y_t|\sigma_t^2) = \sigma_t^2$$

as his point-forecast. (The above relation follows from the fact that $\sigma_t^2 Y_t \sim \chi^2(1)$.)

Predicting asset prices

- The optimist always predicts $x_t = 5$.
- The pessimist always issues the point forecast $x_t = 0.05$.
- The figure below shows the data realized for 100 successive trading days, along with the three forecasts.



Predicting asset prices

- Remember that the $\chi^2(1)$ distribution is highly skewed and its median is approximately equal to 0.47 times its mean value. This explains why there are more observations below the blue line than above it.
- Without a doubt, the statistician is the most skilled forecaster of the three.
- We are going to compare the predictive performance of the three forecasters under the scoring functions given below.

Table: Some commonly used scoring functions.

$S(x, y) = (x - y)^2$	squared error (SE)
$S(x, y) = \lvert x - y \rvert$	absolute error (SE)
$S(x, y) = \lvert x - y \rvert / y$	absolute percentage error (APE)
$S(x, y) = \lvert x - y \rvert / x$	relative error (RE)

Predicting asset prices

The table below provides a formal evaluation of the three forecasters for a sequence of $n = 100,000$ sequential forecasts using different scoring functions.

Table: The mean errors of the three forecasters in our simulation study.

Forecaster	SE	AE	APE	RE
Statistician	5.07	0.97	258000	0.97
Optimist	22.73	4.35	139600	0.87
Pessimist	7.61	0.96	14000	19.24

The results are disconcerting and counterintuitive in that the pessimist has the best (lowest) score both under the absolute error and the absolute percentage error. In terms of relative error, the optimist performs best.

Yet, what we have done here is common practice in academia and business, in that point forecasts are evaluated by means of these scoring functions.

Predicting asset prices

The source of these disconcerting results is explained in a recent article by Engelberg, Manski, and Williams (2009, p. 30):

"Our concern is prediction of real-valued outcomes such as firm profit, GDP, growth, or temperature. In these cases, the users of point predictions sometimes presume that forecasters report the means of their subjective probability distributions; that is, their best point predictions under square loss. However, forecasters are not specifically asked to report subjective means. Nor are they asked to report subjective medians or modes, which are best predictors under other loss functions. Instead, they are simply asked to 'predict' the outcome or to provide their 'best prediction', without definition of the word 'best'. In the absence of explicit guidance, forecasters may report different distributional features as their point predictions. Some may report subjective means, others subjective medians or modes, and still others, applying asymmetric loss functions, may report various quantiles of their subjective probability distributions."

Formal definitions

- Let $\mathcal{X}, \mathcal{Y}$ be measurable subsets of $\mathbb{R}^n$ and $\mathcal{Y}$ be the range of a random variable Y.
- If $\Delta(\mathcal{Y})$ is the set of all probability measures on $\mathcal{Y}$, consider a mapping $\Gamma : \Delta(\mathcal{Y}) \to \mathcal{X}$.
- For example, Γ may be the mean value, the median, the variance of a probability measure P, or a any other vector-valued property of P.
- A scoring function is a mapping $S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$.

Formal definitions

Definition

A scoring function S is consistent for a statistical functional $\Gamma : \Delta(\mathcal{Y}) \to \mathcal{X}$ if

$$\Gamma(P) \in \arg\min_{x \in \mathcal{X}} \mathbb{E}_{Y \sim P} S(x, y),$$

for each $P \in \Delta(\mathcal{Y})$. Moreover, S is strictly consistent for Γ if $\Gamma(P)$ is the unique minimizer of the above expression, that is,

$$\Gamma(P) = \arg\min_{x \in \mathcal{X}} \mathbb{E}_{Y \sim P} S(x, y),$$

for each $P \in \Delta(\mathcal{Y})$.

Definition

A statistical functional for which there is a (strictly) consistent scoring function is called (strictly) elicitable.

Some examples

- The mean value is strictly elicitable with a strictly consistent scoring function: the squared error, $S(x, y) = (x - y)^2$.
- The median is strictly elicitable with a strictly consistent scoring function: the absolute error, $S(x, y) = |x - y|$.
- The mode is strictly elicitable with a strictly consistent scoring function: the zero-one function, $S(x, y) = \mathbb{1}_{\{|x-y|>\epsilon\}}$, where $\epsilon > 0$.
- The higher moments, the weighted median are also strictly elicitable.
- The variance is not elicitable.

Conclusion

- PSRs measure the distance between probability distributions through their associated score divergences

$$D(P, Q) = S(P, Q) - S(Q, Q).$$

- Minimizing the divergence $D(P, Q)$ in P is the same as minimizing the expected score $S(P, Q)$ in P.
- The class of divergences PSRs generate is a subset of the so called Bregman divergences, a fundamental concept in convex optimization.
- The expected score of a forecast gives a combined measure of the degree to which the forecast is calibrated and the residual uncertainty of the predicted variable around the forecast.
- Unlike divergences, PSRs may be implemented as payment schemes to forecasts as they give an individual score to each forecast and subsequent observation.