Scoring Rules

Jason Konek[a] Ben Levinstein[b]

19 August 2022

[a]Department of Philosophy
Bristol

[b]Department of Philosophy
UIUC

Scoring rules can be thought of as:

- Tools to elicit credences
- Tools to evaluate forecasts
- Generalized loss functions
- Generalized information measures
- Measures of the **expected disutility** of a forecast.

Now: a primer on one way to construct scoring rules.

Will help us understand:

- What they are
- How they're useful
- How to pick a good one

Scoring rules are often thought of as measures of inaccuracy:

- Quantify divergence from truth.
- The higher the probability assigned to actually true events, the better the score.

Local Scoring Rule

A function $G : [0,1] \times \{0,1\} \to [0,\infty]$ is a **local scoring rule** if $g(\cdot, 1)$ and $g(\cdot, 0)$ are monotonically decreasing and increasing respectively.

Examples

$$\begin{aligned} abs(x,i) &= |i - x| \\ br(x,i) &= (i - x)^2 \\ log(x,i) &= -\ln(|1 - i - x|) \\ sph(x,i) &= -|1 - i - x|/(x^2 + (1 - x)^2)^{1/2} \end{aligned}$$

We also want to quantify how inaccurate a bunch of forecasts are as a whole.

- Easiest to start with local scoring rules and add them up.
- But more generally, if we're measuring inaccuracy, we want to ensure that sets of forecasts that assign uniformly higher probability to truths get better scores.

Weak Truth-Directedness

A function $\mathcal{I}$ from a set of probability functions and states to $[0, \infty]$ is **truth-directed** if $|P(X) - \omega(X)| \leq |P'(X) - \omega(X)|$ for all $X \in \mathcal{F}$ then $\mathcal{I}(P, \omega) \leq \mathcal{I}(P', \omega)$.

Global Scoring Rule

$\mathcal{I}$ is a **global scoring rule** if it is truth-directed.

Examples

$$
\begin{aligned}
Abs(\mathrm{Pr}, \omega) &= \sum_{X \in \mathcal{F}} |P(X) - \omega(X)| \\
Euc(\mathrm{Pr}, \omega) &= \left(\sum_{X \in \mathcal{F}} (P(X) - \omega(X))^2\right)^{1/2} \\
Br(\mathrm{Pr}, \omega) &= \sum_{X \in \mathcal{F}} (P(X) - \omega(X))^2
\end{aligned}
$$

Some scoring rules have a special property called **propriety.**

- For a proper scoring rule, each probability function expects itself to do best.
- One interpretation: if rewarding people based on their inaccuracy, proper measures **incentivize honesty.**

Propriety

A scoring rule $\mathcal{I}$ is **proper** if for any probability functions Pr, Pr′:

$$E_P(\mathcal{I}(P, \cdot)) \leq E_P(\mathcal{I}(P', \cdot))$$

$\mathcal{I}$ is **strictly proper** if the inequality is strict.

The Brier Score $\sum(x - i)^2$ is strictly proper.

But the Euclidean and Absolute Value scores are not.

Exercise

Show the Brier Score is strictly proper and the Absolute Value score is improper.

In philosophy, this is kind of a problem.

- Big goal: derive fundamental norms (like probabilism, conditionalization, principle of indifference, Principal Principle) based purely on the goal of **accuracy** along with some decision-theoretic norm.
- Most accuracy arguments need the measures to be strictly proper.
- But there aren't great independent arguments for propriety.

Some Properties

With strictly proper rules, you can **elicit** credences.

- Charge $\$\mathrm{Brier}(x, i)$ based on announced forecast and actual outcome.

Different scores measure different types of 'goodness' of forecasts.

	A	B	C	D
Alice	.005	.275	.230	.490
Bob	.033	.127	.137	.703

Figure: Alice and Bob's credences that any particular ball will be drawn. If A is drawn, Brier prefers Alice, but Log prefers Bob.

Schervish Construction

We'll now look at a different reason to think of proper scoring rules as special.

- Your scoring rule depends on which practical decisions you expect to make.
- I.e., encode expectations about decisions you'll be making.
- Uncertainty over the nature of these decisions determines which scoring rule represents you.
- Roughly: The inaccuracy of a credence of x in X when X is true (false) is the **expected disutility** of having a credence of x given that X is true (false).

Your evaluation of a different forecast y is (a function of) the expected disutility of using **that** forecast and **your** preferences to make decisions.

- How well off I expect to be if I set my credence in X equal to y
- Kept my utility function the same.

Alice must decide whether to take an umbrella.

- If it rains, it’s better to have it.
- If it’s dry, it’s better to leave it at home.

Assume Alice will maximize expected utility.

Suppose Alice's utility function is:

	Rain	No Rain
Umbrella	-1	-2
No Umbrella	-4	0

Note that all that **all** that goes into determining how much utility Alice actually gets:

- Whether it rains.
- Whether x is $>$ or $\leq 2/5$.

In particular, a credence of .5 will result in the same outcomes as a credence of .9.

To make this problem easier to work with, we make three changes:

1. We represent Alice as **minimizing expected loss** instead of maximizing expected utility.
2. We **normalize** the problem so that the loss of the better action at each state of the world is 0.
3. We rewrite the problem by **dividing out** the sum of the possible losses

	Rain	No Rain
Umbrella	0	$2/5 \cdot 5$
No Umbrella	$(1 - 2/5) \cdot 5$	0

For now, we can even forget about the 5 and make the potential losses sum to 1.

Suppose Alice is assessing the expected loss of using a possibly alternative credence y along with her utility function to decide whether to bring an umbrella.

- All that matters is how likely it is y will lead Alice into the wrong decision and how bad making the wrong decision would be.

- EL(x) is weakly increasing with expected inaccuracy.
- The scoring rule

$$g_1(x) = \begin{cases} 3/5 & x \leq 2/5 \\ 0 & x > 2/5 \end{cases}$$

$$g_0(x) = \begin{cases} 0 & x \leq 2/5 \\ 2/5 & x > 2/5 \end{cases}$$

is merely proper.

Under the assumption that Alice is an EL-minimizer, we can reformulate the problem so that:

- $L(d_1, X) = L(d_0, \bar{X}) = 0$
- For some $q \in [0, 1], W \in (0, \infty]$:
  - $L(d_1, \bar{X}) = q \cdot W$
  - $L(d_0, X) = (1 - q) \cdot W$

We'll be ignoring W for a while.

- Alice will perform d_1 just in case the forecast probability for X is $> q$.
- Knowing q is then sufficient to characterize the problem!
- Call any 2-Decision problem such that $L(d_1, \bar{X}) = W \cdot q$ a **q-problem**.

This is actually very general:

- Can reduce compound gambles.

For any particular value of q, Alice sees no difference between two forecasts on the same side of q.

$$g_1(x) = \begin{cases} 1 - q & x \leq q \\ 0 & x > q \end{cases}$$

$$g_0(x) = \begin{cases} 0 & x \leq q \\ q & x > q \end{cases}$$

So far, Alice has known the value of q.

- But you don't always know which decision problems you'll end up facing.
- So we now need to account for uncertainty about the value of q.

There are two factors that need to be taken account of:

1. First, the probability density that she'll face a q-problem for any particular q.
2. How important such problems are expected to be relative to one another.

Suppose Alice knows she'll face either Q1 or Q2, with $P(Q1) = P(Q2) = .5$.

Q1

	X	$\bar{X}$
d_1	0	$1/2$
d_0	$1/2$	0

Q2

	X	$\bar{X}$
d_1^*	0	$2/3 \cdot 15$
d_0^*	$1/3 \cdot 15$	0

Much more important she make the right decision in Q2 than in Q1.

In the finite case (where she knows she'll face one of finitely many problems), Alice's **expected loss conditional on** X is:

$$g_1(x) = \sum_{x \leq q} (1 - q) \cdot \mathrm{E}(W \mid q) \cdot P(q)$$

And on $\neg X$:

$$g_0(x) = \sum_{q < x} q \cdot \mathrm{E}(W \mid q) \cdot P(q)$$

So, her overall expected loss is: $x g_1(x) + (1 - x) g_0(x)$

We can think of $E(W \mid q)P(q)$ as the **expected importance** of having her credence on the right side of q.

- In the continuous case, this becomes $m(q) := E(W \mid q) \cdot f(q)$, where f is her density function.
- Will refer to this as her **support function.**

Simplified Theorem (Schervish 1989)

Let $m(q) := E(W \mid q) \cdot f(q)$ be a support function with

$$g_1(x) = \int_x^1 (1 - q) m(q) \, \mathrm{d}q$$
$$g_0(x) = \int_0^x q m(q) \, \mathrm{d}q$$

Then $G = (g_1, g_0)$ is a proper scoring rule. Furthermore, if m is non-degenerate ($m(x) > 0$ a.e.), then G is strictly proper.

Theorem (Schervish 1989)

Let $G = (g_1, g_0)$ be a left-continuous scoring rule such that $g_i(j) = \lim_{t\to j}(t)$ for $i, j = 0, 1$ and having $g_1(1)$ and $g_0(0)$ finite. G is proper iff there exists a measure μ on $[0, 1)$ such that

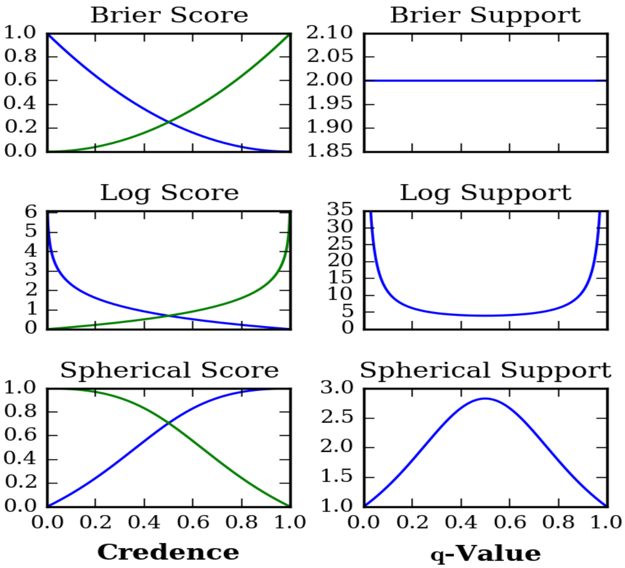
$$g_1(x) = g_1(1) + \int_x^1 (1 - q)\,\mathrm{d}\mu$$

$$g_0(x) = g_0(0) + \int_0^x q\,\mathrm{d}\mu$$

for all x. G is strictly proper iff μ gives a positive measure to every non-degenerate interval.

g_i measures the expected loss of using a forecast x given that X's truth-value is i.

- For any rational agent, G will always be proper.
- (Roughly): If the agent thinks that any bet is at least possible, then she'll use a *strictly* proper rule to measure her expected loss.


Brier Score
Brier Support
Log Score
Log Support
Spherical Score
Spherical Support
Credence
q-Value

Exercise

Construct a strictly proper scoring rule such that $g_0(x) = x + c$ for some constant c.

Benefits of This Approach

Viewing scoring rules this way allows us to tailor them to the views you have about the problems you might face.

Suppose you are a doctor with credence .7 that a bacterium is Graham-negative.

- What is the expected disutility of this credence?
- Well, you don't yet have any real actions that ride on it.
- But there's a chance there could be such an action in the future.

Whatever that decision will be, it will be some q-problem.

Suppose you think:

- The stakes are exponentially distributed and independent of q, so $\mathrm{E}(W) = \lambda$.
- $f(q)$ is constant.

Together, these imply that $m(q) = c$ for some constant c, which we'll let $= 2$. So,

$$g_1(x) = 2\int_x^1 (1 - q)\,\mathrm{d}q$$
$$g_o(x) = 2\int_0^x q\,\mathrm{d}q$$

So, $G = (i - x)^2$, which is the **Brier Score.**

Other scoring rules are tailored to different problems.

- Sometimes the difference between a .9 and .99 and .999 credence will, in expectation, matter quite a lot.
- E.g., in buying insurance.
- In that case, a logarithmic scoring rule, which has more support near the ends of the spectrum may be more appropriate.
- $m(q) = 1/(q(1-q))$

Other times, it might matter quite a bit that your credence is on the right side of .5.

Generalized Entropy and the Value of Information

The self-expected score of any strictly proper role is a **generalized entropy function.**

Log $-\sum P(x_i)\ln(x_i)$

Brier $\sum 1 - P(x_i)^2$

Different notions of information and entropy will lead to different kinds of information seeking activities:

- Which data to gather
- Which experiment to design

This makes sense–information is valuable insofar as it is useful for expected future action.

Let $P^{\mathcal{E}}$ be a random object denoting your (as-yet-unknown) credence after performing an experiment $\mathcal{E}$.

The **value** of performing the experiment is:

$$\mathrm{Val}(\mathcal{E}) = E(G(P)) - E(G(P^{\mathcal{E}}))$$

In general, this will change with which rule you use.

However, some experiments will be better than others regardless of which rule you use.

- Polling 10 random people.
- Polling 10,000 random people.

In this case, you expect your credences after performing the second experiment will be more accurate than your credences after performing the first on every SPSR.

- Superior regardless of your practical interests.
- Compare: Blackwell's Theorem

Can also compare forecasters.

- A is superior to B according to P if A gets a better score in expectation than B on a particular rule.
- A is superior to B according to P if A gets a better score on **every SPSR**.

Let X be some event, and $A = x$ mean Alice assigns probability x to X.

You **reflect** Alice if for all x, $P(X \mid A = x) = x$.

Unsurprisingly, if you reflect Alice, you expect her to do better than you on all SPSRs.

But the converse does not hold.

You **simply trust** Alice if for all x, $P(X \mid A \geq x) \geq x$ and $P(X \mid A \leq x) \leq x$.

Simply trusting someone is equivalent to expecting her to do better on all SPSRs.

Compare what P thinks *of itself* to what it thinks of A (some unknown forecast).

- Doing better on every SPSR requires an intermediate form of deference.

A	$P(X \mid A = x)$	$P(A = x)$
1	1	1/15
.75	.7	1/3
.25	.25	2/5
0	0	1/5

Figure: P simply trusts A.

A	$P(X \mid A = x)$	$P(A = x)$
.75	.7	1/2
.25	.3	1/2

Figure: *P* does not simply trust *A*.

Exercise

Construct scoring rules where:

1. P expects A to be less inaccurate than P.
2. P expects A to be more inaccurate than P.

Scoring rules:

- Lots of applications
- Useful to think of as generalized loss functions encoding expectations about which decisions you'll be making.
- Thereby used to generalize Shannon information to determine best use of information-gathering resources.