Conformal Scoreline Sets

A set of outcomes that holds the result 90% of the time, whatever the model gets wrong.

Split conformal prediction applied to a scoreline grid. The promise is distribution-free: it needs the future to look like the calibration window, and nothing else.

SPORT
Any

FAMILY
The machine layer

LICENCE
£499
one-off, perpetual

RUNNING IN THE PLATFORM
https://edgeworthарena.com/arena/api/ml/match?home=arsenal&away=chelsea&comp=E0

Inputs

NAME	TYPE	UNITS	WHERE IT COMES FROM
calibration	list of (score matrix priced before the match, actual score)	-	a walk-forward run of any scoreline model
alpha (optional)	float	-	your risk level; 0.1 gives a 90% set
window (optional)	int	matches	how far back to calibrate; 380 is a season of a big league

Outputs

NAME	TYPE	UNITS	NOTES
q_hat	float	nats	the nonconformity quantile
set	list of scorelines	-	
mass	float	probability	the model's own probability for the set, which is not the coverage
coverage	float	fraction	measured out of sample

Method

1 Score each calibration match by its nonconformity, −log P(result), under the price it was given before kick-off.

THIS IS THE SAMPLE
The method runs to 3 steps; one is shown. The assumptions, the limits, the module layout and the reference implementation come with the licence. Everything on the page below this -- what the model was measured at -- is the same in both versions, because that is the part you should not have to take on trust.

What it was measured at

Walk-forward across every covered league, calibrated on a rolling window and checked on the matches after it.

MEASURE	RESULT	READING
Leagues	1	
Calibration window	380 matches	
90% promised	held 90.3%	average set of 13.7 scorelines out of 49
80% promised	held 80.8%	average set of 10.3 scorelines out of 49
67% promised	held 68.2%	average set of 7.1 scorelines out of 49

Assumptions

What has to be true for the numbers to mean what they say.

- Exchangeability: the coming match is drawn from the same distribution as the calibration window. A rule change or a mid-season shift breaks it, which is what the changepoint detector is for.

Limits

Where it is known to be wrong. Published because a model without stated limits is one nobody has pushed on.

- It gives coverage, not sharpness. A badly calibrated model gets a huge set, not a broken promise.