Table of Contents

Measuring Predictive Model Performance

Time

Measuring Predictive Model Performance

Time

$m_1(\boldsymbol{x})$ better than $m_2(\boldsymbol{x})$?

Measuring Predictive Model Performance

Time

Stricly Consistent
Scoring Function S

$m_1(\boldsymbol{x})$ better than $m_2(\boldsymbol{x})$?

Notation

- **Observation domain** O, which comprises the potential outcomes of a future observation.
- Convex class $\mathcal{F}$ of probability measures on the observation domain O (equipped with a suitable σ-algebra), which constitutes a family of probability distributions for the future observation.
- **Action domain** A, which comprises the potential actions of a decision maker.
- **Scoring or loss function** $S : \mathrm{A} \times \mathrm{O} \to \mathbb{R}$, where $S(a, o)$ represents the monetary or societal cost when the decision maker takes the action (or point forecast) $a \in \mathrm{A}$ and the observation $o \in \mathrm{O}$ materialises.
- A **scoring rule** is a function $\mathbf{S} : \mathcal{F} \times \mathrm{O} \to \mathbb{R}$, $\mathbf{S}(F, o)$ is the penalty for probabilistic prediction $F \in \mathcal{F}$ and observation o.
- Statistical functional $T : \mathcal{F} \to \mathrm{D}$, potentially set-valued.

Assumption

Common domain $\mathrm{D} = \mathrm{O} = \mathrm{A} \subseteq \mathbb{R}^d$,

Note: We could have chosen $S : \mathrm{A} \times \mathrm{O} \to [0, \infty)$, w.l.o.g.

Scoring Functions

Repetition

- A scoring function S measures the deviation of the model prediction $m(\boldsymbol{X})$ from T using observations Y: $S(m(\boldsymbol{X}), Y)$.
- Convention: The smaller S, the better.

Purpose

- Assess predictive performance of the predictions of a model.
- Compare the predictiveness of different models.

Scoring rules

A scoring rule $\mathbf{S}$ is in principle the same as a scoring function but for probabilistic predictions: model goal is F (or pdf f).

Model Comparison

We estimate the expected score $\mathbb{E}[S(m(\boldsymbol{X}), Y)]$, earlier called statistical risk $R(m)$, as

$$\overline{S}(m; D) = \frac{1}{n} \sum_{(\boldsymbol{x}_i, y_i) \in D} S(m(\boldsymbol{x}_i), y_i) \,, \tag{10}$$

which we called empirical risk before. Model m_A is deemed to have an inferior predictive performance than model m_B in terms of the score S (and on the sample D) if

$$\overline{S}(m_A; D) - \overline{S}(m_B; D) = \frac{1}{n} \sum_{(\boldsymbol{x}_i, y_i) \in D} S(m_A(\boldsymbol{x}_i), y_i) - S(m_B(\boldsymbol{x}_i), y_i) > 0. \tag{11}$$

Statistical test

- With i.i.d. data, simple t-test.
- With (time / serial) correlation, Diebold-Mariano test.

Consistency & Elicitability

Definition (Consistency)

Let $\mathcal{F}$ be a class of probability distributions where the functional T is defined on. A scoring function $S(z, y)$ is a function in a forecast z and an observation y. It is $\mathcal{F}$*-consistent* for T if

$$\int S(t, y) \,\mathrm{d}F(y) \leq \int S(z, y) \,\mathrm{d}F(y) \qquad \text{for all } t \in T(F),\ z \in \mathrm{D},\ F \in \mathcal{F}. \tag{12}$$

The score is *strictly* $\mathcal{F}$-consistent for T if it is $\mathcal{F}$-consistent for T and if equality in (12) implies that $z \in T(F)$.

Definition (Elicitability)

A functional T is *elicitable* on F if there is a strictly $\mathcal{F}$-consistent scoring function for it.

Why Consistency Matters?

Consistency

- It ensures that we get what we want: $m^\star = T(Y|\boldsymbol{X})$.
- At least in the large sample limit (Law of Large Numbers arguments).
- Compare with a repeated game where each forecaster gets penalty / loss $S(z, y)$.

Counter example: Use of absolute error $|z - y|$ when we aim for the expectation.

Elicitability

- Tells us if there exists a consistent scoring function for the functional T.
- Model comparison and backtesting is (partially) pointless for non-elicitable T.

Counter examples: Mode (for general F), variance (alone) and expected shortfall (alone) are not elicitable.

Proper Scoring Rules

Definition (Propriety)

The scoring rule **S** is *proper* relative to the class $\mathcal{F}$ if

$$\mathbb{E}_G[\mathbf{S}(G, Y)] \leq \mathbb{E}_G[\mathbf{S}(F, Y)]$$

for all $F, G \in \mathcal{F}$. It is *strictly proper*, if equality holds (if and) only if $F = G$.

Theorem (Gneiting 2011 Theorem 3)

Suppose that the scoring function S is $\mathcal{F}-$consistent for the functional T. Foreach $F \in \mathcal{F}$, let $t_F \in T(F)$. Then $\mathbf{S}(F, y) = S(t_F, y)$ is a proper scoring rule relative to $\mathcal{F}$.

Examples of proper scoring rules

- quadratic score $\mathbf{S}(F, y) = -2f(y) + \int f^2(x)\,\mathrm{d}x$
- logarithmic score $\mathbf{S}(F, y) = -\log f(y) \Rightarrow$ compare MLE
- continuous ranked probability score (CRPS)
  $\mathbf{S}(F, y) = \int (F(x) - \mathbb{1}\{y \leq x\})^2\,\mathrm{d}x = \mathbb{E}_F[|Y - y|] - \frac{1}{2}\mathbb{E}_F[|Y - Y'|]$, $Y, Y' \overset{\text{i.i.d}}{\sim} F$
- Dawid-Sebastiani score $\mathbf{S}(F, y) = \frac{(y-\mu_F)^2}{\sigma_F^2} + 2\log\sigma_F$ with $\mu_F = \mathbb{E}_F[X]$, $\sigma_F^2 = \mathrm{Var}_F[X]$

Order Sensitivity

Given a one-dimensional, real-valued T ($\mathrm{D} \subseteq \mathbb{R}$).

Definition

T is $\mathcal{F}$-*order sensitive* if for any $F \in \mathcal{F}$ and any $z_1, z_2 \in \mathrm{A}$ with either $z_1 > z_2 > T(F)$ or $z_1 < z_2 < T(F)$ one has $\mathbb{E}_F[S(z_1, Y)] > \mathbb{E}_F[S(z_2, Y)]$.

Implications

- Order sensitivity of S implies consistency.
- (Under weak regularity conditions:) Strict consistency of S implies order sensitivity.

Note: For T with $\mathrm{D} \in \mathbb{R}^k$, different notions of order sensitivity arise, e.g., component-wise order sensitivity.

Convex Level Sets

Theorem (Osband 1985)

If a one-dimensional ($\mathsf{D} \subseteq \mathbb{R}$) *functional T is elicitable, then its level sets are convex in the following sense: If $F_0, F_1 \in \mathcal{F}$ and $p \in (0,1)$ are such that $F_p = pF_0 + (1-p)F_1 \in \mathcal{F}$, then $t \in T(F_0)$ and $t \in T(F_1)$ imply $t \in T(F_p)$.*

Proof.

For $t \in T(F_0)$ and $t \in T(F_1)$, we have $\mathbb{E}_F[S(t,Y)] \leq \mathbb{E}_F[S(z,Y)]$ for all $z \in \mathsf{A}$ and $F \in \{F_0, F_1\}$. Then $\mathbb{E}_{F_p}[S(t,Y)] = p\,\mathbb{E}_{F_0}[S(t,Y)] + (1-p)\,\mathbb{E}_{F_1}[S(t,Y)] \leq p\,\mathbb{E}_{F_0}[S(z,Y)] + (1-p)\,\mathbb{E}_{F_1}[S(z,Y)] = \mathbb{E}_{F_p}[S(z,Y)]$. $\square$

Application

Proof that variance is not elicitable:
We have $\mathrm{Var}[\delta_x] = \mathrm{Var}[\delta_y] = \mathrm{Var}_{\delta_y}[Y] = 0$. But
$\mathrm{Var}[p\delta_x + (1-p)\delta_y] = \mathbb{E}_{p\delta_x+(1-p)\delta_y}[(Y-(px+(1-p)y))^2] =$
$p((1-p)x-(1-p)y)^2 + (1-p)(px-py)^2 = p(1-p)(x-y)^2 \neq 0$

Revelation Principle

Theorem (Osband 1985)

Suppose that the class $\mathcal{F}$ is concentrated on the domain D, and let $g : \mathrm{D} \to \mathrm{D}$ be a one-to-one mapping. Then the following holds.

1. *If T is elicitable, then $T_g = g \circ T$ is elicitable.*
2. *If S is consistent for T, then the scoring function $S_g(x, y) = S(g^{-1}(x), y)$ is consistent for T_g.*
3. *If S is strictly consistent for T, then S_g is strictly consistent for T_g.*

Scoring Functions with Weighted Densities

Some Assumptions

- Functional T is defined on class $\mathcal{F}$ of probability distributions which admit a density, f, with respect to some dominating measure on the domain D.
- Weight function $w : \mathrm{D} \to [0, \infty)$
- $\mathcal{F}^{(w)} \subseteq \mathcal{F}$ denotes subclass of probability distributions in $\mathcal{F}$ which are such that $\int_{\mathrm{D}} w(y) f(y) \, \mathrm{d}y < \infty$, and the probability measure $F^{(w)}$ with density proportional to $w(y)f(y)$ belongs to $\mathcal{F}$. On this subclass $\mathcal{F}^{(w)}$, define the functional

$$T^{(w)} : \mathcal{F}^{(w)} \to I \subseteq \mathbb{R} \qquad F \to T^{(w)}(F) = T(F^{(w)}) \, .$$

Theorem (Gneiting 2011 Theorem 5)

Given the above assumptions, the following holds.

1. *If T is elicitable, then $T^{(w)}$ is elicitable.*
2. *If S is consistent for T relative to $\mathcal{F}$, then $S^{(w)}(z, y) = w(y)S(z, y)$ is consistent for $T^{(w)}$ relative to $\mathcal{F}^{(w)}$.*

Characterisation

Expectation: Bregman functions

$$S(z, y) = \phi(y) - \phi(z) + \phi'(z)(z - y) + a(y) \tag{13}$$

with (strictly) convex ϕ and arbitrary a are (strictly) consistent for $T = \mathbb{E}$

Quantiles: generalised piecewise linear (GPL)

$$S(z, y) = (\mathbb{1}\{y \leq z\} - \alpha)(g(z) - g(y)) + a(y) \tag{14}$$

with (strictly) increasing g and arbitrary a are (strictly) consistent for $T = q_\alpha$

Expectiles

$$S(z, y) = 2|\mathbb{1}\{y \leq z\} - \alpha|(\phi(y) - \phi(z) + \phi'(z)(z - y)) + a(y) \tag{15}$$

with (strictly) convex ϕ and arbitrary a are (strictly) consistent for $T = e_\alpha$.

Mode: zero-one loss

$$S(z, y) = \lambda \mathbb{1}\{z \neq y\} + a(y) \quad \lambda > 0 \tag{16}$$

is strictly consistent for categorical $Y \in \{0, \ldots, k - 1\}$.

Examples of Strictly Consistent Scoring Functions

Functional	Scoring Function	Formula $S(z, y)$	Domain
expectation	squared error	$(y - z)^2$	$y, z \in \mathbb{R}$
	Poisson deviance	$2\left(y \log \frac{y}{z} + z - y\right)$	$y \geq 0, z > 0$
	Gamma deviance	$2\left(\log \frac{z}{y} + \frac{y}{z} - 1\right)$	$y, z > 0$
	Tweedie deviance	$2\left(\frac{y^{2-p}}{(1-p)\cdot(2-p)}\right.$	$y, z > 0$
	$p \in \mathbb{R} \setminus \{1, 2\}$	$\left.-\frac{y \cdot z^{1-p}}{1-p} + \frac{z^{2-p}}{2-p}\right)$	$y \geq 0$ for $p < 2$
	homogeneous score	$\lvert y\rvert^a - \lvert z\rvert^a$	$y, z \in \mathbb{R}$
	$a > 1$	$-a \operatorname{sign}(z)\lvert z\rvert^{a-1}(y - z)$	
	log loss	$-y \log z - (1 - y) \log(1 - z)$	$0 \leq y \leq 1$
		$+y \log y + (1 - y) \log(1 - y)$	$0 < z < 1$
α-expectile	APQSF[3]	$\lvert\mathbb{1}\{z \geq y\} - \alpha\rvert(z - y)^2$	$\mathbb{R}$
median	absolute error	$\lvert y - z\rvert$	$\mathbb{R}$
α-quantile	pinball loss	$(\mathbb{1}\{z \geq y\} - \alpha)(z - y)$	$\mathbb{R}$

[3] asymmetric piecewise quadratic scoring function

Scoring Functions with Weighted Densities

Example

- On $\mathrm{D} = (0, \infty)$, $S(z, y) = |z^{-\beta} - y^{-\beta}|$ and $w(y) = y^{\beta}$ produce

$$S_{\beta}(z, y) = \left|1 - \left(\frac{y}{z}\right)^{\beta}\right| \tag{17}$$

- $S(z, y)$ is consistent for the median, see Eq. (14) with $g(x) = \mathrm{sign}(b)x^{b}$.
- By Theorem (Gneiting 2010 Th. 5), $S_{\beta}(z, y)$ is consistent for the β-median, $\mathrm{med}^{(\beta)}(F)$, i.e. the median of the distribution with density proportional to $y^{\beta} f(y)$, and f the density of F.

Special cases

- $\beta = -1$: absolute percentage error (APE) $S_{-1}(z, y) = |\frac{z-y}{y}|$
- $\beta = 1$: relative error (RE) $S(z, y) = |\frac{z-y}{z}|$

Elementary Scoring Functions

With identification function V for quantile or expectile T, see (22) on slide 70, the elementary scoring function

$$S_\theta(z,y) = (\mathbb{1}\{\theta \le z\} - \mathbb{1}\{\theta \le y\})\, V(\theta, y) \tag{18}$$

is consistent for T.
Any (strictly) consistent scoring function admits a mixture representation

$$S(z,y) = \int S_\theta(z,y)\,\mathrm{d}H(\theta) + a(y) \tag{19}$$

for non-negative (positive[4]) measure H on $\mathbb{R}$, with $\mathrm{d}H(\theta) = \mathrm{d}g(\theta)$ for quantiles and $\mathrm{d}H(\theta) = \mathrm{d}\phi'(\theta)$ for expectiles.

Note: $V(z,y) = z - y$ for $T = \mathbb{E}$.

[4] H gives positive measure to every non-degenerate interval.

Forecast Dominance

Definition

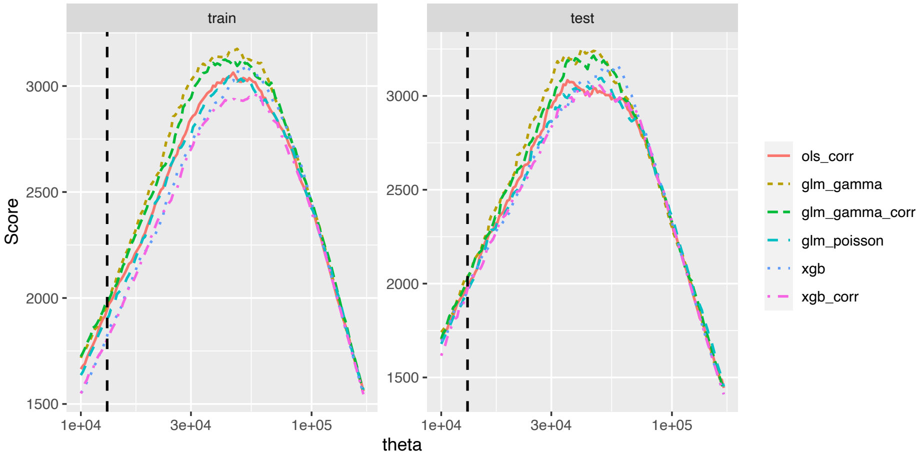
Prediction/forecast z_1 dominates z_2 if $\mathbb{E}[S(z_1, Y)] < \mathbb{E}[S(z_2, Y)]$ for all (strictly) consistent scoring functions.

Quantiles and expectiles

For quantiles and expectiles this is equivalent to $\mathbb{E}[S_\theta(z_1, Y)] < \mathbb{E}[S_\theta(z_2, Y)]$ for all $\theta \in \mathbb{R}$.

Murphy Diagram

Compare many scoring functions (sliding parameter θ) at once.
Assess forecast dominance.



Elementary scoring function for $\mathbb{E}$: $S_\theta(z, y) = |\theta - y| \mathbb{1}\{\min(z, y) \leq \theta < \max(z, y)\}$

Which One to Choose?

Use a strictly consistent scoring function!

But: Which one out of the infinitely many ones (for elicitable T)?

Further criteria

- Domain / Range of target Y.
- Degree of homogeneity: $S(tz, ty) = t^h S(z, y)$ for all $t > 0$ and for all z, y
- Efficiency: How fast is the large sample convergence?
- Forecast dominance: Is one model dominating for many/all scoring functions? Assess with Murphy diagrams.

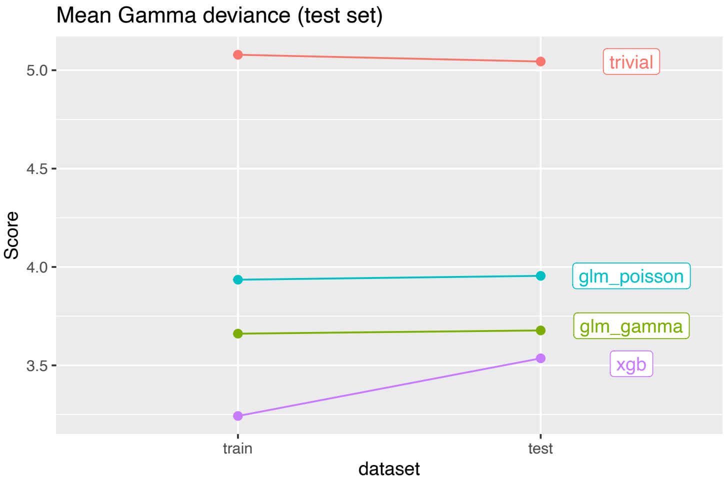
Squared error: $h = 2$

Gamma deviance:
Degree of homogeneity is $h = 0 \Rightarrow$ It only cares about relative differences:
$S(1, 10) = S(10, 100) = S(100, 1000) = 13.39$

Model Comparison

Compare empirical mean scores: $\overline{S}(m) = \frac{1}{n}\sum_i S(m(\boldsymbol{x}_i), y_i)$

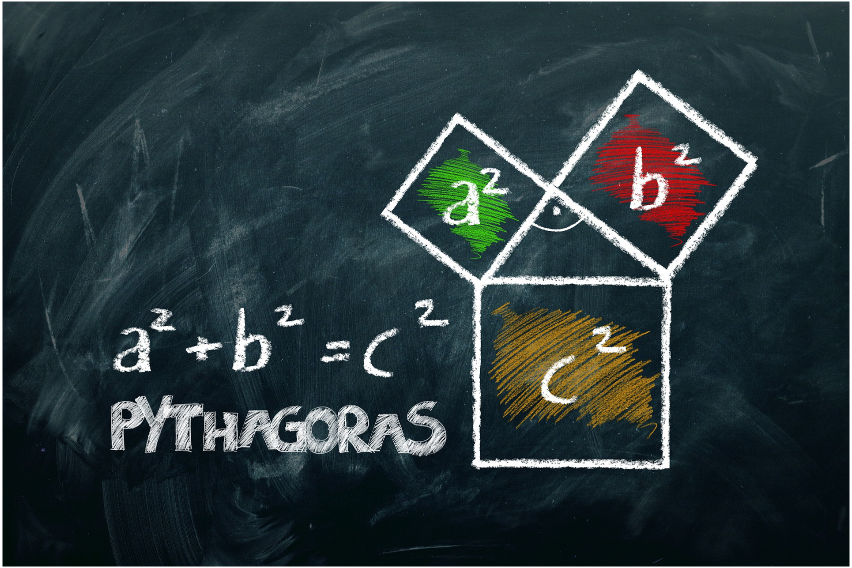
Gamma deviance for workers compensation



Models:

1. Trivial model always predicts mean(y) of the training set.
2. Poisson GLM with canonical log-link.
3. Gamma GLM with log-link.
4. XGBoost model with Gamma deviance and log-link.

Additive Score Decomposition



Score Decomposition

$$\mathbb{E}[S(m(\boldsymbol{X}),Y)] = \Big\{ \underbrace{\mathbb{E}[S(m(\boldsymbol{X}),Y)] - \mathbb{E}[S(T(Y|\boldsymbol{X}),Y)]}_{\text{conditional miscalibration}} \Big\} \tag{20}$$

$$- \Big\{ \underbrace{\mathbb{E}[S(T(Y),Y)] - \mathbb{E}[S(T(Y|\boldsymbol{X}),Y)]}_{\text{conditional resolution / conditional discrimination}} \Big\} + \underbrace{\mathbb{E}[S(T(Y),Y)]}_{\text{uncertainty / entropy}}$$

can be estimated

$$\left\{ \begin{aligned} &= \Big\{ \underbrace{\mathbb{E}[S(m(\boldsymbol{X}),Y)] - \mathbb{E}[S(T(Y|m(\boldsymbol{X})),Y)]}_{\text{auto-miscalibration}} \Big\} \\ &- \Big\{ \underbrace{\mathbb{E}[S(T(Y),Y)] - \mathbb{E}[S(T(Y|m(\boldsymbol{X})),Y)]}_{\text{auto-resolution / auto-discrimination}} \Big\} + \underbrace{\mathbb{E}[S(T(Y),Y)]}_{\text{uncertainty / entropy}} \end{aligned} \right.$$

Note: Minimising consistent scores amounts to *jointly* minimising miscalibration and maximising resolution!

Squared Error / Brier Score

$$\mathbb{E}[(m(\boldsymbol{X}) - Y)^2] = \underbrace{\mathbb{E}[(m(\boldsymbol{X}) - \mathbb{E}[Y|\boldsymbol{X}])^2]}_{\text{conditional miscalibration}} - \underbrace{\mathrm{Var}[\mathbb{E}[Y|\boldsymbol{X}]]}_{\text{conditional resolution}} + \underbrace{\mathrm{Var}[Y]}_{\text{uncertainty}} \tag{21}$$

Score Decomposition of Gamma Deviance

Again for workers compensation

Model	Mean deviance	Auto-miscalibration	Auto-resolution	Uncertainty
Trivial	5.04	0	0	5.04
GLM Gamma	3.68	0.190	1.56	5.04
GLM Poisson	3.95	0.482	1.57	5.04
XGB	**3.54**	**0.124**	**1.63**	5.04

Isotonic regression

For $T = \mathbb{E}$, one can estimate $\mathbb{E}[Y|m(\boldsymbol{x})]$ by isotonic regression (PAV algorithm) of y_i against $m(\boldsymbol{x}_i)$.

New results[5] show how to extend PAV to quantiles, expectiles and more.

[5] A.I. Jordan, A. Mühlemann & J.F. Ziegel (2022)"Characterizing the optimal solutions to the isotonic regression problem for identifiable functionals" Ann Inst Stat Math 74, 489-514. doi:10.1007/s10463-021-00808-0

Exercises

Exercise 10

Compute the Bayes rule for the scoring functions in Eq. (14) and (15). Remember Ex. 2.

Exercise 11

Device a betting game with a wager $\rho_L > 0$ and pay-off scheme depending on a random outcome y such that the optimal strategy in expectation is a quantile. Hint: Have a look at the elementary scoring function.

Exercise 12

Derive the decomposition of the squared error in Eq. (21).

Exercise 13

Calculate the score decomposition of the Gamma deviance for your models on the Workers Compensation dataset.