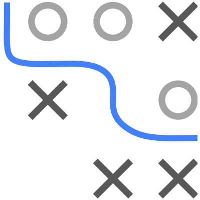
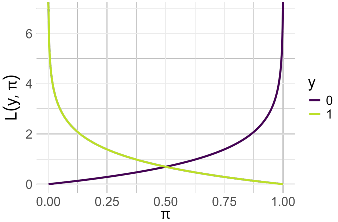
Introduction to Machine Learning



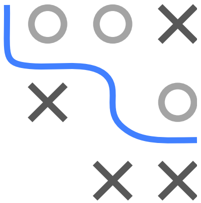
Advanced Risk Minimization
Proper Scoring Rules



Learning goals

- Honest probabilistic forecasts
- Proper scoring rules
- log score
- Brier score

PROBABILISTIC FORECASTS ▸ Gneiting and Raftery 2007



Scoring rules $S(P, y)$ assess the quality of probabilistic forecasts by assigning a score based on the predictive distribution P and the realized event y. The expected score w.r.t. the RV $y \sim Q$ is denoted as

$$S(P, Q) = \mathbb{E}_{y \sim Q}[S(P, y)]$$

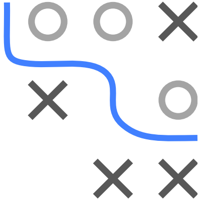
A scoring rule is **proper** if the forecaster maximizes the expected score for an observation drawn from Q if he or she issues the forecast Q rather than $P \neq Q$:

$$S(Q, Q) \geq S(P, Q) \text{ for all } P, Q$$

S is **strictly proper** when equality holds iff $P = Q$. (Strictly) proper scores ensure the forecaster has an incentive to predict Q and is encouraged to report his or her true belief.

NB: scores are typically positively oriented (maximization) while losses are negatively oriented (minimization). Scores could also be defined negatively oriented.

BINARY CLASSIFICATION SCORES



For simplicity, we will only look at binary targets $y \sim \text{Bern}(p)$.
We want to find out if using a loss $L(y, \pi)$ (negative score) incentivizes honest forecasts $\pi = p$ for any $p \in [0, 1]$.

For any loss L, its expectation w.r.t. y is

$$\mathbb{E}_y[L(y, \pi)] = p \cdot L(1, \pi) + (1 - p) \cdot L(0, \pi)$$

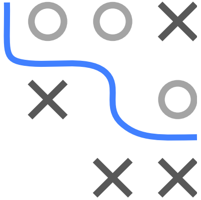
Let's first look at a negative example. Assuming the **L1 loss** $L(y, \pi) = |y - \pi|$, we obtain

$$\mathbb{E}_y[L(y, \pi)] = p|1 - \pi| + (1 - p)\pi = p + \pi(1 - 2p)$$

The expected loss is linear in π, hence we minimize it by setting $\pi = 1$ for $p > 0.5$ and $\pi = 0$ for $p < 0.5$.

The score $S(\pi, y) = -L(y, \pi)$ is therefore not proper.

BINARY CLASSIFICATION SCORES



The **0/1 loss** $L(y, \pi) = \mathbb{1}_{\{y \neq h_\pi\}}$ using the discrete classifier $h_\pi = \mathbb{1}_{\{\pi > 0.5\}}$ yields in expectation over y:

$$\begin{aligned} \mathbb{E}_y[L(y, \pi)] &= p \cdot L(1, \pi) + (1 - p) \cdot L(0, \pi) \\ &= \begin{cases} p & \text{if } h_\pi = 0 \\ 1 - p & \text{if } h_\pi = 1 \end{cases} \end{aligned}$$

- For $p > 0.5$ we minimize the expected loss by choosing $h_\pi = 1$, which holds true for any $\pi \in (0.5, 1)$
- Likewise for $p \leq 0.5$, any $\pi \in (0, 0.5]$ minimizes the expected loss

The **0/1 score** (negative 0/1 loss) is therefore proper but not strictly proper since there is no unique maximum.

BINARY CLASSIFICATION SCORES

To find strictly proper scores/losses, we can ask: Which functions have the property such that $\mathbb{E}_y[L(y, \pi)]$ is minimized at $\pi = p$? We have

$$\mathbb{E}_y[L(y, \pi)] = p \cdot L(1, \pi) + (1 - p) \cdot L(0, \pi)$$

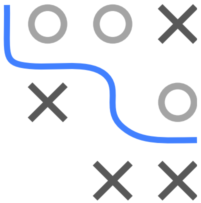
Let's further assume that $L(1, \pi)$ and $L(0, \pi)$ can not be arbitrary, but are the same function evaluated at π and $1 - \pi$: $L(1, \pi) = L(\pi)$ and $L(0, \pi) = L(1 - \pi)$. Then

$$\mathbb{E}_y[L(y, \pi)] = p \cdot L(\pi) + (1 - p) \cdot L(1 - \pi)$$

Setting the derivative w.r.t. π to 0 and requiring $\pi = p$ at the optimum (**propriety**), we get the following first-order condition (F.O.C.):

$$p \cdot L'(p) \overset{!}{=} (1 - p) \cdot L'(1 - p)$$

BINARY CLASSIFICATION SCORES / 2



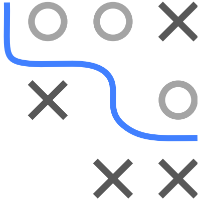
- F.O.C.: $p \cdot L'(p) \overset{!}{=} (1-p) \cdot L'(1-p)$
- One natural solution is $L'(p) = -1/p$, resulting in $-p/p = -(1-p)/(1-p) = -1$ and the antiderivative $L(p) = -\log(p)$.
- This is the **log loss**

$$L(y, \pi) = -(y \cdot \log(\pi) + (1-y) \cdot \log(1-\pi))$$

- The corresponding scoring rule (maximization) is the strictly proper **logarithmic scoring rule**

$$S(\pi, y) = y \cdot \log(\pi) + (1-y) \cdot \log(1-\pi)$$

BINARY CLASSIFICATION SCORES / 3



- F.O.C.: $p \cdot L'(p) \stackrel{!}{=} (1-p) \cdot L'(1-p)$
- A second solution is $L'(p) = -2(1-p)$, resulting in $-2p(1-p) = -2(1-p)p$ and the antiderivative $L(p) = (1-p)^2 = \frac{1}{2}((1-p)^2 + (0-(1-p))^2)$
- This is also called the **Brier score** and is effectively the **MSE loss** for probabilities

$$L(y, \pi) = \frac{1}{2}\sum_{i=1}^{2}(y_i - \pi_i)^2$$

  (with $y_1 = y, y_2 = 1 - y$ and likewise $\pi_1 = \pi, \pi_2 = 1 - \pi$)
- Using positive orientation (maximization), this gives rise to the **quadratic scoring rule**, which for two classes is $S(\pi, y) = -\frac{1}{2}\sum_{i=1}^{2}(y_i - \pi_i)^2$