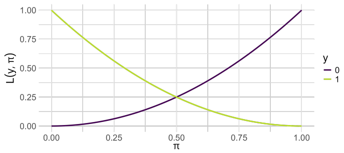
Introduction to Machine Learning

Advanced Risk Minimization
Brier Score



Learning goals

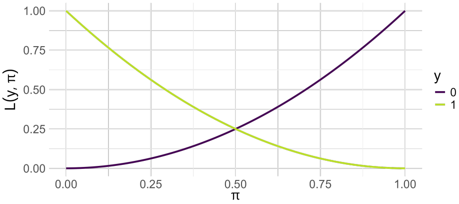
- Know the Brier score
- Derive the risk minimizer
- Derive the optimal constant model

BRIER SCORE

The binary Brier score is defined on probabilities $\pi \in [0, 1]$ and 0-1-encoded labels $y \in \{0, 1\}$ and measures their squared distance ($L2$ loss on probabilities).

$$L(y, \pi) = (\pi - y)^2$$

As the Brier score is a proper scoring rule, it can be used for calibration. Note that is is not convex on probabilities anymore.



BRIER SCORE: RISK MINIMIZER

The risk minimizer for the (binary) Brier score is

$$\pi^*(\mathbf{x}) = \eta(\mathbf{x}) = \mathbb{P}(y \mid \mathbf{x} = \mathbf{x}),$$

which means that the Brier score attains its minimum if the prediction equals the “true” probability of the outcome.

The risk minimizer for the multiclass Brier score is

$$\pi^*(\mathbf{x}) = \mathbb{P}(y = k \mid \mathbf{x} = \mathbf{x}).$$

Proof: We only show the proof for the binary case. We need to minimize

$$\mathbb{E}_x \left[L(1, \pi(\mathbf{x})) \cdot \eta(\mathbf{x}) + L(0, \pi(\mathbf{x})) \cdot (1 - \eta(\mathbf{x})) \right],$$

BRIER SCORE: RISK MINIMIZER / 2

which we do point-wise for every $\mathbf{x}$. We plug in the Brier score

$$
\begin{aligned}
& \arg\min_c \quad L(1, c)\eta(\mathbf{x}) + L(0, c)(1 - \eta(\mathbf{x})) \\
= \ & \arg\min_c \quad (c-1)^2\eta(\mathbf{x}) + c^2(1 - \eta(\mathbf{x})) \quad |+\eta(\mathbf{x})^2 - \eta(\mathbf{x})^2 \\
= \ & \arg\min_c \quad (c^2 - 2c\eta(\mathbf{x}) + \eta(\mathbf{x})^2) - \eta(\mathbf{x})^2 + \eta(\mathbf{x}) \\
= \ & \arg\min_c \quad (c - \eta(\mathbf{x}))^2.
\end{aligned}
$$

The expression is minimal if $c = \eta(\mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x} = \mathbf{x})$.

BRIER SCORE: OPTIMAL CONSTANT MODEL

The optimal constant probability model $\pi(\mathbf{x}) = \theta$ w.r.t. the Brier score for labels from $\mathcal{Y} = \{0, 1\}$ is:

$$
\begin{aligned}
\min_{\theta} \mathcal{R}_{\text{emp}}(\theta) &= \min_{\theta} \sum_{i=1}^{n} \left(y^{(i)} - \theta\right)^2 \\
\Leftrightarrow \frac{\partial \mathcal{R}_{\text{emp}}(\theta)}{\partial \theta} &= -2 \cdot \sum_{i=1}^{n} (y^{(i)} - \theta) = 0 \\
\hat{\theta} &= \frac{1}{n} \sum_{i=1}^{n} y^{(i)}.
\end{aligned}
$$

This is the fraction of class-1 observations in the observed data.
(This also directly follows from our $L2$ proof for regression).

Similarly, for the multiclass brier score the optimal constant is
$\hat{\theta}_k = \frac{1}{n} \sum\limits_{i=1}^{n} [y = k]$.