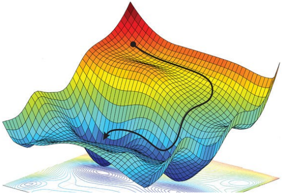
Introduction to Machine Learning

Brier Score



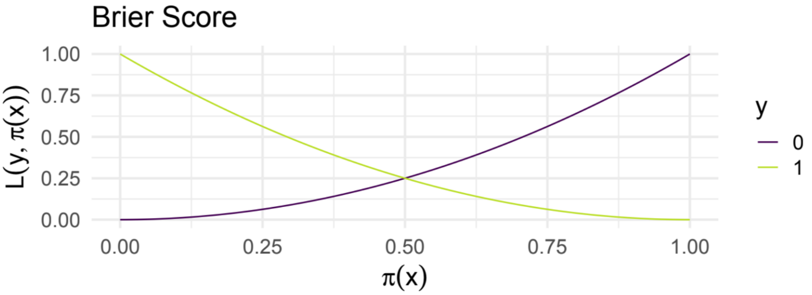
Learning goals

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

BRIER SCORE

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

$$L(y, \pi(\mathbf{x})) = (\pi(\mathbf{x}) - y)^2$$



BRIER SCORE: RISK MINIMIZER

The risk minimizer for the Brier score is

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

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

Proof: We have seen that the (theoretical) optimal prediction c for an arbitrary loss function at fixed point $\mathbf{x}$ is

$$\arg\min_c \sum_{k \in \mathcal{Y}} L(k, c)\, \mathbb{P}(y = k \mid \mathbf{x} = \mathbf{x})\,.$$

BRIER SCORE: RISK MINIMIZER

We plug in the Brier score

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

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:

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

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

BRIER SCORE MINIMIZATION = GINI SPLITTING

Splitting a classification tree w.r.t. the Gini index is equivalent to minimizing the Brier score in each node.

To prove this we show that

$$\mathcal{R}(\mathcal{N}) = n_{\mathcal{N}} I(\mathcal{N})$$

where I is the Gini impurity

$$I(\mathcal{N}) = \sum_{k \neq k\prime} \pi_k^{(\mathcal{N})} \pi_{k\prime}^{(\mathcal{N})} = \sum_{k=1}^{g} \pi_k^{(\mathcal{N})} (1 - \pi_k^{(\mathcal{N})}),$$

and $\mathcal{R}(\mathcal{N})$ is calculated w.r.t. the Brier score

$$L(y, \pi(\mathbf{x})) = \sum_{k=1}^{g} ([y = k] - \pi_k(\mathbf{x}))^2 .$$

BRIER SCORE MINIMIZATION = GINI SPLITTING

$$
\begin{aligned}
\mathcal{R}(\mathcal{N}) &= \sum_{(\mathbf{x},y)\in\mathcal{N}} \sum_{k=1}^{g} ([y = k] - \pi_k(\mathbf{x}))^2 \\
&= \sum_{k=1}^{g} \sum_{(\mathbf{x},y)\in\mathcal{N}} ([y = k] - \pi_k(\mathbf{x}))^2 \\
&= \sum_{k=1}^{g} n_{\mathcal{N},k} \left(1 - \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}\right)^2 + (n_{\mathcal{N}} - n_{\mathcal{N},k}) \left(\frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}\right)^2
\end{aligned}
$$

In the last step, we plugged in the optimal prediction w.r.t. the Brier score (the fraction of class k observations):

$$\hat{\pi}_k(\mathbf{x}) = \pi_k^{(\mathcal{N})} = \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}.$$

BRIER SCORE MINIMIZATION = GINI SPLITTING

We further simplify the expression to

$$
\begin{aligned}
\mathcal{R}(\mathcal{N}) &= \sum_{k=1}^{g} n_{\mathcal{N},k} \left(\frac{n_{\mathcal{N}} - n_{\mathcal{N},k}}{n_{\mathcal{N}}} \right)^2 + (n_{\mathcal{N}} - n_{\mathcal{N},k}) \left(\frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}} \right)^2 \\
&= \sum_{k=1}^{g} \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}} \frac{n_{\mathcal{N}} - n_{\mathcal{N},k}}{n_{\mathcal{N}}} (n_{\mathcal{N}} - n_{\mathcal{N},k} + n_{\mathcal{N},k}) \\
&= n_{\mathcal{N}} \sum_{k=1}^{g} \pi_k^{(\mathcal{N})} \cdot (1 - \pi_k^{(\mathcal{N})}) = n_{\mathcal{N}} I(\mathcal{N}).
\end{aligned}
$$