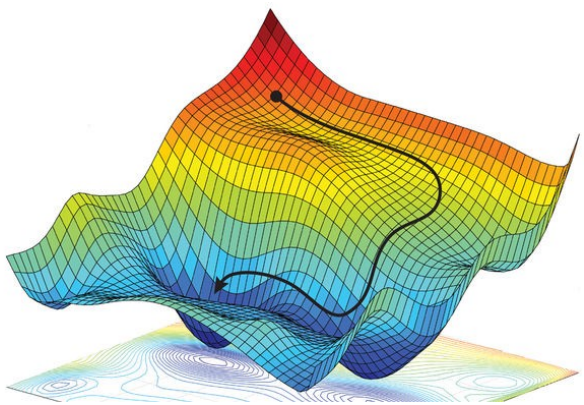
Introduction to Machine Learning

Advanced Classification Losses



Learning goals

- Understand that an ML model is simply a parametrized curve
- Understand that the hypothesis space lists all admissible models for a learner
- Understand the relationship between the hypothesis space and the parameter space

RISK MINIMIZATION FOR CLASSIFICATION

Let y be categorical with g classes, i. e. $\mathcal{Y} = \{1, ..., g\}$ and let $f : \mathcal{X} \to \mathbb{R}^g$. We assume our model f outputs a g-dimensional vector of scores or probabilities, one per class.

Note: In this section, we will consider loss for **binary classification** tasks, so $f(\mathbf{x})$ and $\pi(\mathbf{x})$ are univariate scalars.

We will (usually) encode labels as $y \in \{-1, 1\}$ for scoring classifiers $f(\mathbf{x})$, and as $y \in \{0, 1\}$ for probabilistic classifiers $\pi(\mathbf{x})$ unless explicitly stated differently.

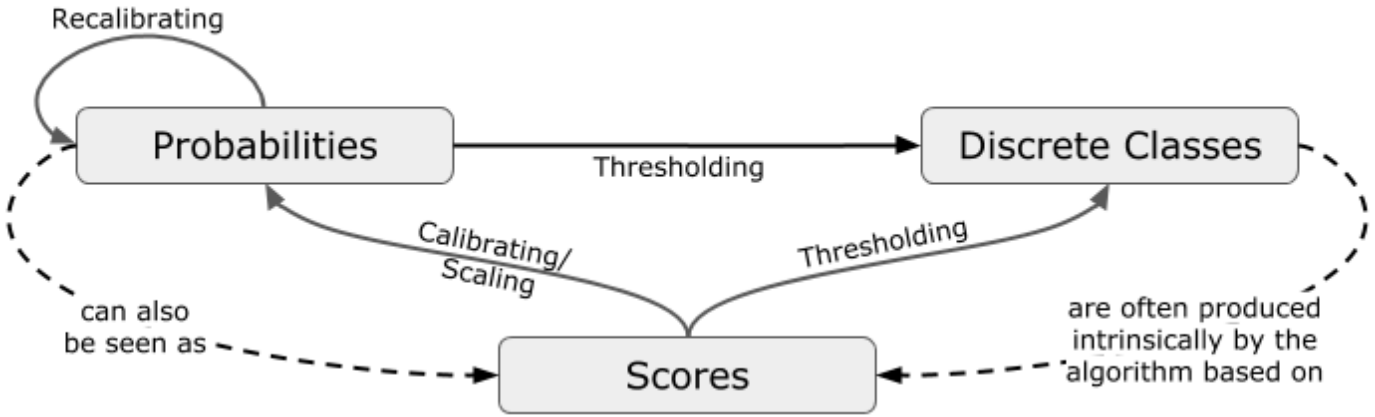
Goal: Find a model f that minimizes the expected loss over random observations $(\mathbf{x}, y) \sim \mathbb{P}_{xy}$

$$\arg\min_{f \in \mathcal{H}} \mathcal{R}(f) = \mathbb{E}_{xy}[L(y, f(\mathbf{x}))] = \int L(y, f(\mathbf{x})) \, \mathrm{d}\mathbb{P}_{xy}.$$

RISK MINIMIZATION FOR CLASSIFICATION

- As for regression before, losses measure prediction errors **point-wise**.
- In classification, however, we need to distinguish the different types of prediction functions:
- Losses can either be defined on
  - hard labels $h(\mathbf{x})$ or
  - (class) scores $f(\mathbf{x})$ or
  - (class) probabilities $\pi(\mathbf{x})$.
- For multiclass classification, loss functions will be defined on vectors of scores $(f_1(\mathbf{x}), ..., f_g(\mathbf{x}))$ or on vectors of probabilities $(\pi_1(\mathbf{x}), ..., \pi_g(\mathbf{x}))$.

RISK MINIMIZATION FOR CLASSIFICATION



Note that for a **binary scoring classifier** $f(\mathbf{x})$,

$$h(\mathbf{x}) = \text{sign}(f(\mathbf{x})) \in \{-1, 1\}$$

and for a **probabilistic classifier** $\pi(\mathbf{x})$

$$h(\mathbf{x}) = \mathbb{1}_{\{\pi(\mathbf{x}) > c\}} \in \{0, 1\}$$

(e.g. $c = 0.5$) will be the corresponding label.

MARGINS

When considering scoring classifiers $f(\mathbf{x})$ we usually define loss functions on the so-called **margin**

$$r = y \cdot f(\mathbf{x}) = \begin{cases} > 0 & \text{if } y = \text{sign}(f(\mathbf{x})) \text{ (correct classification)}, \\ < 0 & \text{if } y \neq \text{sign}(f(\mathbf{x})) \text{ (misclassification)}, \end{cases}$$

$|f(\mathbf{x})|$ is called **confidence**.

0-1-Loss

0-1-LOSS

- Let us first consider a classifier $h(\mathbf{x})$ that outputs discrete classes directly.
- The most natural choice for $L(y, h(\mathbf{x}))$ is of course the 0-1-loss that counts the number of misclassifications

$$L(y, h(\mathbf{x})) = \mathbb{1}_{\{y \neq h(\mathbf{x})\}} = \begin{cases} 1 & \text{if } y \neq h(\mathbf{x}) \\ 0 & \text{if } y = h(\mathbf{x}) \end{cases} .$$

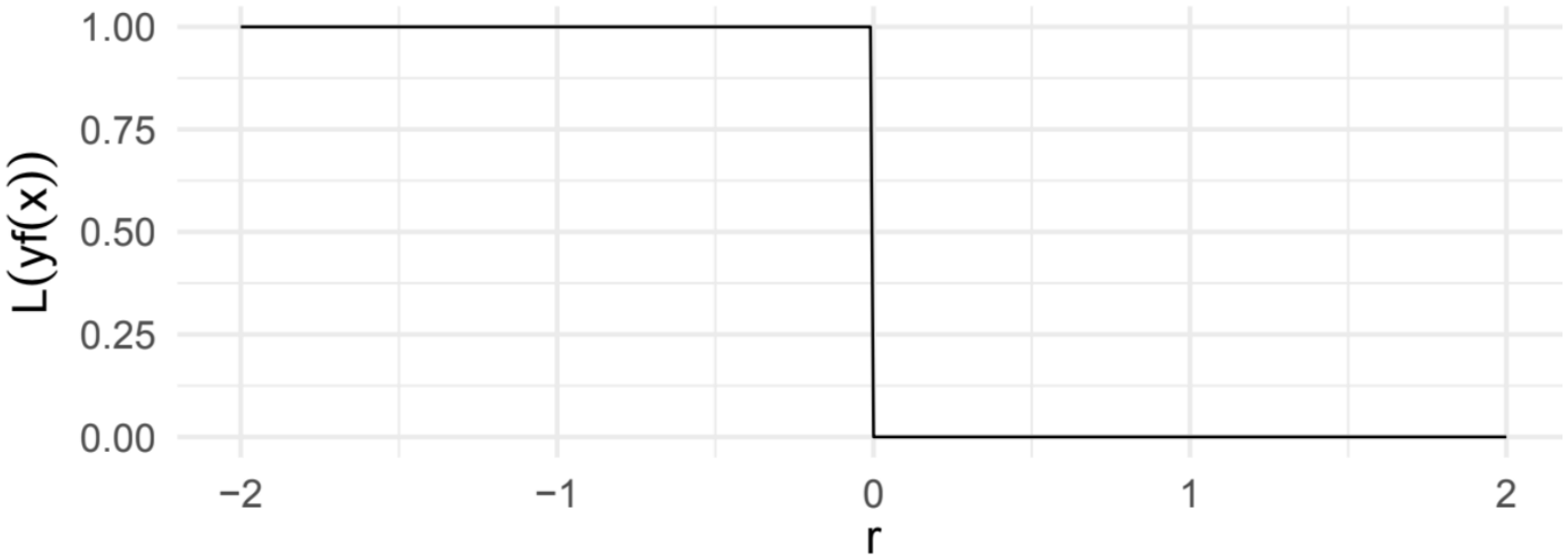
- We can express the 0-1-loss also for a scoring classifier $f(\mathbf{x})$ based on the margin r

$$L(r) = \mathbb{1}_{\{r<0\}} = \mathbb{1}_{\{yf(\mathbf{x})<0\}} = \mathbb{1}_{\{y \neq h(\mathbf{x})\}}.$$

0-1-LOSS

$$L(r) = \mathbb{1}_{\{r<0\}} = \mathbb{1}_{\{yf(\mathbf{x})<0\}} = \mathbb{1}_{\{y \neq h(\mathbf{x})\}}$$

- Intuitive, often what we are interested in.
- Analytic properties: Not continuous, even for linear f the optimization problem is NP-hard and close to intractable.

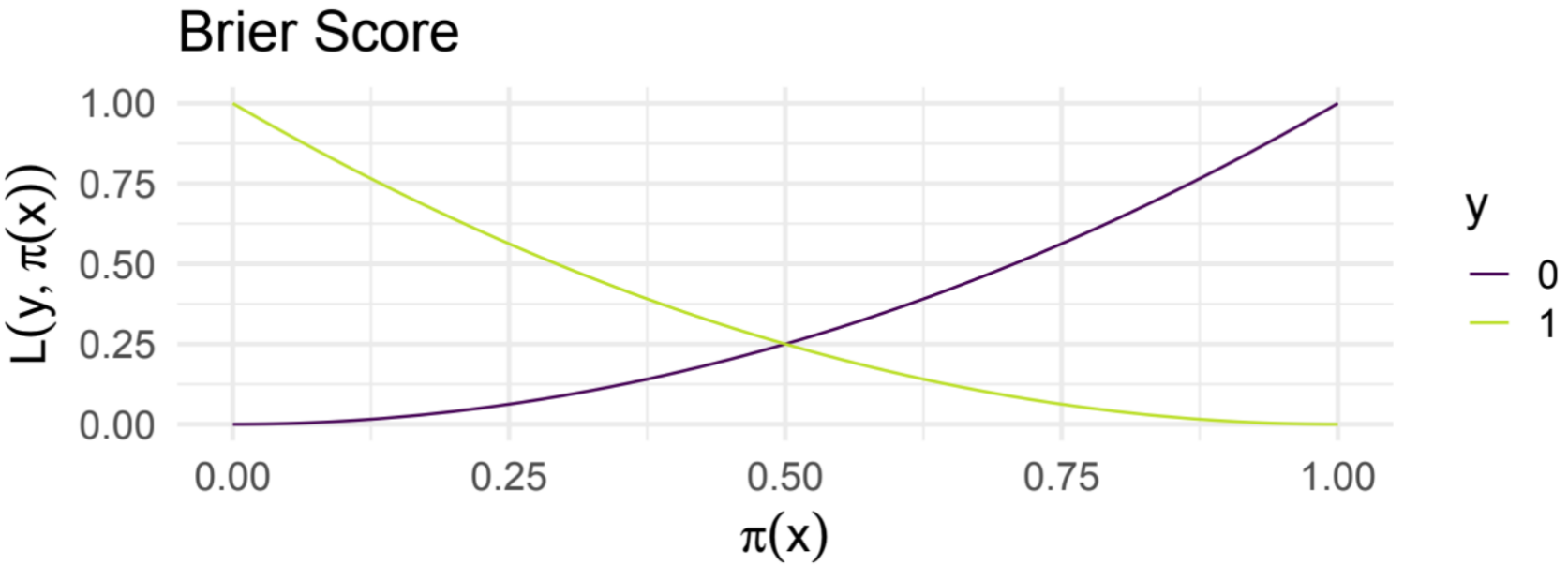


Brier Score

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).

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

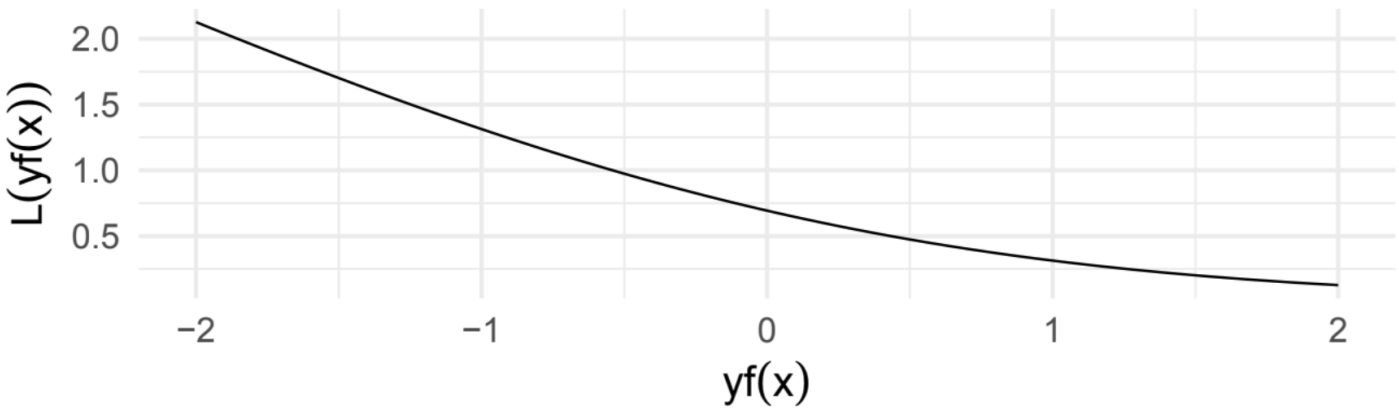


Bernoulli Loss

BERNOULLI LOSS

$$
\begin{aligned}
L_{-1,+1}(y, f(\mathbf{x})) &= \ln(1 + \exp(-yf(\mathbf{x}))) \\
L_{0,1}(y, f(\mathbf{x})) &= -y \cdot f(\mathbf{x}) + \log(1 + \exp(f(\mathbf{x}))).
\end{aligned}
$$

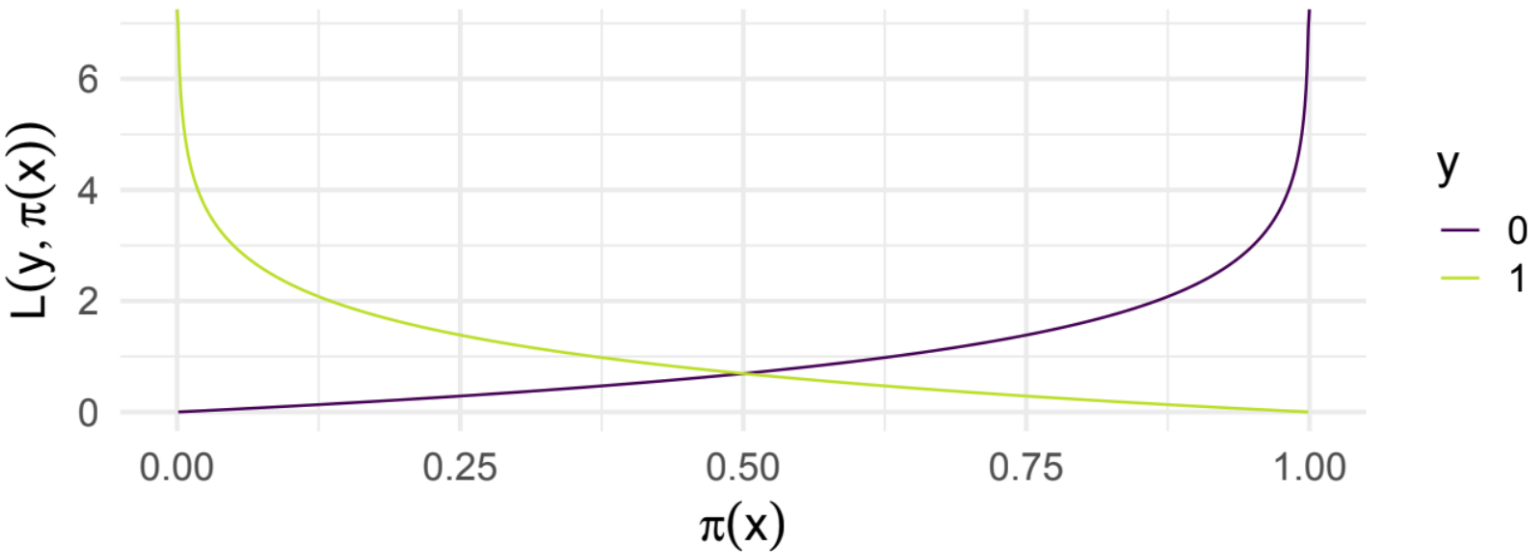
- Two equivalent formulations: Labels $y \in \{-1, 1\}$ or $y \in \{0, 1\}$
- Negative log-likelihood of Bernoulli model, e.g., logistic regression
- Convex, differentiable
- Pseudo-Residuals (0,1 case): $\tilde{r} = y - \frac{1}{1+\exp(-f(\mathbf{x}))}$
  Interpretation: $L1$ distance between 0/1-labels and posterior prob!



BERNOULLI LOSS ON PROBABILITIES

If scores are transformed into probabilities by the logistic function $\pi(\mathbf{x}) = (1 + \exp(-f(\mathbf{x})))^{-1}$, we arrive at another equivalent formulation of the loss

$$L_{0,1}(y, \pi(\mathbf{x})) = -y \log(\pi(\mathbf{x})) - (1 - y) \log(1 - \pi(\mathbf{x})).$$



Via this form it is easy to show that the point-wise optimum for probability estimates is $\hat{\pi}(\mathbf{x}) = \mathbb{P}(y \mid \mathbf{x} = \mathbf{x})$.

Summary

SUMMARY OF LOSS FUNCTIONS

Name	Formula	Differentiable
0-1	$L(y, h(\mathbf{x})) = [y \neq h(\mathbf{x})]$	✗
Brier	$L(y, \pi(\mathbf{x})) = (\pi(\mathbf{x}) - y)^2$	✓
Bernoulli	$L_{-1+1}(y, f(\mathbf{x})) = \ln[1 + \exp(-yf(\mathbf{x}))]$	✓
Bernoulli	$L_{0,1}(y, f(\mathbf{x})) = -yf(\mathbf{x}) + \log(1 + \exp(f(\mathbf{x})))$	✓
Bernoulli	$L_{0,1}(y, \pi(\mathbf{x})) = -y \log \pi(\mathbf{x}) - (1 - y) \log(1 - \pi(\mathbf{x}))$	✓

Name	Point-wise Opt.	Optimal Constant
0-1	$\hat{h}(\mathbf{x}) = \arg\max_{l \in \mathcal{Y}} \mathbb{P}(y = l \mid \mathbf{x})$	$h(\mathbf{x}) = \text{mode}\left\{y^{(i)}\right\}$
Brier	$\hat{\pi}(\mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x} = \mathbf{x})$	$\hat{\pi} = \frac{1}{n} \sum_{i=1}^{n} y^{(i)}$
Bernoulli	$\hat{\pi}(\mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x} = \mathbf{x})$	$\hat{\pi} = \frac{1}{n} \sum_{i=1}^{n} y^{(i)}$
Bernoulli	$\hat{f}(\mathbf{x}) = \log\left(\frac{\mathbb{P}(y=1 \mid \mathbf{x})}{1-\mathbb{P}(y=1 \mid \mathbf{x})}\right)$	$\hat{f} = \ln \frac{n_{+1}}{n_{-1}}$

SUMMARY OF LOSS FUNCTIONS

There are other loss functions for classification tasks, for example:

- Hinge-Loss
- Exponential-Loss

As for regression, loss functions might also be customized to an objective that is defined by an application.