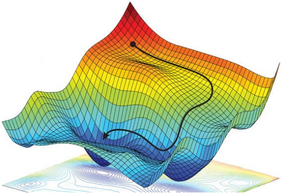
Introduction to Machine Learning

Advanced Classification Losses



Learning goals

- Know advanced classification losses

MARGINS

Losses can be defined on

- Hard labels $h(\mathbf{x}) \in \mathcal{Y}$
- Scores $f(\mathbf{x}) \in \mathbb{R}$
- Probabilities $\pi(\mathbf{x})$

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

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

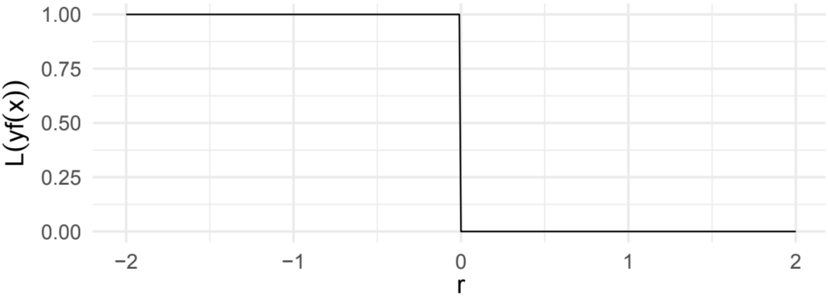
- We can express the 0-1-loss also for a scoring classifier $f(\mathbf{x})$ based on the margin r

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

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.

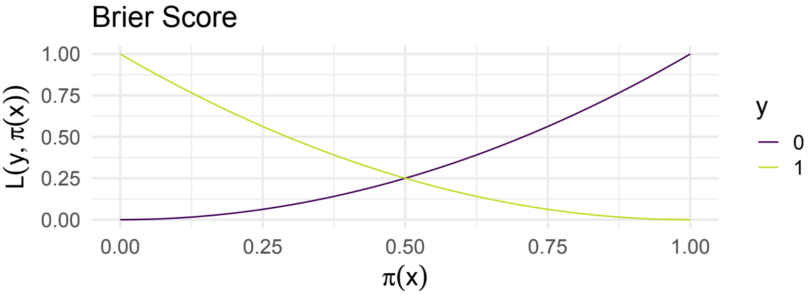


Brier Score

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).

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



BRIER SCORE: POINT-WISE OPTIMUM

The minimizer of the (theoretical) risk $\mathcal{R}(f)$ for the Brier score

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

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

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

BRIER SCORE: POINT-WISE OPTIMUM

We plug in the Brier score

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

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

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).

GINI SPLITTING = BRIER SCORE MINIMIZATION

Interestingly, 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

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

GINI SPLITTING = BRIER SCORE MINIMIZATION

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

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

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

GINI SPLITTING = BRIER SCORE MINIMIZATION

We further simplify the expression to

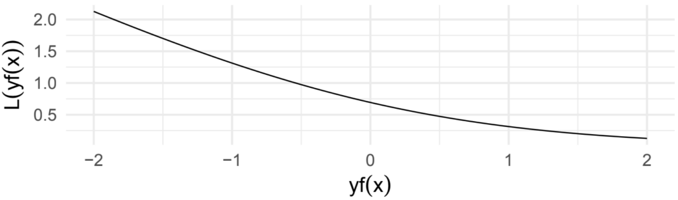
$$
\begin{aligned}
\mathcal{R}(\mathcal{N}) &= \sum_{k=1}^{g} n_{\mathcal{N},k} \left(\frac{n_{\mathcal{N}} - n_{\mathcal{N},k}}{n_{\mathcal{N}}}\right)^2 + (n_{\mathcal{N}} - n_{\mathcal{N},k}) \left(\frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}\right)^2 \\
&= \sum_{k=1}^{g} \frac{n_{\mathcal{N},k}}{n_{\mathcal{N}}}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

Bernoulli Loss

BERNOULLI LOSS

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

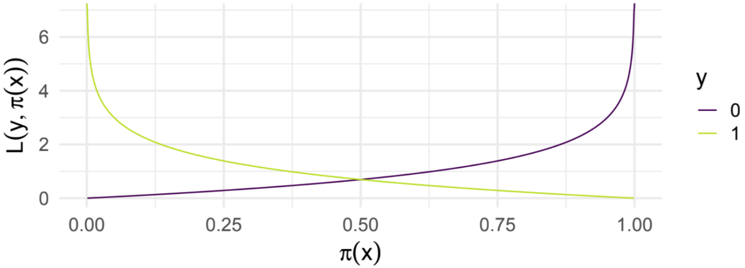
- Two equivalent formulations: Labels $y \in \{-1, 1\}$ or $y \in \{0, 1\}$
- Negative log-likelihood of Bernoulli model, e.g., logistic regression
- Convex, differentiable
- Pseudo-Residuals (0,1 case): $\tilde{r} = y - \frac{1}{1+\exp(-f(\mathbf{x}))}$
  Interpretation: $L1$ distance between 0/1-labels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

BERNOULLI LOSS: POINT-WISE OPTIMUM

The theoretical point-wise optimum for scores under Bernoulli loss is actually the point-wise log-odds:

$$\hat{f}(\mathbf{x}) \quad = \quad \ln\left(\frac{\mathbb{P}(y \mid \mathbf{x} = \mathbf{x})}{1 - \mathbb{P}(y \mid \mathbf{x} = \mathbf{x})}\right).$$

The function is undefined when $P(y \mid \mathbf{x} = \mathbf{x}) = 1$ or $P(y \mid \mathbf{x} = \mathbf{x}) = 0$, but predicts a smooth curve which grows when $P(y \mid \mathbf{x} = \mathbf{x})$ increases and equals 0 when $P(y \mid \mathbf{x} = \mathbf{x}) = 0.5$.

Proof: We consider the case $\mathcal{Y} = \{-1, 1\}$. We have seen that the (theoretical) optimal prediction c for an arbitrary loss function at fixed point $\mathbf{x}$ is

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

BERNOULLI LOSS: POINT-WISE OPTIMUM

We plug in the Bernoulli loss

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

Setting the derivative w.r.t. c to zero yields

$$\begin{aligned} 0 \quad &= \quad -\frac{\exp(-c)}{1 + \exp(-c)} p + \frac{\exp(c)}{1 + \exp(c)} (1 - p) \\ &= \quad -\frac{\exp(-c)}{1 + \exp(-c)} p + \frac{1}{1 + \exp(-c)} (1 - p) \\ &= \quad -p + \frac{1}{1 + \exp(-c)} \\ \Leftrightarrow p \quad &= \quad \frac{1}{1 + \exp(-c)} \\ \Leftrightarrow c \quad &= \quad \ln\left(\frac{p}{1 - p}\right) \end{aligned}$$

BERNOULLI: OPTIMAL CONSTANT MODEL

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

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} y^{(i)}$$

Again, this is the fraction of class-1 observations in the observed data. We can simply prove this again by setting the derivative of the risk to 0 and solving for θ.

BERNOULLI: OPTIMAL CONSTANT MODEL

The optimal constant score model $f(\mathbf{x}) = \theta$ w.r.t. the Bernoulli loss labels from $\mathcal{Y} = \{-1, +1\}$ or $\mathcal{Y} = \{0, 1\}$ is:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \ln \frac{n_{+1}}{n_{-1}} = \ln \frac{n_{+1}/n}{n_{-1}/n}$$

where n_{-1} and n_{+1} are the numbers of negative and positive observations, respectively.

This again shows a tight (and unsurprising) connection of this loss to log-odds.

Proving this is also a (quite simple) exercise.

BERNOULLI-LOSS: NAMING CONVENTION

We have seen three loss functions that are closely related. In the literature, there are different names for the losses:

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

are referred to as Bernoulli, Binomial or logistic loss.

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

is referred to as cross-entropy or log-loss.

For simplicity, we will call all of them **Bernoulli loss**, and rather make clear whether they are defined on labels $y \in \{0, 1\}$ or $y \in \{-1, 1\}$ and on scores $f(\mathbf{x})$ or probabilities $\pi(\mathbf{x})$.

ENTROPY SPLITTING = LOG LOSS MINIMIZATION

The logarithmic loss for multiple classes $y \in \{1, 2, ..., g\}$ is defined as

$$L(y, \pi_k(\mathbf{x})) = \sum_{k=1}^{g} [y = k] \cdot \log\left(\pi_k(\mathbf{x})\right).$$

$$\begin{aligned}
\mathcal{R}(\mathcal{N}) &= \sum_{(\mathbf{x},y)\in\mathcal{N}} \sum_{k=1}^{g} [y = k] \log \pi_k(\mathbf{x}) \\
&= \sum_{k=1}^{g} \sum_{(\mathbf{x},y)\in\mathcal{N}} [y = k] \log \pi_k^{(\mathcal{N})} \\
&= \sum_{k=1}^{g} n_{\mathcal{N}k} \log \pi_k^{(\mathcal{N})} = n_{\mathcal{N}} \sum_{k=1}^{g} \pi_k^{(\mathcal{N})} \log \pi_k^{(\mathcal{N})} = n_{\mathcal{N}} I(\mathcal{N})
\end{aligned}$$

Plugging in the optimal constant $\pi_k(\mathbf{x}) = \pi_k^{(\mathcal{N})}$.

ENTROPY SPLITTING = LOG LOSS MINIMIZATION

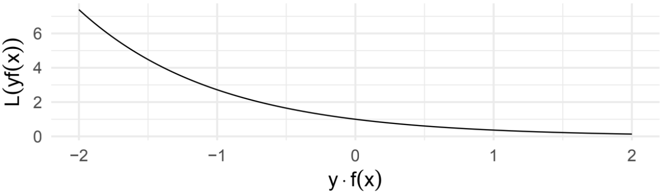
Conclusion:
Stumps/trees with entropy splitting use the same loss function as logistic regression (binary) / softmax regression (multiclass). While logistic regression is based on the hypothesis space of **linear functions**, stumps/trees use **step functions** as hypothesis spaces.

Exponential Loss

CLASSIFICATION LOSSES: EXPONENTIAL LOSS

Another possible choice for a (binary) loss function that is a smooth approximation to the 0-1-loss:

- $L(y, f(\mathbf{x})) = \exp(-yf(\mathbf{x}))$, used in AdaBoost
- Convex, differentiable (thus easier to optimize than 0-1-loss)
- The loss increases exponentially for wrong predictions with high confidence; if the prediction is right with a small confidence only, there, loss is still positive
- No closed-form analytic solution to empirical risk minimization



AUC Loss

CLASSIFICATION LOSSES: AUC-LOSS

- Often AUC is used as an evaluation criterion for binary classifiers
- Let $Y \in \{-1, 1\}$ with observations n_{-1} number of negative and n_1 of positive samples
- The AUC can then be defined as

$$AUC = n_{-1}^{-1} n_1^{-1} \sum_{i:y_i=1} \sum_{j:y_j=-1} I(f_i > f_j)$$

- This is not differentiable wrt f due to $I(f_i > f_j)$
- But the indicator function can be approximated by the distribution function of the triangular distribution on $[-1, 1]$ with mean 0
- However, direct optimization of the AUC is usually not as good as optimization wrt a common loss and tuning via AUC in practice

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.

RISK MINIMIZING FUNCTIONS

Overview of binary classification losses and the corresponding risk minimizing functions:

loss name	loss formula	minimizing function
0-1	$[y \neq h(\mathbf{x})]$	$\hat{h}(\mathbf{x}) = \begin{cases} 1 & \text{if } \pi(\mathbf{x}) > 1/2 \\ -1 & \pi(\mathbf{x}) < 1/2 \end{cases}$
Hinge	$\max\{0, 1 - yf(\mathbf{x})\}$	$\hat{f}(x) = \begin{cases} 1 & \text{if } \pi(\mathbf{x}) > 1/2 \\ -1 & \pi(\mathbf{x}) < 1/2 \end{cases}$
Logistic	$\ln(1 + \exp(-yf(\mathbf{x})))$	$\hat{f}(x) = \ln\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right)$
Cross entropy	$-y\ln(\pi(\mathbf{x}))$ $-(1-y)\ln(1-\pi(\mathbf{x}))$	
Exponential	$\exp(-yf(\mathbf{x}))$	