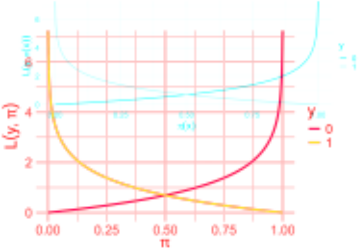
Introduction to Machine Learning

Advanced Risk Minimization

Bernoulli Loss



Learning goals

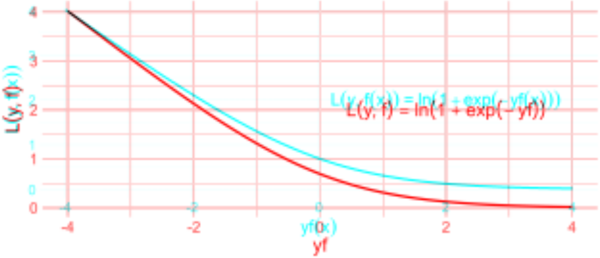
- Know the Bernoulli loss and related losses (log-loss, logistic loss, Binomial loss)
- Derive the risk minimizer
- Derive the optimal constant model
- Understand the connection between log-loss and entropy splitting

BERNOULLI LOSS

$$L(y, f) = \log(1 + \exp(-yf)) \quad \text{for } y \in \{-1, +1\}$$

$$L(y, f) = -yf + \log(1 + \exp(f)) \quad \text{for } y \in \{0, 1\}$$

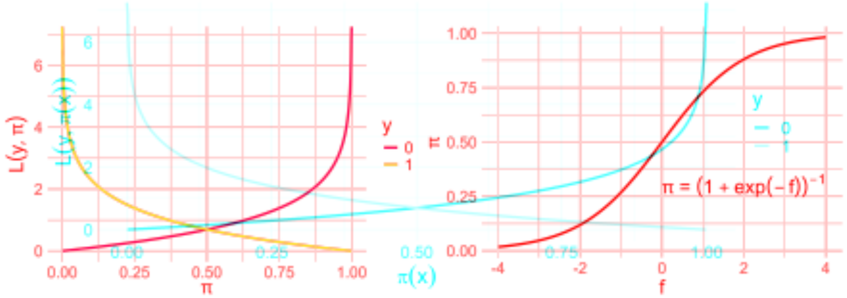
- Two equivalent formulations for different label encodings
- Negative log-likelihood of Bernoulli model, e.g., logistic regression
- Convex, differentiable



BERNOULLI LOSS ON PROBABILITIES

If scores are transformed into probabilities by the logistic function $\pi = (1 + \exp(-f))^{-1}$ (or equivalently if $f = \log\left(\frac{\pi}{1-\pi}\right)$ are the log-odds of π), we arrive at another equivalent formulation of the loss, where y is again encoded as $\{0, 1\}$:

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



BERNOULLI LOSS: RISK MINIMIZER

The risk minimizer for the Bernoulli loss defined for probabilistic classifiers $\pi(\mathbf{x})$ and on $y \in \{0, 1\}$ is

$$\pi^*(\mathbf{x}) = \eta(\mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x} = \mathbf{x}).$$

Proof: We can write the risk for binary y as follows:

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

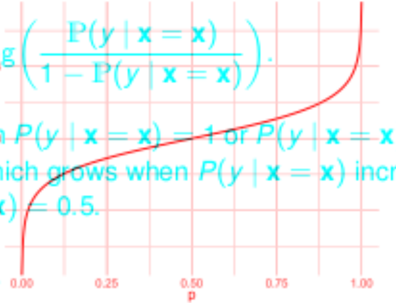
with $\eta(\mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x})$ (see section on the 0-1-loss for more details).
For a fixed $\mathbf{x}$ we compute the point-wise optimal value c by setting the derivative to 0:

$$\begin{aligned}
\frac{\partial}{\partial c} \left(- \log c \cdot \eta(\mathbf{x}) - \log(1 - c) \cdot (1 - \eta(\mathbf{x})) \right) &= 0 \\
- \frac{\eta(\mathbf{x})}{c} + \frac{1 - \eta(\mathbf{x})}{1 - c} &= 0 \\
\frac{-\eta(\mathbf{x}) + \eta(\mathbf{x})c + c - \eta(\mathbf{x})c}{c(1 - c)} &= 0 \\
c &= \eta(\mathbf{x}).
\end{aligned}$$

BERNOULLI LOSS: RISK MINIMIZER / 2

The risk minimizer for the Bernoulli loss defined on $y \in \{-1, 1\}$ and scores $f(\mathbf{x})$ is the point-wise log-odds; i.e. the logit function (inverse of logistic function) of $p(\mathbf{x}) = \mathbb{P}(y \mid \mathbf{x} = \mathbf{x})$:

$$f^*(\mathbf{x}) = \log\left(\frac{\mathbb{P}(y \mid \mathbf{x} = \mathbf{x})}{1 - \mathbb{P}(y \mid \mathbf{x} = \mathbf{x})}\right).$$



$f^*(\mathbf{x}) = \log(\frac{p(\mathbf{x})}{1-p(\mathbf{x})})$

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: As before we minimize

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 $P(y \mid \mathbf{x} = \mathbf{x}) = 0.5$.

$$\begin{aligned}\mathcal{R}(f) &= \mathbb{E}_x\left[L(1, f(\mathbf{x})) \cdot \eta(\mathbf{x}) + L(-1, f(\mathbf{x})) \cdot (1 - \eta(\mathbf{x}))\right] \\ &= \mathbb{E}_x\left[\log(1 + \exp(-f(\mathbf{x})))\eta(\mathbf{x}) + \log(1 + \exp(f(\mathbf{x})))(1 - \eta(\mathbf{x}))\right]\end{aligned}$$

BERNOULLI LOSS: RISK MINIMIZER / 3

Proof: As before we minimize

$$
\begin{aligned}
\mathcal{R}(f) &= \mathbb{E}_x\left[L(1, f(\mathbf{x})) \cdot \eta(\mathbf{x}) + L(-1, f(\mathbf{x})) \cdot (1 - \eta(\mathbf{x}))\right] \\
&= \mathbb{E}_x\left[\log(1 + \exp(-f(\mathbf{x})))\eta(\mathbf{x}) + \log(1 + \exp(f(\mathbf{x})))(1 - \eta(\mathbf{x}))\right]
\end{aligned}
$$

For a fixed $\mathbf{x}$ we compute the point-wise optimal value c by setting the derivative to 0:

$$
\begin{aligned}
\frac{\partial}{\partial c} \log(1 + \exp(-c))\eta(\mathbf{x}) + \log(1 + \exp(c))(1 - \eta(\mathbf{x})) &= 0 \\
-\frac{\exp(-c)}{1 + \exp(-c)}\eta(\mathbf{x}) + \frac{\exp(c)}{1 + \exp(c)}(1 - \eta(\mathbf{x})) &= 0 \\
-\frac{\exp(-c)\eta(\mathbf{x}) - 1 + \eta(\mathbf{x})}{1 + \exp(-c)} &= 0 \\
-\eta(\mathbf{x}) + \frac{1}{1 + \exp(-c)} &= 0 \\
c &= \log\left(\frac{\eta(\mathbf{x})}{1 - \eta(\mathbf{x})}\right)
\end{aligned}
$$

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:

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

- 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.
- 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:

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

- where n_- and n_+ 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:

$$L(y, f) = \log(1 + \exp(-yf)) \quad \text{for } y \in \{-1, +1\}$$
$$L(y, f) = -y \cdot f + \log(1 + \exp(f)) \quad \text{for } y \in \{0, 1\}$$
$$L(y, \pi) = -y \log(\pi) - (1 - y) \log(1 - \pi) \quad \text{for } y \in \{0, 1\}$$
$$L(y, \pi) = -\frac{1 + y}{2} \log(\pi) - \frac{1 - y}{2} \log(1 - \pi) \quad \text{for } y \in \{-1, +1\}$$

are equally referred to as Bernoulli, Binomial, logistic, log loss, or cross-entropy (showing equivalence is a simple exercise).

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

where n_- and n_+ are the number 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:

$$L(y, f(\mathbf{x})) = \log(1 + \exp(-yf(\mathbf{x}))) \quad \text{for } y \in \{-1, +1\}$$
$$L(y, f(\mathbf{x})) = -y \cdot f(\mathbf{x}) + \log(1 + \exp(f(\mathbf{x}))) \quad \text{for } y \in \{0, 1\}$$

are referred to as Bernoulli, Binomial or logistic loss.

$$L(y, \pi(\mathbf{x})) = -y \log(\pi(\mathbf{x})) - (1 - y) \log(1 - \pi(\mathbf{x})) \quad \text{for } y \in \{0, 1\}$$

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

We usually refer to all of them as **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 $\pi(\mathbf{x})$.

BERNOULLI LOSS MIN = ENTROPY SPLITTING

When fitting a tree we minimize the risk within each node $\mathcal{N}$ by risk minimization and predict the optimal constant. Another approach that is common in literature is to minimize the average node impurity $\text{Imp}(\mathcal{N})$.

Claim: Entropy splitting $\text{Imp}(\mathcal{N}) = -\sum_{k=1}^{g} \pi_k^{(\mathcal{N})} \log \pi_k^{(\mathcal{N})}$ is equivalent to minimize risk measured by the Bernoulli loss.

Note that $\pi_k^{(\mathcal{N})} := \frac{1}{n_\mathcal{N}} \sum\limits_{(\mathbf{x},y)\in\mathcal{N}} [y = k]$.

Proof: To prove this we show that the risk related to a subset of observations $\mathcal{N} \subseteq \mathcal{D}$ fulfills

$$\mathcal{R}(\mathcal{N}) = n_\mathcal{N} \text{Imp}(\mathcal{N}),$$

where $\mathcal{R}(\mathcal{N})$ is calculated w.r.t. the (multiclass) Bernoulli loss

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

BERNOULLI LOSS MIN = ENTROPY SPLITTING / 2

$$\begin{aligned}
\mathcal{R}(\mathcal{N}) &= \sum_{(\mathbf{x},y)\in\mathcal{N}} \left(-\sum_{k=1}^{g} [y=k] \log \pi_k(\mathbf{x}) \right) \\
&\overset{(*)}{=} -\sum_{k=1}^{g} \sum_{(\mathbf{x},y)\in\mathcal{N}} [y=k] \log \pi_k^{(\mathcal{N})} \\
&= -\sum_{k=1}^{g} \log \pi_k^{(\mathcal{N})} \underbrace{\sum_{(\mathbf{x},y)\in\mathcal{N}} [y=k]}_{n_{\mathcal{N}} \cdot \pi_k^{(\mathcal{N})}} \\
&= -n_{\mathcal{N}} \sum_{k=1}^{g} \pi_k^{(\mathcal{N})} \log \pi_k^{(\mathcal{N})} = n_{\mathcal{N}} \operatorname{Imp}(\mathcal{N}),
\end{aligned}$$

where in $^{(*)}$ the optimal constant per node $\pi_k^{(\mathcal{N})} = \frac{1}{n_{\mathcal{N}}} \sum_{(\mathbf{x},y)\in\mathcal{N}} [y=k]$ was plugged in.