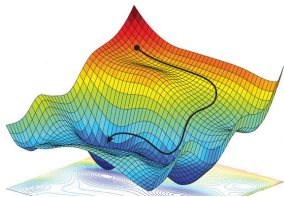


# Introduction to Machine Learning

## Bernoulli Loss



### Learning goals

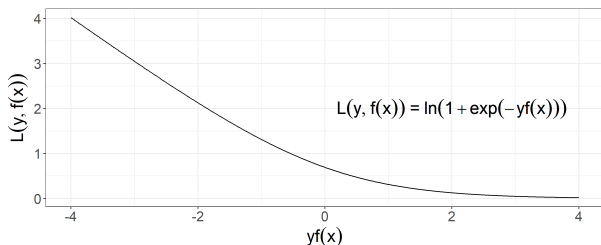
- Know the Bernoulli loss and related losses (log-loss, logistic loss, Binomial loss)
- Derive the risk minimizer
- Derive the optimal constant model

# BERNOULLI LOSS

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

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

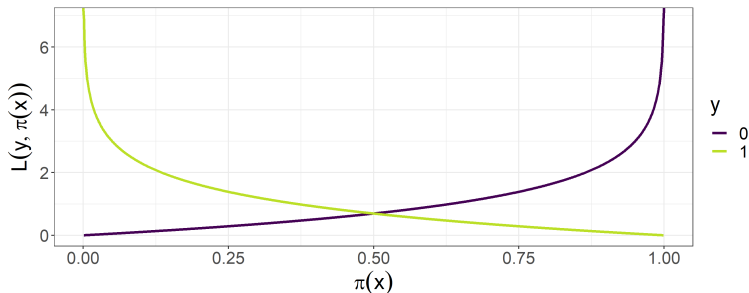
- Two equivalent formulations for different label encodings
- Negative log-likelihood of Bernoulli model, e.g., logistic regression
- Convex, differentiable
- Pseudo-residuals (0/1 case):  $\tilde{r} = y - \frac{1}{1 + \exp(-f(\mathbf{x}))}$   
Interpretation: L1 distance between 0/1-labels and posterior 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, where  $y$  is again encoded as  $\{0, 1\}$ :

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



# BERNOULLI LOSS: RISK MINIMIZER

To derive the risk minimizer for the Bernoulli loss for the formulation

$$L(y, \pi(\mathbf{x})) = -y \log(\pi(\mathbf{x})) - (1 - y) \log(1 - \pi(\mathbf{x})),$$

we again make use of the law of total expectation

$$\begin{aligned}\mathcal{R}(f) &= \mathbb{E}_{xy} [L(y, f(\mathbf{x}))] = \mathbb{E}_x [\mathbb{E}_{y|x} [L(y, f(\mathbf{x}))]] \\ &= \mathbb{E}_x \left[ \sum_{k \in \mathcal{Y}} L(k, f(\mathbf{x})) \mathbb{P}(y = k \mid \mathbf{x} = \mathbf{x}) \right].\end{aligned}$$

In the binary case this becomes

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

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

# BERNOULLI LOSS: RISK MINIMIZER

We fix a specific  $\mathbf{x}$  and compute the point-wise optimal value  $c$  by setting the derivative to 0:

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

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

# BERNOULLI LOSS: RISK MINIMIZER

To derive the risk minimizer for the Bernoulli loss we again compute the point-wise optimum for a fixed  $\mathbf{x}$ . The point-wise log-odds:

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

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

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

# BERNOULLI LOSS: RISK MINIMIZER

We plug in the Bernoulli loss

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

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

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

# 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  $\theta$ .



# 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:

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

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:

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

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

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.

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  $\pi(\mathbf{x})$ .

# LOG LOSS MINIMIZATION = ENTROPY SPLITTING

Entropy splitting in trees is equivalent to minimizing the log-loss of a node. The logarithmic loss for multiple classes  $y \in \{1, 2, \dots, g\}$  is defined as

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

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

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