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

BERNOULLI LOSS

$$\begin{array}{lll} L(y,f) &=& \log(1+\exp(-y \cdot f)) & \text{for } y \in \{-1,+1\} \\ L(y,f) &=& -y \cdot f + \log(1+\exp(f)) & \text{for } y \in \{0,1\} \end{array}$$

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



× 0 0 × 0 × ×

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 $\{0, 1\}$:

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



× × 0 × ×

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:

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

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

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

× 0 0 × × ×

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 $p(\mathbf{x}) = \mathbb{P}(y | \mathbf{x} = \mathbf{x})$:



× 0 0 × × ×

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

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 **x** we compute the point-wise optimal value c by setting the derivative to 0:

$$\frac{\partial}{\partial c} \log(1 + \exp(-c))\eta(\mathbf{x}) + \log(1 + \exp(c))(1 - \eta(\mathbf{x})) = 0$$

$$-\frac{\exp(-c)}{1 + \exp(-c)}\eta(\mathbf{x}) + \frac{\exp(c)}{1 + \exp(c)}(1 - \eta(\mathbf{x})) = 0$$

$$-\frac{\exp(-c)\eta(\mathbf{x}) - 1 + \eta(\mathbf{x})}{1 + \exp(-c)} = 0$$

$$-\eta(\mathbf{x}) + \frac{1}{1 + \exp(-c)} = 0$$

$$c = \log\left(\frac{\eta(\mathbf{x})}{1 - \eta(\mathbf{x})}\right)$$

X

 $\times \times$

©

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:

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

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:

$$\hat{ heta} = rgmin_{ heta} \mathcal{R}_{ ext{emp}}(oldsymbol{ heta}) = \log rac{n_+}{n_-} = \log rac{n_+/n}{n_-/n}$$

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:

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

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

× × ×