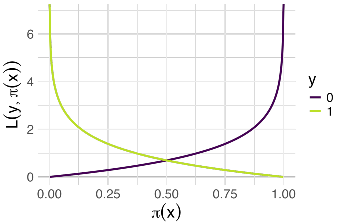
Introduction to Machine Learning

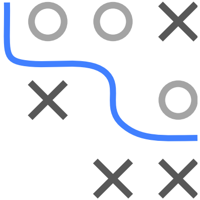
Advanced Risk Minimization
Optimal constant model for the empirical log loss risk (Deep-Dive)



Learning goals

- Derive the optimal constant model for the binary empirical log loss risk
- Derive the optimal constant model for the empirical multiclass log loss risk

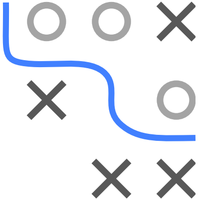
BINARY LOG LOSS: EMP. RISK MINIMIZER



Given $n \in \mathbb{N}$ observations $y^{(1)}, \cdots, y^{(n)} \in \mathcal{Y} = \{0, 1\}$ we want to determine the optimal constant model for the empirical log loss risk.

$$\underset{\theta \in (0,1)}{\arg\min}\, \mathcal{R}_{\text{emp}} = \underset{\theta \in (0,1)}{\arg\min} - \sum_{i=1}^{n} y^{(i)} \log(\theta) + (1 - y^{(i)}) \log(1 - \theta).$$

BINARY LOG LOSS: EMP. RISK MINIMIZER



The minimizer can be found by setting the derivative to zero, i.e.,

$$
\begin{aligned}
\frac{d}{d\theta}\mathcal{R}_{\text{emp}} &= -\sum_{i=1}^{n} \frac{y^{(i)}}{\theta} - \frac{1-y^{(i)}}{1-\theta} \overset{!}{=} 0 \\
&\iff -\sum_{i=1}^{n} y^{(i)}(1-\theta) - \theta(1-y^{(i)}) \overset{!}{=} 0 \\
&\iff -\sum_{i=1}^{n} \left(y^{(i)} - \theta\right) \overset{!}{=} 0
\end{aligned}
$$

$$\Rightarrow \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} y^{(i)} \quad \in \quad (0,1) \checkmark \text{(assuming both labels occur).}$$

MULTICLASS LOG LOSS: EMP. RISK MINIMIZER

Given $n \in \mathbb{N}$ observations $y^{(1)}, \cdots, y^{(n)} \in \mathcal{Y} = \{1, \dots, g\}$ with $g \in \mathbb{N}_{>1}$ we want to determine the optimal constant model $\theta \in (0,1)^g$ for the empirical log loss risk

$$\begin{aligned} \underset{\theta \in (0,1)^g}{\arg\min} \mathcal{R}_{\text{emp}} &= \underset{\theta \in (0,1)^g}{\arg\min} - \sum_{i=1}^{n} \sum_{j=1}^{g} \mathbb{1}_{\{y^{(i)}=j\}} \log(\theta_j) \\ \text{s.t.} &\quad \sum_{j=1}^{g} \theta_j = 1. \end{aligned}$$

We can solve this constrained optimization problem by plugging the constraint into the risk (we could also use Lagrange multipliers), i.e., we replace θ_g (this is an arbitrary choice) such that $\theta_g = 1 - \sum_{j=1}^{g-1} \theta_j$.

MULTICLASS LOG LOSS: EMP. RISK MINIMIZER

With this, we find the equivalent optimization problem

$$
\begin{aligned}
\underset{\theta \in (0,1)^{g-1}}{\arg\min}\ \mathcal{R}_{\text{emp}} \quad = \quad & \underset{\theta \in (0,1)^{g-1}}{\arg\min} - \sum_{i=1}^{n} \sum_{j=1}^{g-1} \mathbb{1}_{\{y^{(i)}=j\}} \log(\theta_j) \\
& + \mathbb{1}_{\{y^{(i)}=g\}} \log\left(1 - \sum\nolimits_{j=1}^{g-1} \theta_j\right) \\
\text{s.t.} \quad & \sum\nolimits_{j=1}^{g-1} \theta_j < 1.
\end{aligned}
$$

For $j \in \{1, \ldots, g-1\}$, the j-th partial derivative of our objective

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} \mathcal{R}_{\text{emp}} \quad &= \quad - \sum_{i=1}^{n} \mathbb{1}_{\{y^{(i)}=j\}} \frac{1}{\theta_j} - \mathbb{1}_{\{y^{(i)}=g\}} \frac{1}{1 - \sum_{j=1}^{g-1} \theta_j} \\
&= \quad - \frac{n_j}{\theta_j} + \frac{n_g}{\theta_g}
\end{aligned}
$$

where n_k with $k \in \{1, \ldots, g\}$ is the number of label k in y and we assume that $n_k > 0$.

MULTICLASS LOG LOSS: EMP. RISK MINIMIZER

For the minimizer, it must hold for $j \in \{1, \ldots, g-1\}$ that

$$
\begin{aligned}
\frac{\partial}{\partial \theta_j} \mathcal{R}_{\text{emp}} &\overset{!}{=} 0 \\
\iff -n_j\theta_g + n_g\theta_j &\overset{!}{=} 0 \\
\Rightarrow \sum_{j=1}^{g-1} (-n_j\theta_g + n_g\theta_j) &\overset{!}{=} 0 \\
\iff -(n - n_g)\theta_g + n_g(1 - \theta_g) &\overset{!}{=} 0 \\
\iff -n\theta_g + n_g &\overset{!}{=} 0 \\
\Rightarrow \hat{\theta}_g = \frac{n_g}{n} & \qquad \in (0, 1) \checkmark \\
\Rightarrow \forall j \in \{1, \ldots, g-1\}: \quad \hat{\theta}_j = \frac{\hat{\theta}_g n_j}{n_g} = \frac{n_j}{n} & \qquad \in (0, 1) \checkmark .
\end{aligned}
$$

$$
\left(\Rightarrow \textstyle\sum_{j=1}^{g-1} \hat{\theta}_j = 1 - \hat{\theta}_g = 1 - \frac{n_g}{n} < 1 \checkmark \right)
$$

CONVEXITY

Finally, we check that we indeed found a minimizer by showing that $\mathcal{R}_{\text{emp}}$ is convex for the multiclass case (binary is a special case of this):

The Hessian of $\mathcal{R}_{\text{emp}}$

$$\nabla^2_\theta \mathcal{R}_{\text{emp}} = \begin{pmatrix} \frac{n_1}{\theta_1^2} & 0 & \dots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \dots & 0 & \frac{n_{g-1}}{\theta_{g-1}^2} \end{pmatrix}$$

is positive definite since all its eigenvalues

$$\lambda_j = \frac{n_j}{\theta_j^2} > 0 \quad \forall j \in \{1, \dots, g-1\}.$$

From this, it follows that $\mathcal{R}_{\text{emp}}$ is (strictly) convex.