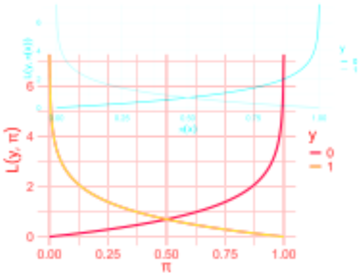
Introduction to Machine Learning

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

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 / 2

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

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

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

MULTICLASS LOG LOSS: EMP. RISK MINIMIZER / 2

With this, we find the equivalent optimization problem

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

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 / 3

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

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

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