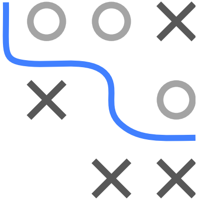
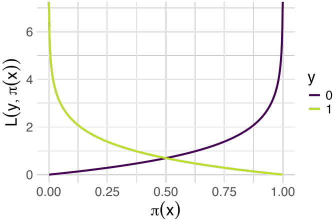
Introduction to Machine Learning



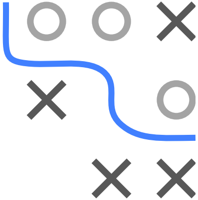
Advanced Risk Minimization
Classification (Deep-Dive)



Learning goals

- Equivalence of different loss formulations
- Risk minimizer on scores
- Optimal constant model for the binary empirical log loss risk
- Optimal constant model for the empirical multiclass log loss risk

EQUIVALENCE OF LOSS FORMULATIONS



- Starting point Bernoulli loss on probs:

$$L\left(y, \pi(\mathbf{x})\right) = -y \log\left(\pi(\mathbf{x})\right) - (1-y)\log\left(1-\pi(\mathbf{x})\right) \;\;,\; y \in \{0,1\}$$

- Loss on scores $f(\mathbf{x}) = \log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) \Leftrightarrow \pi(\mathbf{x}) = (1+\exp(-f(\mathbf{x})))^{-1}$:

$$\begin{aligned}
L\left(y, \pi(\mathbf{x})\right) &= -y\big(\log\left(\pi(\mathbf{x})\right) - \log\left(1-\pi(\mathbf{x})\right)\big) - \log\left(1-\pi(\mathbf{x})\right) \\
&= -y\log\left(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right) - \log\left(1 - \frac{1}{1+\exp(-f(\mathbf{x}))}\right) \\
&= -yf(\mathbf{x}) - \log\left(\frac{\exp(-f(\mathbf{x}))}{1+\exp(-f(\mathbf{x}))}\right) \\
&= -yf(\mathbf{x}) - \log\left(\frac{1}{1+\exp(f(\mathbf{x}))}\right) \\
&= -yf(\mathbf{x}) + \log\left(1+\exp(f(\mathbf{x}))\right)
\end{aligned}$$

- Yields equivalent loss formulation

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

EQUIVALENCE OF LOSS FORMULATIONS

- For $y \in \{-1, +1\}$ convert labels using $y' = (y + 1)/2$
- Bernoulli loss on probs with $y \in \{-1, +1\}$:

$$L(y, \pi(\mathbf{x})) = -\frac{1+y}{2}\log(\pi(\mathbf{x})) - \frac{1-y}{2}\log(1-\pi(\mathbf{x})), \quad y \in \{-1,+1\}$$

- For $y \in \{-1, +1\}$ loss on scores becomes:

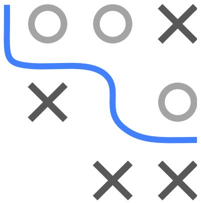
$$L(y, f(\mathbf{x})) = \log(1 + \exp(-y \cdot f(\mathbf{x})))$$

- For $y = -1$ plug $y' = 0$ in $L(y, f(\mathbf{x}))$ for $y \in \{0, 1\}$ loss:
  $L(0, f(\mathbf{x})) = \log(1 + \exp(f(\mathbf{x}))$ ✓
- For $y = y' = 1$:

$$\begin{aligned} L(1, f(\mathbf{x})) &= -1 \cdot f(\mathbf{x}) + \log(1 + \exp(f(\mathbf{x}))) \\ &= \log(1 + \exp(f(\mathbf{x}))) - \log(\exp(f(\mathbf{x}))) \\ &= \log(1 + \exp(-f(\mathbf{x})) \quad \checkmark \end{aligned}$$

NAMING CONVENTIONS

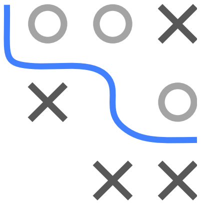
We have seen several closely related loss functions:



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

They are equally referred to as Bernoulli-, Binomial-, logistic-, log-, or cross-entropy loss

PROOF RISK MINIMIZER ON SCORES



For $y \in \{0, 1\}$ the pointwise RM on scores is

$$f^*(\mathbf{x}) = \log\left(\eta(\mathbf{x})/(1 - \eta(\mathbf{x}))\right)$$

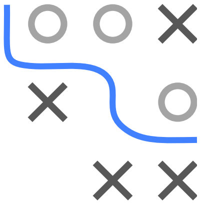
Proof: As before we minimize

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

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

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

BINARY LOG LOSS: EMP. RISK MINIMIZER



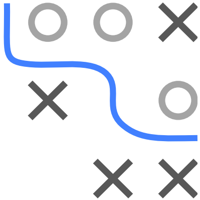
Given $n \in \mathbb{N}$ observations $y^{(1)}, \cdots, y^{(n)} \in \mathcal{Y} = \{0, 1\}$ we want to determine the optimal constant model for the empirical log loss risk.

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

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

$$\begin{aligned}
\frac{d}{d\theta}\mathcal{R}_{\text{emp}} &= -\sum_{i=1}^{n} \frac{y^{(i)}}{\theta} - \frac{1 - y^{(i)}}{1 - \theta} && \overset{!}{=} 0 \\
\iff & -\sum_{i=1}^{n} y^{(i)}(1 - \theta) - \theta(1 - y^{(i)}) && \overset{!}{=} 0 \\
\iff & -\sum_{i=1}^{n} (y^{(i)} - \theta) && \overset{!}{=} 0 \\
\Rightarrow \hat{\theta} &= \frac{1}{n}\sum_{i=1}^{n} y^{(i)} \qquad \in (0, 1) \checkmark \text{(assuming both labels occur)}
\end{aligned}$$

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

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

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

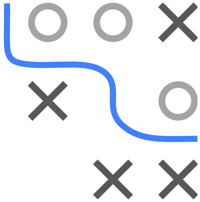
$$
\begin{aligned}
\underset{\theta \in (0,1)^{g-1}}{\arg\min} \mathcal{R}_{\text{emp}} = \underset{\theta \in (0,1)^{g-1}}{\arg\min} & - \sum_{i=1}^{n} \sum_{j=1}^{g-1} \mathbb{1}_{\{y^{(i)}=j\}} \log(\theta_j) \\
& + \mathbb{1}_{\{y^{(i)}=g\}} \log(1 - \textstyle\sum_{j=1}^{g-1} \theta_j) \\
\text{s.t.}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

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

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

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

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

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 $\mathcal{R}_{\text{emp}}$

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

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