Introduction to Machine Learning

Multiclass Classification and Losses

Multiclass Classification and Losses

Learning goals

- Know what multiclass means and which types of classifiers exist
- Know the MC 0-1-loss
- Know the MC brier score
- Know the MC logarithmic loss

Learning goals

- Know what multiclass means and which types of classifiers exist
- Know the MC 0-1-loss
- Know the MC brier score
- Know the MC logarithmic loss

REVISION: RISK FOR CLASSIFICATION

Goal: Find a model $f : \mathcal{X} \to \mathbb{R}^g$, where g is the number of classes, that minimizes the expected loss over random variables $(\mathbf{x}, y) \sim P_{xy}$

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

The optimal model for a loss function $L(y, f(\mathbf{x}))$ is

$$\hat{f}(\mathbf{x}) = \arg\min_{f \in \mathcal{H}} \sum_{k \in \mathcal{Y}} L(k, f(\mathbf{x})) \mathbb{P}(y = k | \mathbf{x} = \mathbf{x}) .$$

Because we usually do not know P_{xy}, we minimize the **empirical risk** as an approximation to the **theoretical** risk

$$\hat{f} = \arg\min_{f \in \mathcal{H}} \mathcal{R}_{\text{emp}}(f) = \arg\min_{f \in \mathcal{H}} \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right) .$$

0-1-LOSS

We have already seen that optimizer $\hat{h}(\mathbf{x})$ of the theoretical risk using the 0-1-loss

$$L(y, h(\mathbf{x})) = \mathbb{1}_{\{y \neq h(\mathbf{x})\}}$$

is the Bayes optimal classifier, with

$$\hat{h}(\mathbf{x}) = \arg\max_{l \in \mathcal{Y}} \mathbb{P}(y = l \mid \mathbf{x} = \mathbf{x})$$

and the optimal constant model (featureless predictor)

$$h(\mathbf{x}) = k; \; k \in \{1; 2; ...; g\}$$

is the classifier that predicts the most frequent class $k \in \{1; 2; ...; g\}$ in the data

$$h(\mathbf{x}) = \text{mode}\left\{y^{(i)}\right\}.$$

MC BRIER SCORE

The (binary) Brier score generalizes to the multiclass Brier score that is defined on a vector of class probabilities $(\pi_1(\mathbf{x}), \ldots, \pi_g(\mathbf{x}))$

$$L(y, \pi(x)) = \sum_{k=1}^{g} \left(\mathbb{1}_{\{y=k\}} - \pi_k(\mathbf{x})\right)^2 .$$

Optimal constant prob vector $\pi(\mathbf{x}) = (\theta_1, \ldots, \theta_g)$:

$$\boldsymbol{\theta} = \underset{\boldsymbol{\theta} \in \mathbb{R}^g, \sum \theta_k = 1}{\arg\min} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) \quad \text{with} \quad \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \left(\sum_{i=1}^{n} \sum_{k=1}^{g} \left(\mathbb{1}_{\{y^{(i)}=k\}} - \theta_k\right)^2\right)$$

We solve this by setting the derivative w.r.t. θ_k to 0

$$\frac{\partial \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})}{\partial \theta_k} = -2 \cdot \sum_{i=1}^{n} (\mathbb{1}_{\{y^{(i)}=k\}} - \theta_k) = 0 \Rightarrow \hat{\pi}_k(\mathbf{x}) = \hat{\theta}_k = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{y^{(i)}=k\}},$$

being the fraction of class-k observations.

NB: We naively ignored the constraints! But since $\sum_{k=1}^{g} \hat{\theta}_k = 1$ holds for the minimizer of the unconstrained problem, we are fine. Could have also used Lagrange multipliers!

MC BRIER SCORE / 2

Claim: For $g = 2$ the MC Brier score is exactly twice as high as the binary Brier score, defined as $(\pi_1(\mathbf{x}) - y)^2$.

Proof:

$$L(y, \pi(x)) = \sum_{k=0}^{1} \left(\mathbb{1}_{\{y=k\}} - \pi_k(\mathbf{x})\right)^2$$

For $y = 0$:

$$\begin{aligned} L(y, \pi(x)) &= (1 - \pi_0(\mathbf{x}))^2 + (0 - \pi_1(\mathbf{x}))^2 = (1 - (1 - \pi_1(\mathbf{x})))^2 + \pi_1(\mathbf{x})^2 \\ &= \pi_1(\mathbf{x})^2 + \pi_1(\mathbf{x})^2 = 2 \cdot \pi_1(\mathbf{x})^2 \end{aligned}$$

For $y = 1$:

$$\begin{aligned} L(y, \pi(x)) &= (0 - \pi_0(\mathbf{x}))^2 + (1 - \pi_1(\mathbf{x}))^2 = (-(1 - \pi_1(\mathbf{x})))^2 + (1 - \pi_1(\mathbf{x}))^2 \\ &= 1 - 2 \cdot \pi_1(\mathbf{x}) + \pi_1(\mathbf{x})^2 + 1 - 2 \cdot \pi_1(\mathbf{x}) + \pi_1(\mathbf{x})^2 \\ &= 2 \cdot (1 - 2 \cdot \pi_1(\mathbf{x}) + \pi_1(\mathbf{x})^2) = 2 \cdot (1 - \pi_1(\mathbf{x}))^2 = 2 \cdot (\pi_1(\mathbf{x}) - 1)^2 \end{aligned}$$

$$L(y, \pi(x)) = \begin{cases} 2 \cdot \pi_1(\mathbf{x})^2 & \text{for } y = 0 \\ 2 \cdot (\pi_1(\mathbf{x}) - 1)^2 & \text{for } y = 1 \end{cases} = 2 \cdot (\pi_1(\mathbf{x}) - y)^2$$

LOGARITHMIC LOSS (LOG-LOSS) / 2

Claim: For $g = 2$ the log-loss is equal to the Bernoulli loss, defined as

$$L_{0,1}(y, \pi_1(\mathbf{x})) = -y log(\pi_1(\mathbf{x})) - (1 - y) log(1 - \pi_1(\mathbf{x}))$$

Proof:

$$\begin{aligned}
L_{0,1}(y, \pi_1(\mathbf{x})) &= -y log(\pi_1(\mathbf{x})) - (1 - y) log(1 - \pi_1(\mathbf{x})) \\
&= -y log(\pi_1(\mathbf{x})) - (1 - y) log(\pi_0(\mathbf{x})) \\
&= -\mathbb{1}_{\{y=1\}} log(\pi_1(\mathbf{x})) - \mathbb{1}_{\{y=0\}} log(\pi_0(\mathbf{x})) \\
&= -\sum_{k=0}^{1} \mathbb{1}_{\{y=k\}} \log(\pi_k(\mathbf{x})) = L(y, \pi(x))
\end{aligned}$$