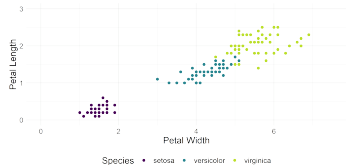


Introduction to Machine Learning

Multiclass Classification

Multiclass Classification and Losses



Learning goals

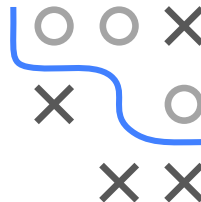
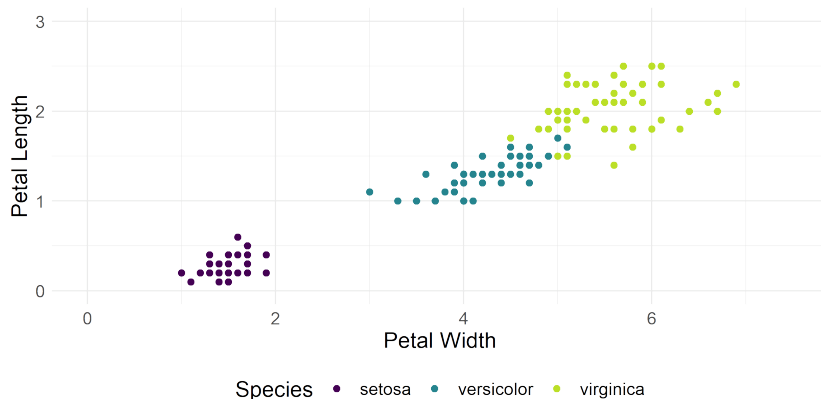
- Know what multiclass means and which types of classifiers exist
- Know the MC 0-1-loss
- Know the MC brier score
- Know the MC logarithmic loss

MULTICLASS CLASSIFICATION

Scenario: Multiclass classification with $g > 2$ classes

$$\mathcal{D} \subset (\mathcal{X} \times \mathcal{Y})^n, \mathcal{Y} = \{1, \dots, g\}$$

Example: Iris dataset with $g = 3$



TYPES OF CLASSIFIERS

- We already saw losses for binary classification tasks. Now we will consider losses for **multiclass classification** tasks.
- For multiclass classification, loss functions will be defined on
 - vectors of scores

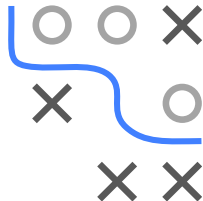
$$f(\mathbf{x}) = (f_1(\mathbf{x}), \dots, f_g(\mathbf{x}))$$

- vectors of probabilities

$$\pi(\mathbf{x}) = (\pi_1(\mathbf{x}), \dots, \pi_g(\mathbf{x}))$$

- hard labels

$$h(\mathbf{x}) = k, k \in \{1, 2, \dots, g\}$$



ONE-HOT ENCODING

- Multiclass outcomes y with classes $1, \dots, g$ are often transformed to g binary (1/0) outcomes using

$$\text{with } \mathbb{1}_{\{y=k\}} = \begin{cases} 1 & \text{if } y = k \\ 0 & \text{otherwise} \end{cases}$$

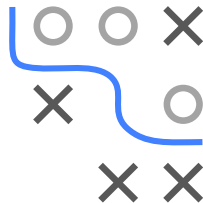
- One-hot encoding does not lose any information contained in the outcome.

Example: Iris

Species	Species.setosa	Species.versicolor	Species.virginica
versicolor	0	1	0
virginica	0	0	1
versicolor	0	1	0
versicolor	0	1	0
setosa	1	0	0
setosa	1	0	0



0-1-Loss



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

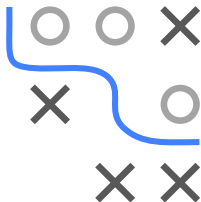
$$\hat{h}(\mathbf{x}) = \arg \max_{l \in \mathcal{Y}} \mathbb{P}(y = l \mid \mathbf{x} = \mathbf{x})$$

and the optimal constant model (featureless predictor)

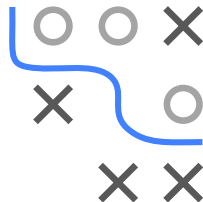
$$h(\mathbf{x}) = k, k \in \{1, 2, \dots, g\}$$

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

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



MC Brier Score



MC BRIER SCORE

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

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

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

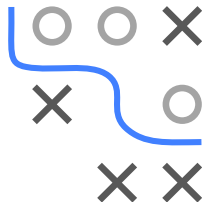
$$\theta = \arg \min_{\theta \in \mathbb{R}^g, \sum \theta_k = 1} \mathcal{R}_{\text{emp}}(\theta) \quad \text{with} \quad \mathcal{R}_{\text{emp}}(\theta) = \left(\sum_{i=1}^n \sum_{k=1}^g (\mathbb{1}_{\{y^{(i)}=k\}} - \theta_k)^2 \right)$$

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

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

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:

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

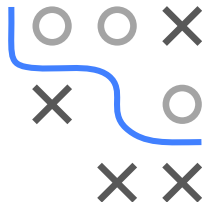
For $y = 0$:

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

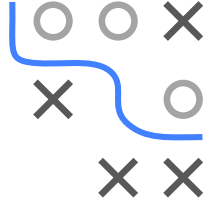
For $y = 1$:

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

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



Logarithmic Loss



LOGARITHMIC LOSS (LOG-LOSS)

The generalization of the Binomial loss (logarithmic loss) for two classes is the multiclass **logarithmic loss** / **cross-entropy loss**:

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

with $\pi_k(\mathbf{x})$ denoting the predicted probability for class k .

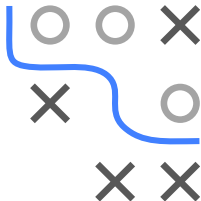
Optimal constant prob vector $\pi(\mathbf{x}) = (\theta_1, \dots, \theta_g)$:

$$\pi_k(\mathbf{x}) = \theta_k = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{y^{(i)}=k\}},$$

being the fraction of class- k observations.

Proof: Exercise.

In the upcoming section we will see how this corresponds to the (multinomial) **softmax regression**.



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:

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

