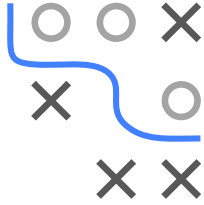
Introduction to Machine Learning

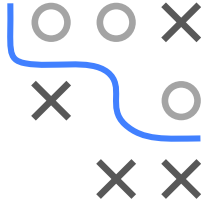


Multiclass Classification
Multiclass Classification and Losses

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

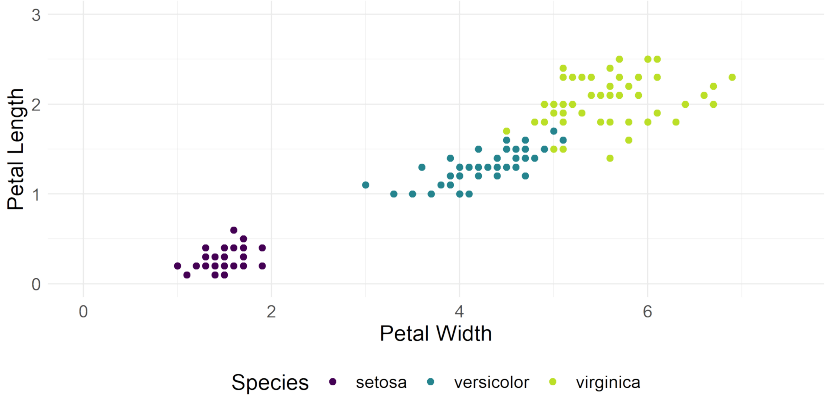
MULTICLASS CLASSIFICATION



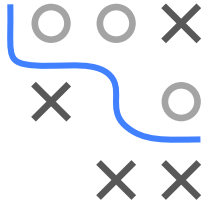
Scenario: Multiclass classification with $g > 2$ classes

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

Example: Iris dataset with $g = 3$



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 $(\mathbf{x}, y) \sim \mathbb{P}_{xy}$

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

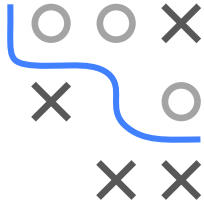
The optimal model for a loss function $L(y, f(\mathbf{x}))$ is

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

Because we usually do not know $\mathbb{P}_{xy}$, we minimize the **empirical risk** as an approximation to the **theoretical** risk

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

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

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

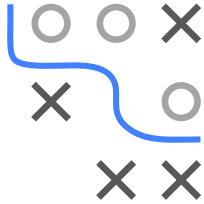
  - vectors of probabilities

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

  - hard labels

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

ONE-HOT ENCODING



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

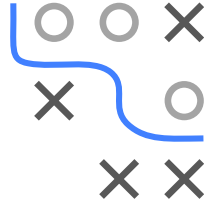
$$\text{with} \quad \mathbb{1}_{\{y=k\}} = \begin{cases} 1 & \text{if } y = k \\ 0 & \text{otherwise} \end{cases}$$

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

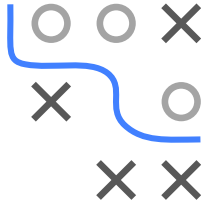
Example: Iris

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

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

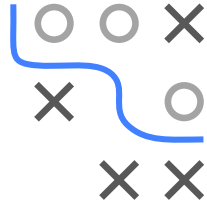
and the optimal constant model (featureless predictor)

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

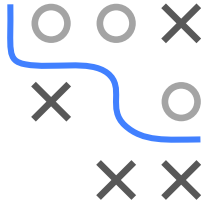
is the classifier that predicts the most frequent class $k \in \{1, 2, ..., g\}$ in the data

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

MC Brier Score



MC BRIER SCORE



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

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

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

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

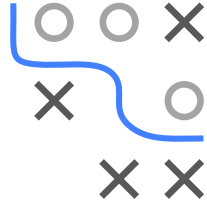
We solve this by setting the derivative w.r.t. θ_k to 0

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

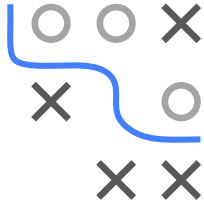
being the fraction of class-k observations.

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!

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**:

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

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

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

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

being the fraction of class-k observations.

Proof: Exercise.

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