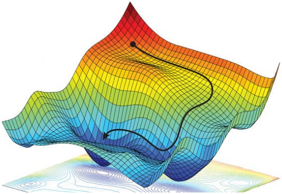
Introduction to Machine Learning

0-1-Loss



Learning goals

- Derive the risk minimizer of the 0-1-loss
- Derive the optimal constant model for the 0-1-loss

0-1-LOSS

- Let us first consider a classifier $h(\mathbf{x}) : \mathcal{X} \to \mathcal{Y}$ with $\mathcal{Y} = \{1, ..., g\}$ that outputs discrete classes directly.
- The most natural choice for $L(y, h(\mathbf{x}))$ is of course the 0-1-loss that counts the number of misclassifications

$$L(y, h(\mathbf{x})) = \mathbb{1}_{\{y \neq h(\mathbf{x})\}} = \begin{cases} 1 & \text{if } y \neq h(\mathbf{x}) \\ 0 & \text{if } y = h(\mathbf{x}) \end{cases} .$$

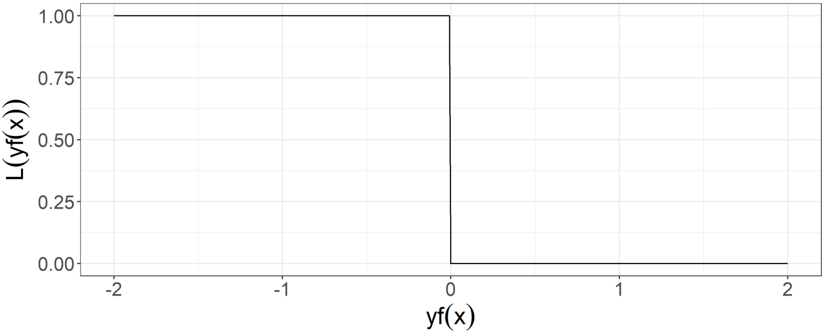
- For the binary case ($g = 2$) we can express the 0-1-loss for a scoring classifier $f(\mathbf{x})$ based on the margin $r := yf(\mathbf{x})$

$$L(r) = \mathbb{1}_{\{r<0\}} = \mathbb{1}_{\{yf(\mathbf{x})<0\}} = \mathbb{1}_{\{y \neq h(\mathbf{x})\}}.$$

0-1-LOSS

$$L(r) = \mathbb{1}_{\{r<0\}} = \mathbb{1}_{\{yf(\mathbf{x})<0\}} = \mathbb{1}_{\{y \neq h(\mathbf{x})\}}$$

- Intuitive, often what we are interested in.
- Analytic properties: Not continuous, even for linear f the optimization problem is NP-hard and close to intractable.



0-1-LOSS: RISK MINIMIZER

By the law of total expection we can in general rewrite the risk as

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

with $\mathbb{P}(y = k|\mathbf{x} = \mathbf{x})$ being the posterior probability for class k.

The risk minimizer for a general loss function $L\left(y, f(\mathbf{x})\right)$ is

$$\hat{f}(\mathbf{x}) = \underset{f:\mathcal{X} \to \mathbb{R}^g}{\arg\min}\, \mathbb{E}_x\left[\sum_{k \in \mathcal{Y}} L(k, f(\boldsymbol{x}))\mathbb{P}(y = k|\mathbf{x} = \mathbf{x})\right].$$

0-1-LOSS: RISK MINIMIZER

We compute the point-wise optimizer of the above term for the 0-1-loss (defined on a discrete classifier $h(\mathbf{x})$):

$$
\begin{aligned}
h^*(\mathbf{x}) &= \arg\min_{l \in \mathcal{Y}} \sum_{k \in \mathcal{Y}} L(k, l) \cdot \mathbb{P}(y = k \mid \mathbf{x} = \mathbf{x}) \\
&= \arg\min_{l \in \mathcal{Y}} \sum_{k \neq l} \mathbb{P}(y = k \mid \mathbf{x} = \mathbf{x}) \\
&= \arg\min_{l \in \mathcal{Y}} 1 - \mathbb{P}(y = l \mid \mathbf{x} = \mathbf{x}) \\
&= \arg\max_{l \in \mathcal{Y}} \mathbb{P}(y = l \mid \mathbf{x} = \mathbf{x}),
\end{aligned}
$$

which corresponds to predicting the most probable class.

Note that some literature refers to $h^*(\mathbf{x})$ as the **Bayes optimal classifier** (without specifying the loss function).

0-1-LOSS: RISK MINIMIZER

The Bayes risk for the 0-1-loss (also: Bayes error rate) is

$$
\begin{aligned}
\mathcal{R}^* &= \mathbb{E}_x \left[\min_{l \in \mathcal{Y}} \left(1 - \mathbb{P}(y = l \mid \mathbf{x} = \mathbf{x})\right) \right] \\
&= 1 - \mathbb{E}_x \left[\max_{l \in \mathcal{Y}} \mathbb{P}(y = l \mid \mathbf{x} = \mathbf{x}) \right].
\end{aligned}
$$

In the binary case ($g = 2$), we define $\eta(\mathbf{x}) := \mathbb{P}(y = 1 \mid \mathbf{x})$ and write risk minimizer and Bayes risk as follows:

$$
\begin{aligned}
h^*(\mathbf{x}) &= \begin{cases} 1 & \eta(\mathbf{x}) \geq \frac{1}{2} \\ 0 & \eta(\mathbf{x}) < \frac{1}{2} \end{cases} \\
\mathcal{R}^* &= \mathbb{E}_x \left[\min(\eta(\mathbf{x}), 1 - \eta(\mathbf{x})) \right] = \mathbb{E}_x \left[\max(\eta(\mathbf{x}), 1 - \eta(\mathbf{x})) \right].
\end{aligned}
$$

0-1-LOSS: RISK MINIMIZER

Example: Assume that $\mathbb{P}(y = 1) = \frac{1}{2}$ and $\mathbb{P}(\mathbf{x} \mid y) = \begin{cases} \phi_{\mu_1,\sigma^2}(\mathbf{x}) \\ \phi_{\mu_2,\sigma^2}(\mathbf{x}) \end{cases}$.

The decision boundary of the Bayes optimal classifier is shown in orange and the Bayes error rate is highlighted as red area.