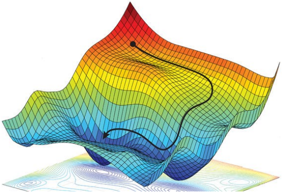
Introduction to Machine Learning

Theoretical Considerations on Classification Losses



Learning goals

- Know the 0-1-loss
- Derive the point-wise optimum for the 0-1-loss
- Understand the concept of the Bayes Optimal Classifier and the Bayes Error

RISK MINIMIZATION FOR CLASSIFICATION

Let y be categorical with g classes, i. e. $\mathcal{Y} = \{1, ..., g\}$ and let $f : \mathcal{X} \to \mathbb{R}^g$. We assume our model f outputs a g-dimensional vector of scores or probabilities, one per class.

Goal: Find a model f that minimizes the expected loss over random observations $(\mathbf{x}, y) \sim \mathbb{P}_{xy}$

$$\arg\min_{f \in \mathcal{H}} \mathcal{R}(f) = \mathbb{E}_{xy}[L\left(y, f(\mathbf{x})\right)] = \int L\left(y, f(\mathbf{x})\right)\, d\mathbb{P}_{xy}.$$

Note: If we exclusively consider **binary classification** tasks

- we (usually) encode labels as $y \in \{-1, 1\}$ for scoring classifiers $f(\mathbf{x})$, and as $y \in \{0, 1\}$ for probabilistic classifiers $\pi(\mathbf{x})$ unless explicitly stated differently.
- $f(\mathbf{x})$ and $\pi(\mathbf{x})$ are univariate scalars.

POINT-WISE OPTIMUM

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.

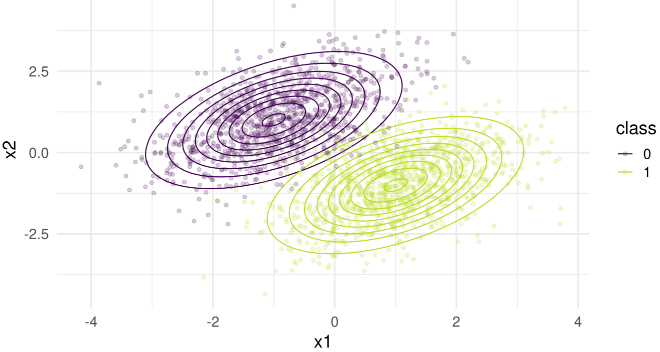
The optimal model for a loss function $L\left(y, f(\mathbf{x})\right)$ is

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

POINT-WISE OPTIMUM

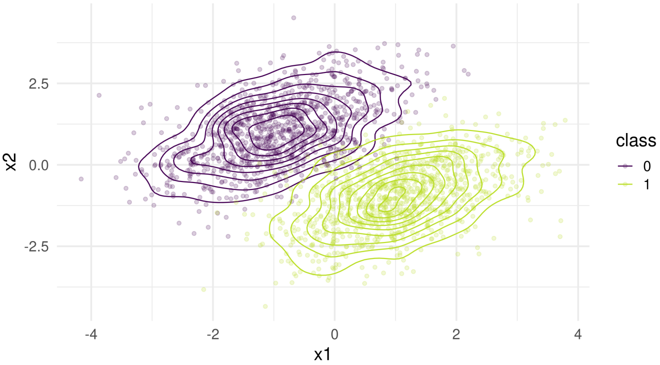
If we can estimate $\mathbb{P}_{xy}$ very well via $\pi_k(\mathbf{x})$ through a stochastic model, we can compute the loss-optimal classifications point-wise.

Example: Assume that our data is generated by a Mixture of Gaussian distributions.



POINT-WISE OPTIMUM

We could try to approximate the $\mathbb{P}(y = k \mid \mathbf{x} = \mathbf{x})$ via a stochastic model $\pi(\mathbf{x})$ (shown as contour lines):



POINT-WISE OPTIMUM

For each new $\mathbf{x}$, we estimate the class probabilities directly with the stochastic model $\pi(\mathbf{x})$, and our best point-wise prediction is

$$\hat{f}(\mathbf{x}) \quad = \quad \arg\min_{f \in \mathcal{H}} \sum_{k \in \mathcal{Y}} L(k, f(\boldsymbol{x}))\pi(\mathbf{x})\,.$$

But usually we directly adapt to the loss via **empirical risk minimization**.

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

0-1-Loss and Bayes Optimal Predictor

0-1-LOSS

- Let us first consider a classifier $h(\mathbf{x})$ 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

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

0-1-LOSS: POINT-WISE OPTIMUM

For an (unrestricted) classifier $h(\mathbf{x})$ and the 0-1-loss:

$$\min_{h \in \mathcal{H}} \mathcal{R}(h) = \mathbb{E}_{xy}[L(y, h(\mathbf{x}))].$$

The (point-wise) solution of the above minimization problem is

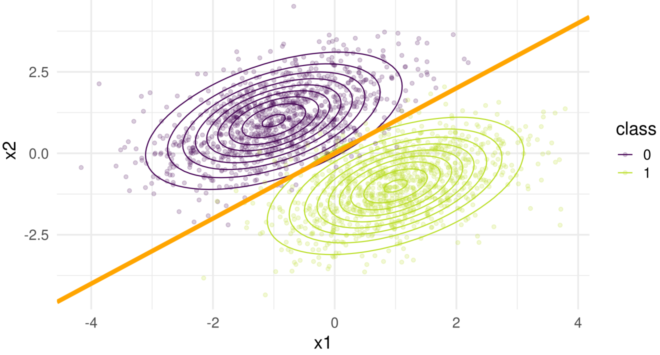
$$\begin{aligned}
\hat{h}(\mathbf{x}) &= \arg\min_{l \in \mathcal{Y}} \sum_{k \in \mathcal{Y}} L(k, l) \cdot \mathbb{P}(y = k \mid \mathbf{x} = \mathbf{x}) \\
&= \arg\min_{l \in \mathcal{Y}} \sum_{k \neq l} \mathbb{P}(y = k \mid \mathbf{x} = \mathbf{x}) = \arg\min_{l \in \mathcal{Y}} 1 - \mathbb{P}(y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

which corresponds to predicting the most probable class.

BAYES OPTIMAL CLASSIFIER

$\hat{h}(\mathbf{x})$ is called the **Bayes optimal classifier** for the 0-1-loss.

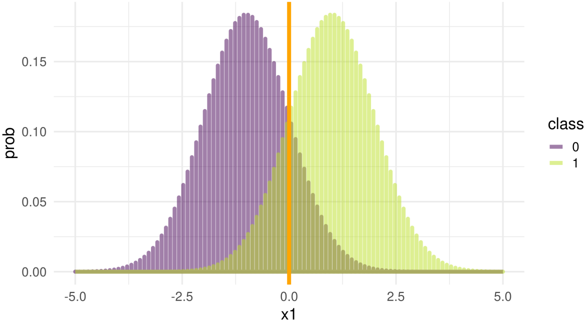
Example: Consider again the previous example, and assume the data generating distribution is known. The decision boundary of the Bayes optimal classifier is shown in orange:



BAYES ERROR RATE

There is an unavoidable error: Even if we know the underlying distribution perfectly, it is possible that a class 1 observations is more likely under $\mathbb{P}_{xy}(\mathbf{x} \mid y = 0)$ than under $\mathbb{P}_{xy}(\mathbf{x} \mid y = 1)$.

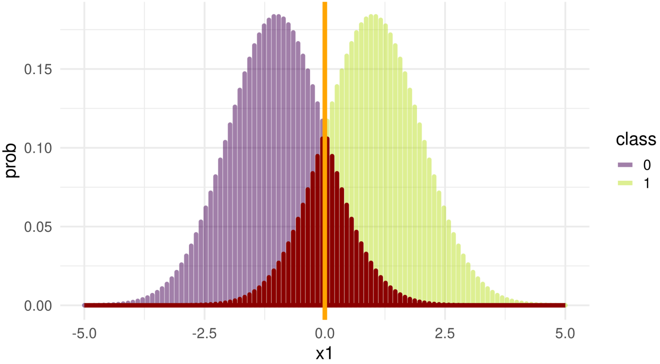
Example: Consider a one-dimensional variant of the Gaussian mixture model. The Bayes optimal classifier is shown in orange.



BAYES ERROR RATE

The expected loss is called **Bayes loss** or **Bayes error rate** for the 0-1-loss.

Example: The Bayes error rate is highlighted as red area.



0-1-LOSS: OPTIMAL CONSTANT MODEL

The optimal constant model (featureless predictor) under 0-1 loss, with $y \in \{-1, +1\}$, either for hard classifiers $h(\mathbf{x})$ or scoring classifiers $f(\mathbf{x})$

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

is the classifier that predicts the most frequent class in the data

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

Proof: Exercise / Trivial.

While the **Bayes error rate** is the theoretically lowest error rate we can achieve for a given data generating process, the above classifier gives usually a lower baseline for the predictive performance of a model.