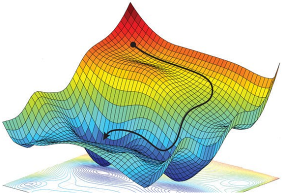
Introduction to Machine Learning

Risk Minimizers



Learning goals

- Know the concepts of the Bayes optimal model (also: risk minimizer, population minimizer)
- Know the Bayes risk
- Know the concept of consistent learners
- Know the concept of the optimal constant model

RISK MINIMIZER

Our goal is to minimize the risk

$$\mathcal{R}_L(f) := \mathbb{E}_{xy}[L(y, f(\mathbf{x}))] = \int L(y, f(\mathbf{x}))\, d\mathbb{P}_{xy}.$$

for a certain hypothesis $f(\mathbf{x}) \in \mathcal{H}$ and a loss $L(y, f(\mathbf{x}))$.

Let us assume we are in an "ideal world":

- The hypothesis space $\mathcal{H}$ is unrestricted. We can choose any $f : \mathcal{X} \to \mathbb{R}^g$.
- We do not care about the optimizer; let us assume every solution in the hypothesis space can be reached efficiently.
- We know $\mathbb{P}_{xy}$.

How should f be chosen?

RISK MINIMIZER

We call the function $f : \mathcal{X} \to \mathbb{R}^g$ that minimizes the risk

$$\begin{aligned} f^* &= \underset{f:\mathcal{X}\to\mathbb{R}^g}{\arg\min}\, \mathcal{R}_L(f) = \underset{f:\mathcal{X}\to\mathbb{R}^g}{\arg\min}\, \mathbb{E}_{xy}\left[L\left(y, f(\mathbf{x})\right)\right] \\ &= \underset{f:\mathcal{X}\to\mathbb{R}^g}{\arg\min} \int L\left(y, f(\mathbf{x})\right) d\mathbb{P}_{xy}. \end{aligned}$$

the **risk minimizer**, **population minimizer**, or **Bayes optimal model**. Note that we search over an unrestricted hypothesis space (that is over all possible functions $f : \mathcal{X} \to \mathbb{R}^g$)!

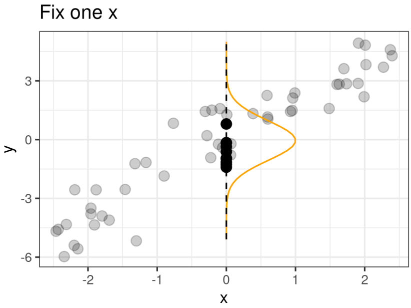
The resulting risk is called **Bayes risk**

$$\mathcal{R}_L^* = \inf_{f:\mathcal{X}\to\mathbb{R}^g} \mathcal{R}_L(f).$$

OPTIMAL POINT-WISE PREDICTIONS

To derive the risk minimizer we usually make use of the following trick:

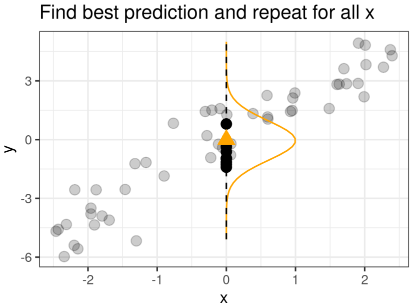
- We can choose $f(\mathbf{x})$ as we want (unrestricted hypothesis space)
- Consequently, for a fixed value $\mathbf{x} \in \mathcal{X}$ we can select **any** value c we want to predict (we are not restricted by any functional form, e.g., a linear function)
- Instead of looking for the optimal f in function space (which is impossible), we compute the **point-wise optimizer** for every $\mathbf{x} \in \mathcal{X}$.



OPTIMAL POINT-WISE PREDICTIONS

To derive the risk minimizer we usually make use of the following trick:

- We can choose $f(\mathbf{x})$ as we want (unrestricted hypothesis space)
- Consequently, for a fixed value $\mathbf{x} \in \mathcal{X}$ we can select **any** value c we want to predict (we are not restricted by any functional form, e.g., a linear function)
- Instead of looking for the optimal f in function space (which is impossible), we compute the **point-wise optimizer** for every $\mathbf{x} \in \mathcal{X}$.



THEORETICAL AND EMPIRICAL RISK

The risk minimizer in general only allows for theoretical considerations:

- In practice we need to restrict the hypothesis space $\mathcal{H}$ such that we can efficiently search over it.
- In practice we (usually) do not know $\mathbb{P}_{xy}$. Instead of $\mathcal{R}(f)$, we are optimizing the empirical risk and

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

Note that according to the **law of large numbers** (LLN), the empirical risk converges to the true risk:

$$\bar{\mathcal{R}}_{\text{emp}}(f) = \frac{1}{n} \sum_{i=1}^{n} L\left(y^{(i)}, f\left(\mathbf{x}^{(i)}\right)\right) \overset{n \to \infty}{\longrightarrow} \mathcal{R}(f).$$

ESTIMATION AND APPROXIMATION ERROR

Goal of learning: Train a model $\hat{f}$ for which the true risk $\mathcal{R}_L\left(\hat{f}\right)$ is close to the Bayes risk $\mathcal{R}_L^*$. In other words, we want the **Bayes regret**

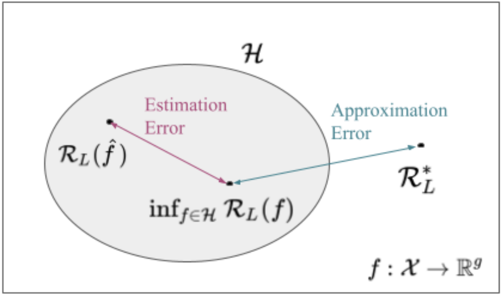
$$\mathcal{R}_L\left(\hat{f}\right) - \mathcal{R}_L^*$$

to be as low as possible.

The Bayes regret can be decomposed as follows:

$$\mathcal{R}_L\left(\hat{f}\right) - \mathcal{R}_L^* = \underbrace{\left[\mathcal{R}_L\left(\hat{f}\right) - \inf_{f \in \mathcal{H}} \mathcal{R}_L(f)\right]}_{\text{estimation error}} + \underbrace{\left[\inf_{f \in \mathcal{H}} \mathcal{R}_L(f) - \mathcal{R}_L^*\right]}_{\text{approximation error}}$$

ESTIMATION AND APPROXIMATION ERROR



- $\mathcal{R}_L\left(\hat{f}\right) - \inf_{f \in \mathcal{H}} \mathcal{R}(f)$ is the **estimation error**. We fit $\hat{f}$ via empirical risk minimization and (usually) use approximate optimization, so we usually do not find the optimal $f \in \mathcal{H}$.
- $\inf_{f \in \mathcal{H}} \mathcal{R}_L(f) - \mathcal{R}_L^*$ is the **approximation error**. We need to restrict to a hypothesis space $\mathcal{H}$ which might not even contain the Bayes optimal model f^*.

A NOTE ON NOTATION

The risk function $\mathcal{R}_L$ depends on the choice of the loss function $L(y, f(\mathbf{x}))$. To make this clear we denote this as subscript:

$$\mathcal{R}_L$$

To keep notation simple, we will omit the subscript most of the time.

Keep in mind that the risk $\mathcal{R}$ always depends on the specific choice of the loss function $L(y, f(\mathbf{x}))$!

(UNIVERSALLY) CONSISTENT LEARNERS

Consistency is an asymptotic property of a learning algorithm, which ensures the algorithm returns **better models** when given **more data**.

Let $\mathcal{I} : \mathbb{D} \times \Lambda \to \mathcal{H}$ be a learning algorithm$^{(*)}$ that takes a training set $\mathcal{D}_{\text{train}} \sim \mathbb{P}_{xy}$ of size n_{train} and estimates a model $\hat{f} : \mathcal{X} \to \mathbb{R}^g$.

The learning method $\mathcal{I}$ is said to be **consistent** w.r.t. a certain distribution $\mathbb{P}_{xy}$ if the risk of the estimated model $\hat{f}$ converges in probability ("$\xrightarrow{p}$") to the Bayes risk $\mathcal{R}^*$ when n_{train} goes to ∞:

$$\mathcal{R}\left(\mathcal{I}\left(\mathcal{D}_{\text{train}}, \boldsymbol{\lambda}\right)\right) \xrightarrow{p} \mathcal{R}_L^* \quad \text{for } n_{\text{train}} \to \infty.$$

$^{(*)}$ $\boldsymbol{\lambda} \in \Lambda$ denote hyperparameters of the learning algorithm.

(UNIVERSALLY) CONSISTENT LEARNERS

Consistency of an algorithm is a statement that is valid for only one particular distribution $\mathbb{P}_{xy}$.

But since we usually do not have any prior knowledge on $\mathbb{P}_{xy}$, consistency of algorithms does not offer much help to choose an algorithm for a particular task.

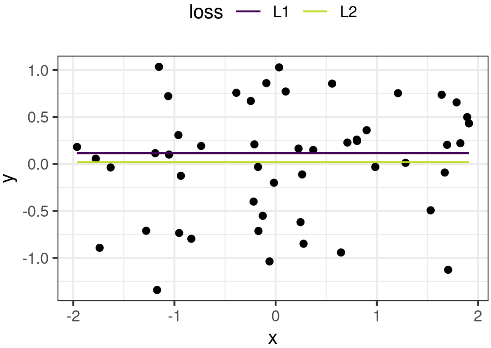
Thus we define the stronger concept of **universal consistency**: An algorithm is said to be universally consistent if it is consistent for **any** distribution $\mathbb{P}_{xy}$.

Note that universal consistency is obviously a desirable property - however, (universal) consistency does not tell us anything about convergence rates ...

OPTIMAL CONSTANT MODEL

While the risk minimizer gives us the (theoretical) optimal solution, the **optimal constant model** (also: featureless predictor) gives us an computable empirical lower baseline solution.

The constant model is the model $f(\mathbf{x}) = \boldsymbol{\theta}$ that optimizes the empirical risk $\mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$.



RISK MINIMIZER AND OPTIMAL CONSTANT

In the following chapters, we will derive the risk minimizers and optimal constant models for different loss functions.

Name	Risk Minimizer	Optimal Constant
L2	$f^*(\mathbf{x}) = \mathbb{E}_{y\|x}\left[y \mid \mathbf{x}\right]$	$\hat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} y^{(i)}$
L1	$f^*(\mathbf{x}) = \text{med}_{y\|x}\left[y \mid \mathbf{x}\right]$	$\hat{f}(\mathbf{x}) = \text{med}(y^{(i)})$
0-1	$h^*(\mathbf{x}) = \arg\max_{l \in \mathcal{Y}} \mathbb{P}(y = l \mid \mathbf{x})$	$\hat{h}(\mathbf{x}) = \text{mode}\left\{y^{(i)}\right\}$
Brier	$\pi^*(\mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x} = \mathbf{x})$	$\hat{\pi}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} y^{(i)}$
Bernoulli (on probs)	$\pi^*(\mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x} = \mathbf{x})$	$\hat{\pi}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} y^{(i)}$
Bernoulli (on scores)	$f^*(\mathbf{x}) = \log\left(\frac{\mathbb{P}(y=1 \mid \mathbf{x})}{1-\mathbb{P}(y=1 \mid \mathbf{x})}\right)$	$\hat{f}(\mathbf{x}) = \log\frac{n_{+1}}{n_{-1}}$