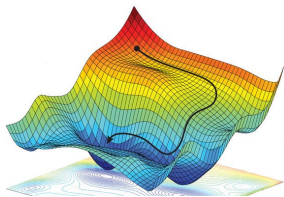


Introduction to Machine Learning

Risk Minimizers



Learning goals

- Know the concepts of the Bayes optimal model (also: risk minimizer, population minimizer)
- Bayes risk
- Consistent learners
- Bayes regret, estimation and approximation error
- Optimal constant model
- Proper scoring rules

RISK MINIMIZER

Our goal is to minimize the risk

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

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

NB: As \mathcal{R}_L depends on loss L , we sometimes make this explicit with a subscript if needed, and omit in other cases.

Let us assume we are in an “ideal world”:

- The hypothesis space \mathcal{H} is unrestricted. We can choose any $f : \mathcal{X} \rightarrow \mathbb{R}^g$.
- We also assume an ideal optimizer; the risk minimization can always be solved perfectly and efficiently.
- We know \mathbb{P}_{xy} .

How should f be chosen?

RISK MINIMIZER

The f with minimal risk across all (measurable) functions is called the **risk minimizer**, **population minimizer** or **Bayes optimal model**.

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

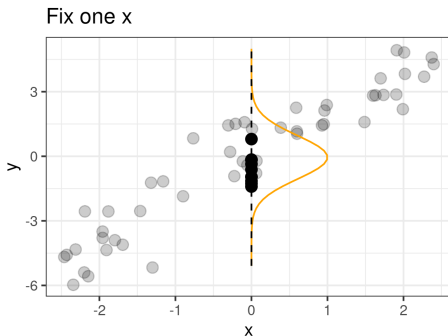
The resulting risk is called **Bayes risk**

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

OPTIMAL POINT-WISE PREDICTIONS

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

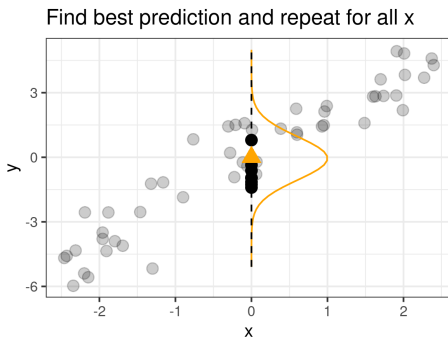
- We can choose $f(\mathbf{x})$ as we want (unrestricted hypothesis space, no assumed functional form)
- Consequently, for a fixed value $\mathbf{x} \in \mathcal{X}$ we can select **any** value c we want to predict
- So we construct the **point-wise optimizer** for every $\mathbf{x} \in \mathcal{X}$.



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, no assumed functional form)
- Consequently, for a fixed value $\mathbf{x} \in \mathcal{X}$ we can select **any** value c we want to predict
- So we construct the **point-wise optimizer** for every $\mathbf{x} \in \mathcal{X}$.



THEORETICAL AND EMPIRICAL RISK

The risk minimizer is mainly a theoretical tool:

- 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

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

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

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

ESTIMATION AND APPROXIMATION ERROR

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

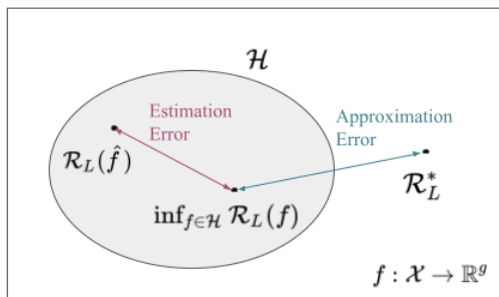
$$\mathcal{R}_L(\hat{f}) - \mathcal{R}_L^*$$

to be as low as possible.

The Bayes regret can be decomposed as follows:

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

ESTIMATION AND APPROXIMATION ERROR



- $\mathcal{R}_L(\hat{f}) - \inf_{f \in \mathcal{H}} \mathcal{R}(f)$ 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^* .

(UNIVERSALLY) CONSISTENT LEARNERS

Consistency is an asymptotic property of a learning algorithm, which ensures the algorithm returns **the correct model** when given **unlimited data**.

Let $\mathcal{I} : \mathbb{D} \times \Lambda \rightarrow \mathcal{H}$ 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 $\hat{f} : \mathcal{X} \rightarrow \mathbb{R}^g$.

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 ∞ :

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

^(*) $\lambda \in \Lambda$ denotes hyperparameters of the learning algorithm.

(UNIVERSALLY) CONSISTENT LEARNERS

Consistency is defined w.r.t. a particular distribution \mathbb{P}_{xy} . But since we usually do not know \mathbb{P}_{xy} , consistency does not offer much help to choose an algorithm for a particular task.

More interesting is the stronger concept of **universal consistency**: An algorithm is universally consistent if it is consistent for **any** distribution.

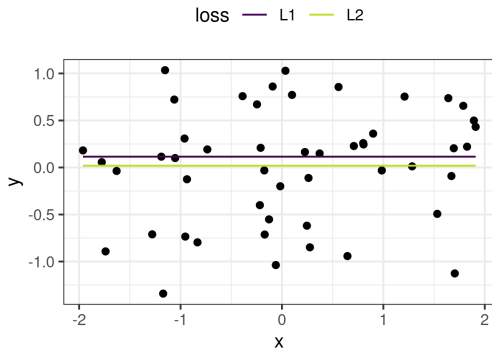
In Stone's famous consistency theorem from 1977, the universal consistency of a weighted average estimator as KNN was proven. Many other ML models have since then been proven to be universally consistent (SVMs, ANNs, etc.).

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



RISK MINIMIZER AND OPTIMAL CONSTANT

Later, we will derive risk minimizers for various losses.

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

We see: For regression, the RMs model the conditional expectation and median of the underlying distribution. This makes intuitive sense, depending on your concept of how to best estimate central location / how robust this location should be.

RISK MINIMIZER AND OPTIMAL CONSTANT

Later, we will derive risk minimizers for various losses.

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

For the 0-1 loss, we the risk minmizer construct the **optimal Bayes decision rule**: We predict the class with maximal posterior probability.

RISK MINIMIZER AND OPTIMAL CONSTANT

Later, we will derive risk minimizers for various losses.

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

For Brier and Bernoulli, we predict the posterior probabilities (of the true DGP!). Losses that have this desirable property are called **proper scoring (rules)**.