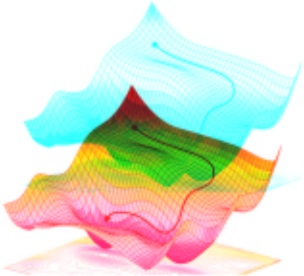
Introduction to Machine Learning

Advanced Risk Minimization

Risk Minimizers



Learning goals

- Bayes optimal model (also: risk minimizer, population minimizer)
- Bayes risk
- Bayes regret, estimation and approximation error
- Optimal constant model
- Consistent learners

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

EMPIRICAL RISK MINIMIZATION

Very often, in ML, we minimize the empirical risk

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

- each observation $\left(\mathbf{x}^{(i)}, y^{(i)}\right) \in \mathcal{X} \times \mathcal{Y}$, so from feature and target space
- $f_{\mathcal{H}} : \mathcal{X} \to \mathbb{R}^g$, f is a model from hypothesis space $\mathcal{H}$; maps a feature vector to output score; sometimes or often we omit $\mathcal{H}$ in the index
- $L : (\mathcal{Y} \times \mathbb{R}^g) \to \mathbb{R}$ is loss; $L(y, f)$ measures distance between label and prediction
- We assume that $(\mathbf{x}, y) \sim P_{xy}$ and $\left(\mathbf{x}^{(i)}, y^{(i)}\right) \overset{\text{i.i.d.}}{\sim} P_{xy}$ P_{xy} is the distribution of the data generating process (DGP)

Let's define (and minimize) loss in expectation, the theoretical risk:

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

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

where

- each observation $\left(\mathbf{x}^{(i)}, y^{(i)}\right) \in \mathcal{X} \times \mathcal{Y}$ and $\mathcal{X}, \mathcal{Y}$ are the feature and target spaces, respectively,
- $f : \mathcal{X} \to \mathbb{R}^g, f \in \mathcal{H}$ is a model which maps a feature vector to a numerically encoded element of $\mathcal{Y}$ and $\mathcal{H}$ is the hypothesis space,
- $L : (\mathcal{Y} \times \mathbb{R}^g) \to \mathbb{R}$ is the loss function which measures the dissimilarity of the model prediction and the true target,
- and we assume that $\left(\mathbf{x}^{(i)}, y^{(i)}\right) \overset{\text{i.i.d.}}{\sim} \mathbb{P}_{xy}$ where $\mathbb{P}_{xy}$ is the distribution of the data generating process (DGP).

What is the theoretical justification for this procedure?

TWO SHORT EXAMPLES

Regression with linear model:

- Model: $f(\mathbf{x}) = \boldsymbol{\theta}^\top \mathbf{x} + \theta_0$
- Squared loss: $L(y, f) = (y - f)^2$
- Hypothesis space:

$$\mathcal{H}_{\text{lin}} = \left\{ \mathbf{x} \mapsto \boldsymbol{\theta}^\top \mathbf{x} + \theta_0 : \boldsymbol{\theta} \in \mathbb{R}^d, \theta_0 \in \mathbb{R} \right\}$$

Binary classification with shallow MLP:

- Model: $f(\mathbf{x}) = \pi(\mathbf{x}) = \sigma(\boldsymbol{w}_2^\top \text{ReLU}(\boldsymbol{W}_1 \mathbf{x} + \boldsymbol{b}_1) + b_2)$
- Binary cross-entropy loss: $L(y, \pi) = -(y \log(\pi) + (1 - y) \log(1 - \pi))$
- Hypothesis space:

$$\mathcal{H}_{\text{MLP}} = \{ \mathbf{x} \mapsto \sigma(\boldsymbol{w}_2^\top \text{ReLU}(\boldsymbol{W}_1 \mathbf{x} + \boldsymbol{b}_1) + b_2) : \mathbf{W}_1 \in \mathbb{R}^{h \times d}, \mathbf{b}_1 \in \mathbb{R}^h, \mathbf{w}_2 \in \mathbb{R}^h, b_2 \in \mathbb{R} \}$$

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

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

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

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

- The hypothesis space $\mathcal{H}$ is unrestricted. We can choose any
- We also assume an ideal optimizer; the risk minimization can always be solved perfectly and efficiently.

How should f be chosen?

OPTIMAL CONSTANTS FOR A LOSS

- Let's assume some RV $z \in \mathcal{Y}$ for a label
- z not RV y, because we want to fiddle with its distribution
- Assume z has distribution Q, so $z \sim Q$
- We can now consider $\arg\min_c \mathbb{E}_{z\sim Q}[L(z, c)]$ so the score-constant which loss-minimally approximates z

We will consider 3 cases for Q

- $Q = P_y$, simply our labels and their marginal distribution in P_{xy}
- $Q = P_{y|x=x}$, conditional label distribution at point $x = \tilde{x}$
- $Q = P_n$, the empirical product distribution for data $y_1, \ldots, y_n$

If we can solve $\arg\min_c \mathbb{E}_{z\sim Q}[L(z, c)]$ for any Q, we will get multiple useful results!

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

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

The resulting risk is called **Bayes risk**

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

OPTIMAL CONSTANT MODEL

- We would like a loss optimal, constant baseline predictor
  - A "featureless" ML model, which always predicts the same value
  - Can use it as baseline in experiments, if we don't beat this with more complex model, that model is useless
  - Will also be useful as component in algorithms and derivations

$$\hat{f}_c = \arg\min_{c\in\mathbb{R}} \mathbb{E}_{xy}\left[L(y,c)\right] = \arg\min_{c\in\mathbb{R}} \mathbb{E}_{y}\left[L(y,c)\right]$$

and $f(\mathbf{x}) = \theta = c$ that optimizes the empirical risk $\mathcal{R}_{\text{emp}}(\theta)$ is denoted as as $\hat{f}_c = \arg\min_{c\in\mathbb{R}} \mathcal{R}_{\text{emp}}(\theta)$.

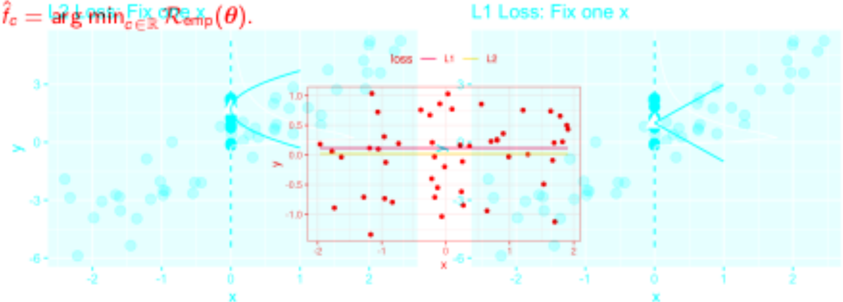
POINT-WISE PREDICTIONS

To derive the risk minimizer, observe that by the law of total expectation

$$\mathcal{R}_L(f) = \mathbb{E}_{xy}\left[L(y, f(\mathbf{x}))\right] = \mathbb{E}_x\left[\mathbb{E}_{y|x}\left[L(y, f(\mathbf{x})) \mid \mathbf{x} = \mathbf{x}\right]\right].$$

- We can choose $f(\mathbf{x})$ as we want (unrestricted hypothesis space, no assumed functional form)
- Hence, 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**

$$f^*(\mathbf{x}) = \text{argmin}_c \mathbb{E}_{y|x}\left[L(y,c) \mid \mathbf{x} = \mathbf{x}\right] \quad \forall \mathbf{x} \in \mathcal{X}.$$



OPTIMAL CONSTANT MODEL

- Let's start with the simplest case, L2 loss
- And we want to find and optimal constant model for

$$\arg\min \mathbb{E}[L(z,c)] =$$
$$\arg\min \mathbb{E}[(z-c)^2] =$$
$$\arg\min \mathbb{E}[z^2] - 2cE[z] + c^2 =$$
$$E[z]$$

- Using $Q = P_y$, this means that, given we know the label distribution, the best constant is $c = E[y]$.
- If we only have data $y_1, \dots y_n$

$$\arg\min \mathbb{E}_{z\sim P_n}[(z-c)^2] = \mathbb{E}_{z\sim P_n}[z] = \frac{1}{n}\sum_{i=1}^{n} y^{(i)} = \bar{y}$$

- And we want to find and optimal constant model for

THEORETICAL VS. 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!):

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

RISK MINIMIZER

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

- The hypothesis space $\mathcal{H} = \mathcal{H}_{all}$ is unrestricted. We can choose any measurable $f : \mathcal{X} \to \mathbb{R}^g$.
- 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?

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:

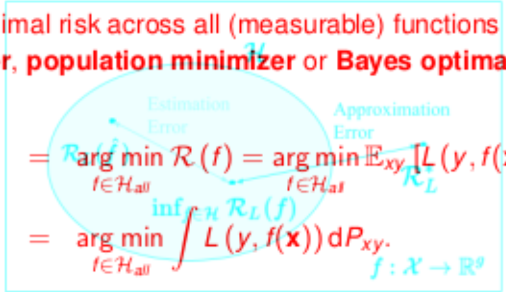
$$\mathcal{R}_L\left(\hat{f}\right) - \mathcal{R}_L^* = \underbrace{\left[\mathcal{R}_L\left(\hat{f}\right) - \inf_{f \in \mathcal{H}} \mathcal{R}_L(f)\right]}_{\text{estimation error}} + \underbrace{\left[\inf_{f \in \mathcal{H}} \mathcal{R}_L(f) - \mathcal{R}_L^*\right]}_{\text{approximation error}}$$

RISK MINIMIZER / 2

ESTIMATION AND APPROXIMATION ERROR / 2

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

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



The resulting risk is called **Bayes risk**: $\mathcal{R}^* = \mathcal{R}(f^*_{\mathcal{H}_{all}})$

Note that if we leave out the hypothesis space in the subscript it becomes clear from the context!

Similarly, we define the risk minimizer over some $\mathcal{H} \subset \mathcal{H}_{all}$ as

$$f^*_{\mathcal{H}} = \arg\min_{f \in \mathcal{H}} \mathcal{R}(f)$$

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

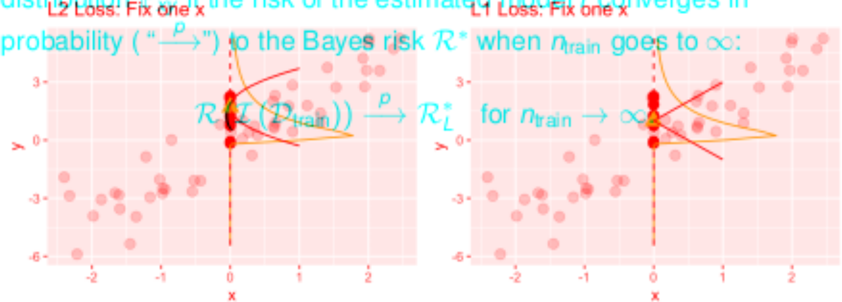
OPTIMAL POINT-WISE PREDICTIONS

To derive the risk minimizer, observe that by law of total expectation

$$\mathcal{R}(f) = \mathbb{E}_{xy}\left[L(y, f(\mathbf{x}))\right] = \mathbb{E}_x\left[\mathbb{E}_{y|x}\left[L(y, f(\mathbf{x})) \mid \mathbf{x}\right]\right].$$

- We can choose $f(\mathbf{x})$ as we want (unrestricted hypothesis space, no assumed functional form)
- Hence, 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**

$$f^*(\tilde{\mathbf{x}}) = \text{argmin}_c \, \mathbb{E}_{y|x}\left[L(y, c) \mid \mathbf{x} = \tilde{\mathbf{x}}\right]$$



(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 $\mathcal{I} : \mathbb{D} \to \mathcal{H}$ 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 ∞:

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

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 P_{xy}. Instead of $\mathcal{R}(f)$, we are optimizing the empirical risk

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

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

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

(UNIVERSALLY) CONSISTENT LEARNERS / 2

Consistency is defined w.r.t. a particular distribution $\mathbb{P}_{xy}$. But since we usually don't 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 ...

ESTIMATION AND APPROXIMATION ERROR

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

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

to be as low as possible.

The Bayes regret can be decomposed as follows:

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

$$= \left[\mathcal{R}(\hat{f}_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}})\right] + \left[\mathcal{R}(f^*_{\mathcal{H}}) - \mathcal{R}(f^*_{\mathcal{H}_{all}})\right]$$

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

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