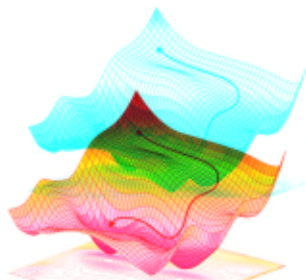


Introduction to Machine Learning

Advanced Risk Minimization

Risk Minimizers



Learning goals

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

EMPIRICAL RISK MINIMIZATION

Very often, in ML, we minimize the empirical risk

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



where each observation $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{X} \times \mathcal{Y}$, so from feature and target space

- $f: \mathcal{X} \rightarrow \mathbb{R}^g$, 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

- $L: (\mathcal{Y} \times \mathbb{R}^g) \rightarrow \mathbb{R}$ is loss;

- $L(y, f)$ measures distance between label and prediction numerically encoded element of \mathcal{Y} and \mathcal{H} is the hypothesis space,

- We assume that $(\mathbf{x}, y) \sim \mathbb{P}_{xy}$ and $(\mathbf{x}^{(i)}, y^{(i)}) \stackrel{i.i.d.}{\sim} \mathbb{P}_{xy}$

- \mathbb{P}_{xy} is the distribution of the data generating process (DGP)

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

- and we assume that $(\mathbf{x}^{(i)}, y^{(i)}) \stackrel{i.i.d.}{\sim} \mathbb{P}_{xy}$ where \mathbb{P}_{xy} is the distribution of the data generating process (DGP).

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

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: $L(y, f) = (y - f)^2 = \int L(y, f(\mathbf{x})) d\mathbb{P}_{xy}$
- Hypothesis space:

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.

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

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

Binary classification with shallow MLP:

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

How should f be chosen?

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



OPTIMAL CONSTANTS FOR A LOSS

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

- 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 $\arg \min_c \mathbb{E}_{z \sim Q} [L(z, c)] = \arg \min_c \mathbb{E}_{xy} [L(y, f(x))]$
so the score-constant which loss-minimally approximates z
 $= \arg \min_{f: \mathcal{X} \rightarrow \mathbb{R}^g} \int L(y, f(x)) d\mathbb{P}_{xy}$.

We will consider 3 cases for Q

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

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

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



OPTIMAL CONSTANT MODELCTIONS

Today we would like a loss optimal, constant baseline predictor of total expectation

- 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

Hence, for a fixed value $x \in \mathcal{X}$ we can select any value c we want to predict. So we restrict the point-wise optimizer

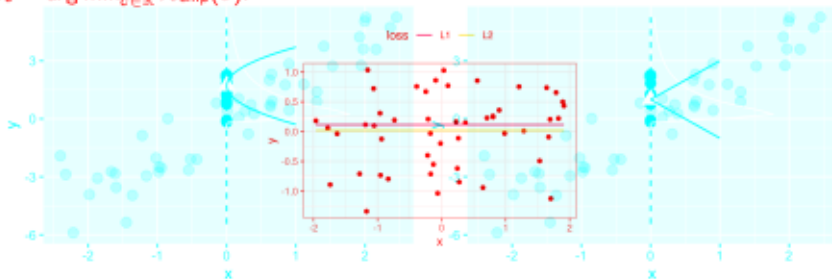
$$f_c = \arg \min_{c \in \mathbb{R}} E_{y|x} [L(y, c)] = \arg \min_{c \in \mathbb{R}} E_y [L(y, c) | x = x]$$

$$f^*(x) = \arg \min_{c \in \mathbb{R}} E_{y|x} [L(y, c) | x = x] \quad \forall x \in \mathcal{X}$$

and $f(x) = \theta = c$ that optimizes the empirical risk $\mathcal{R}_{\text{emp}}(\theta)$ is denoted as as

$$\hat{f}_c = \arg \min_{c \in \mathbb{R}} \mathcal{R}_{\text{emp}}(\theta)$$

L1 Loss: Fix one x



OPTIMAL CONSTANT MODEL RISK

The risk minimizer is mainly a theoretical tool:

- Let's start with the simplest case, L2 loss
- 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}_y . Instead of $\mathcal{R}(f)$, we are optimizing the empirical risk

$$\arg \min_c \mathbb{E}[(z - c)^2] =$$

$$\arg \min_c \mathbb{E}[z^2] - 2c\mathbb{E}[z] + c^2 =$$

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

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

- If we only have data y_1, \dots, y_n

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

- And we want to find an optimal constant model for



RISK MINIMIZER AND APPROXIMATION ERROR

Goal of learning: Train a model \hat{f} for which the true risk $\mathcal{R}_L(\hat{f})$ is

Let us assume we are in an "ideal world":
close to the Bayes risk \mathcal{R}_L^* . In other words, we want the **Bayes regret**

- The hypothesis space $\mathcal{H} = \mathcal{H}_{all}$ is unrestricted. We can choose any measurable $f: \mathcal{X} \rightarrow \mathbb{R}^q$.
 - We also assume an ideal optimizer; the risk minimization can always be solved perfectly and efficiently.
 - We know \mathbb{P}_L .
- The Bayes regret can be decomposed as follows:

How should f be chosen?

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



RISK MINIMIZER / 2 APPROXIMATION ERROR

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

$$\begin{aligned}
 f_{\mathcal{H}_{all}}^* &= \arg \min_{f \in \mathcal{H}_{all}} \mathcal{R}(f) = \arg \min_{f \in \mathcal{H}_{all}} \mathbb{E}_{xy} [L(y, f(\mathbf{x}))] \\
 &= \arg \min_{f \in \mathcal{H}_{all}} \int L(y, f(\mathbf{x})) d\mathbb{P}_{xy}.
 \end{aligned}$$

Diagram illustrating the decomposition of the risk minimization problem:

- The overall risk minimization is shown as $\arg \min_{f \in \mathcal{H}_{all}} \mathcal{R}(f)$.
- This is equivalent to $\arg \min_{f \in \mathcal{H}_{all}} \mathbb{E}_{xy} [L(y, f(\mathbf{x}))]$.
- The diagram shows a light blue circle containing the expression $\inf_{f \in \mathcal{H}} \mathcal{R}_L(f)$.
- Arrows indicate the relationship between the terms:
 - Estimation Error** is the difference between $\mathcal{R}(f)$ and $\mathcal{R}_L(f)$.
 - Approximation Error** is the difference between $\mathcal{R}(f)$ and \mathcal{R}^* .
- The function space is defined as $f: \mathcal{X} \rightarrow \mathbb{R}^q$.



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

- $\mathcal{R}_L(\hat{f}) - \inf_{f \in \mathcal{H}} \mathcal{R}(f)$ is the **estimation error**. We fit \hat{f} via

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

empirical risk minimization and (usually) use approximate optimization, so we usually do not find the optimal $f \in \mathcal{H}$.

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

- $\inf_{f \in \mathcal{H}} \mathcal{R}_L(f) - \mathcal{R}_L^*$ is the **approximation error**. We need to restrict to a hypothesis space \mathcal{H} which might not even contain the Bayes optimal model.

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

OPTIMAL POINT-WISE PREDICTIONS

To derive the risk minimizer, observe that by law of total expectation, consistency is an asymptotic property of a learning algorithm, which ensures that the algorithm returns the correct model when given unlimited data.

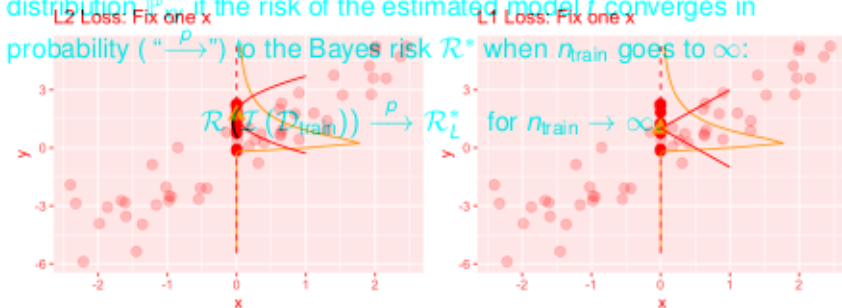
$$\mathcal{R}(f) = \mathbb{E}_{xy} [L(y, f(x))] = \mathbb{E}_x [\mathbb{E}_{y|x} [L(y, f(x)) | \mathbf{x}]] .$$

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

Let $\mathcal{I} : \mathcal{D} \rightarrow \mathcal{H}$ be a learning algorithm that takes a training set

- 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**

The learning method $f^*(\tilde{\mathbf{x}}) = \operatorname{argmin}_c \mathbb{E}_{yx} [L(y, c) | \mathbf{x} = \tilde{\mathbf{x}}]$ is said to be consistent 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 ∞ :



THEORETICAL AND EMPIRICAL RISKS / 2

The risk minimizer is mainly a theoretical tool.

Consistency is defined w.r.t. a particular distribution \mathbb{P}_{xy} . But since we

usually 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

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, KNNs, etc.).

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

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

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

ESTIMATION AND APPROXIMATION ERROR

Goal of learning: Train a model \hat{f}_H for which the true risk $\mathcal{R}(\hat{f}_H)$ is close to the Bayes risk \mathcal{R}^* (also, featureless predictor) gives us a computable empirical lower baseline solution.

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

to be as low as possible.

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

loss — L1 — L2



The Bayes regret can be decomposed as follows:

$$\begin{aligned} \mathcal{R}(\hat{f}_H) - \mathcal{R}^* &= \underbrace{\left[\mathcal{R}(\hat{f}_H) - \inf_{f \in \mathcal{H}} \mathcal{R}(f) \right]}_{\text{estimation error}} + \underbrace{\left[\inf_{f \in \mathcal{H}} \mathcal{R}(f) - \mathcal{R}^* \right]}_{\text{approximation error}} \\ &= \left[\mathcal{R}(\hat{f}_H) - \mathcal{R}(f_{H,x}^*) \right] + \left[\mathcal{R}(f_{H,x}^*) - \mathcal{R}(f_{H,\text{all}}^*) \right] \end{aligned}$$