Lecture 4: Empirical Risk Minimization

Isabel Valera

Machine Learning Group
Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

28.04.2021

Outline

1. Bibliograhy
2. Statistical Learning
3. Empirical Risk Minimization
4. Regularized ERM
5. Bayesian Interpretation
6. Summary

Main references

- "Learning with Kernels." Schölkopf & Smola, MPIT Press 1998 – Chapters 3 & 4.
- "Principles of Risk Minimization for Learning Theory." Vapnik, NeurIPS 1991.
- "The nature of Statistical Learning Theory (2nd edition)." Vapnik, Springer 1999 – Chapters 1 & 4 (for a learning theory perspective).

Outline

1. Bibliograhy
2. Statistical Learning
3. Empirical Risk Minimization
4. Regularized ERM
5. Bayesian Interpretation
6. Summary

Motivation

- Bayesian decision theory allows us to make optimal decisions under uncertainty.
- So far we have only considered the uncertainty that comes from the stochastic nature of the problem, i.e., we only accounted for the fact that the outcomes $y \in \mathcal{Y}$ are **non-deterministic** given the features x.
- In other words, we have assumed that the **probability measure** P on $\mathcal{X} \times \mathcal{Y}$ is known.
- However, in practice we do not have access to the probability measure, but instead we only observe training data generated from such a probability measure.
- **Can we still make "optimal" decisions? In other words, can we still learn a stable (under small changes in the training data) function that maps features into outputs, i.e.,** $y = f(x)$**?**

Statistical Learning Problems

Goal: Reason on the outcome value y for given features x.
Assumption: Only an **independently and identically distributed (i.i.d.)** sample $(X_i, Y_i)_{i=1}^n$ (*training data*) of the probability measure P on $\mathcal{X} \times \mathcal{Y}$ is available. Thus,

- $(X_i, Y_i)_{i=1}^n$ are random variables,
- **independent**: joint density factorizes

$$p\big((x_1, y_1); (x_2, y_2); \ldots; (x_n, y_n)\big) = \prod_{i=1}^{n} p_i(x_i, y_i).$$

- **identically distributed**:

$$p_i(x, y) = p_j(x, y), \qquad \forall i, j \in \{1, \ldots, n\}.$$

and $p(x, y)$ is the density of the data-generating measure P on $\mathcal{X} \times \mathcal{Y}$.

Discriminative versus generative learning

We usually distinguish between two main types of approaches to solve supervised statistical learning problems:

- **Generative Learning:** Estimate the joint distribution $p(x, y)$ (inference) and then use Bayes rule to compute the conditional probability $p(y|x)$.
- **Discriminative Learning:** Directly approximate the conditional distribution $p(y|x)$.

Usually estimating $p(x, y)$ is a harder poblem than approximating the conditional probability $p(y|x)$, thus many methods adopt a discriminative approach. The main advantage of generative learning is that it allows you to sample new (synthetic data) as well as handle uncertainties on the obverved features (e.g., to handle missing values or outliers).

General Principle in Statistics

Statistics: Given an i.i.d. sample $(X_i)_{i=1}^n$, use the empirical measure

$$\mathrm{P}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{x=X_i}$$

to approximate quantities of the data generating measure.

- **Empirical mean**: $\mathbb{E}_{P_n}[X] = \frac{1}{n}\sum_{i=1}^n x\, \mathbb{1}_{x=X_i} = \frac{1}{n}\sum_{i=1}^n X_i$.
- **Empirical variance**: $\mathrm{Var}[X] = \frac{1}{n}\sum_{i=1}^n \left(X_i - \frac{1}{n}\sum_{j=1}^n X_j\right)^2$.
- **Empirical covariance**: $\mathrm{Cov}(X,Y) = \frac{1}{n}\sum_{i=1}^n X_i Y_i - \frac{1}{n}\sum_{i=1}^n X_i \,\frac{1}{n}\sum_{i=1}^n Y_i$.

$\boxed{\mathrm{P}_n \textbf{ approximates } \mathrm{P}}$

Outline

1. Bibliograhy
2. Statistical Learning
3. Empirical Risk Minimization
4. Regularized ERM
5. Bayesian Interpretation
6. Summary

Empirical risk minimization

Definition

Let $(X_i, Y_i)_{i=1}^n$ be an i.i.d. sample of P on $\mathcal{X} \times \mathcal{Y}$, which we call the **training sample**. The **empirical loss** is defined as

$$\mathbb{E}_{\mathrm{P}_n}[L(Y, f(X))] = \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)).$$

Given a class of functions $\mathcal{F}$, **empirical risk minimization** is defined as

$$f_n = \operatorname*{arg\,min}_{f \in \mathcal{F}} \mathbb{E}_{\mathrm{P}_n}[L(Y, f(X))] = \operatorname*{arg\,min}_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)),$$

where f_n is then the optimal learning rule based on the training sample.

Classification

Natural loss: 0-1-loss $L(y, f(x)) = \mathbb{1}_{y \neq f(x)}$.

Empirical risk minimization: minimize the number of errors on the training set:

$$\frac{1}{n}\sum_{i=1}^{n} L(Y_i, f(X_i)) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{Y_i \neq f(X_i)}.$$

Problem: For several classes of functions, minimizing the above empirical risk leads to NP-hard (nondeterministic polynomial time) problems.
Solution: Use of convex margin-based loss functions (refer to Lecture 3).

Regression

Standard loss used in practice: squared loss $L(y, f(x)) = (y - f(x))^2$.

$$f_n = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

- $\mathcal{F} = \{f(x) = \langle w, x \rangle \mid w \in \mathbb{R}^d\}$, **linear least squares regression** (Lecture 5).

Outline

1 Bibliograhy

2 Statistical Learning

3 Empirical Risk Minimization

4 Regularized ERM

5 Bayesian Interpretation

6 Summary

Problems of ERM

- **Problems of ERM:**
  - function class $\mathcal{F}$ too large → overfitting.
  - function class $\mathcal{F}$ too small → underfitting.
- The mapping “data” to “learning rule” can be seen as an *inverse problem*. A **well-posed problem** fulfills that:
  - a solution exists,
  - the solution is unique,
  - the solution depends continuously on the data.

  A problem which does not have one of these properties is called **ill-posed**. In particular the last two properties are most of the time not fulfilled in empirical risk minimization. In order to make problems well-posed one uses **regularization**.

Solution: Assume a large function class $\mathcal{F}$ and use regularization.

Illustration of over/under-fitting.

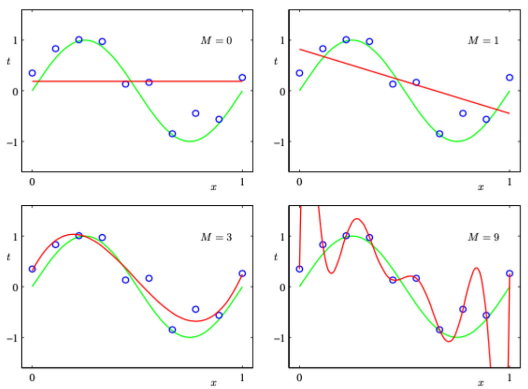

Figure 1.4 Plots of polynomials having various orders M, shown as red curves, fitted to the data set shown in Figure 1.2.

Figure: Image from Bishop.

Regularized empirical risk minimization I

Definition (Tikhonov regularization)

Let

- $(X_i, Y_i)_{i=1}^n$ be the training sample,
- $\mathcal{F}$ a fixed function class,
- $L(y, f(x))$ the loss function,
- $\Omega : \mathcal{F} \to \mathbb{R}_+$ the **regularization functional**.

Then **regularized empirical risk minimization** is defined as

$$f_{n,\lambda} = \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)) + \lambda \Omega(f),$$

where $\lambda \in \mathbb{R}_+$ is called the **regularization parameter**.

Observation: the regularization parameter λ trades-off between **fit of the data** and **complexity of the learning rule**.

Regularized empirical risk minimization II

Proposition (Ivanov regularization)

If the loss $L(y, f(x))$ and the regularization function $\Omega(f)$ are convex in f *and the set $\{f \mid \Omega(f) < r\}$ is non-empty for every $r > 0$ and $\mathcal{F}$ is a convex set, then regularized empirical risk minimization is equivalent to the following problem:*

$$\arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)) \tag{1}$$

$$\textit{subject to } \Omega(f) \leq r \tag{2}$$

in the sense that there exists for each r a corresponding λ such that $f_{n,r} = f_{n,\lambda}$.

Proof: use of duality in convex optimization (refer to Block III).

Regularized empirical risk minimization III

Regularization parameter λ: controls trade-off between data fitting and model complexity.

Limits:

$\lambda \to 0$: selects the least complex function among the "optimal" ones (note that $\lambda \neq 0$), i.e.,

$$\arg\min_{f \in \mathcal{F}^*} \Omega(f), \text{ with } \mathcal{F}^* = \{f \in \mathcal{F} | \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i))\}.$$

$\lambda \to \infty$: considers only functions of zero complexity, $\Omega(f) = 0$, and selects the function which has the smallest loss, i.e.,

$$\arg\min_{f \in \mathcal{F}^*} \frac{1}{n} \sum_{i=1}^{n} L(Y_i, f(X_i)) \text{ with } \mathcal{F}^* = \{f \in \mathcal{F} | \Omega(f) = 0\}.$$

Example: $f : \mathbb{R}^d \to \mathbb{R}$, and $\Omega(f) = \sup_{x \in \mathbb{R}^d} \max_{i=1,\dots,d} \left| \frac{\partial f}{\partial x^i}(x) \right|$

$\Omega(f) = 0 \iff \exists\, c \in \mathbb{R}, \text{ such that } f(x) = c, \forall x \in \mathbb{R}^d.$

Structural risk minimization

Structural risk minimization proposed by Vapnik considers:

- empirical risk minimization over nested function classes $\mathcal{F}_n$, such that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots$,
- as the size of the sample n increases one also allows more complex functions.

Example: start with the linear functions and then add polynomials of increasing order as the number of observations n increases.

Occam's razor

"Occam's razor" (William of Ockham, 1287-1347):

> *"Pluralitas non est ponenda sine necessitas."*
> (Plurality should not be posited without necessity.)

Or similarly:

> *"The simplest explanation is usually the right one."*

In ERM: Between two functions with same loss (risk), select the least complex one measured by Ω.
In ML: Between two ML models with similar performance, select the simpler one (Block II - performance metrics and model selection).

Outline

1. Bibliograhy
2. Statistical Learning
3. Empirical Risk Minimization
4. Regularized ERM
5. Bayesian Interpretation
6. Summary

Bayesian Interpretation

Relation I:

Empirical risk minimization
corresponds to
maximum likelihood estimation.

Relation II:

Regularized empirical risk minimization
corresponds to
maximum a posteriori estimation.

Maximum Likelihood Estimation (MLE)

Problem: Given i.i.d. samples $x_1, \ldots, x_n$ from an unknown probability density $p(x)$, estimate $p(x)$.

MLE Solution:

1. Assume a parametric model of the data generating probability density $p(x \,|\, \theta)$ (i.e., a likelihood model), such that we can evaluate the likelihood (which is a function of the parameters) as:

$$p(x_1, \ldots, x_n \,|\theta) = \prod_{i=1}^{n} p(x_i \,|\, \theta).$$

2. Find parameter $\theta \in \Theta$ by maximizing the likelihood (resp. the log-likelihood), i.e.,

$$\begin{aligned}\operatorname*{arg\,max}_{\theta\in\Theta} \prod_{i=1}^{n} p(x_i \,|\, \theta) &= \operatorname*{arg\,max}_{\theta\in\Theta} \ \log\left(\prod_{i=1}^{n} p(x_i \,|\, \theta)\right) \\ &= \operatorname*{arg\,max}_{\theta\in\Theta} \sum_{i=1}^{n} \log\left(p(x_i \,|\, \theta)\right)\end{aligned}$$

Example - Gaussian model

Gaussian likelihood: $p(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, the variance σ^2 is assumed to be known.

Maximum likelihood estimation of μ:

$$\begin{aligned}\arg\max_{\mu\in\mathbb{R}} \sum_{i=1}^{n} \log\left(p(x_i \mid \mu)\right) &= \arg\max_{\mu\in\mathbb{R}} \sum_{i=1}^{n} \Big(- \frac{\log\left(2\pi\sigma^2\right)}{2} - \frac{(x_i-\mu)^2}{2\sigma^2}\Big) \\ &= \arg\min_{\mu\in\mathbb{R}} \sum_{i=1}^{n} (x_i-\mu)^2\end{aligned}$$

The mean parameter μ^* maximizing the likelihood is (*Exercise*-Proof!):

$$\mu^* = \frac{1}{n}\sum_{i=1}^{n} x_i .$$

Observation: the log-likelihood is convex with respect to μ, thus there exist a unique global minimum (Block III).

ERM vs MLE I

Next, we aim to approximate the conditional distribution (**likelihood**) $p(y|x, f)$, where f denotes the (parameters of the) model, and $\mathcal{F}$ the family of considered functions (characterized by its set of parameters).

Definition

The **maximum likelihood** solution to this problem f_{ML} is then defined as

$$f_{ML} = \arg\max_{f \in \mathcal{F}} \mathrm{P}(D|f) = \arg\max_{f \in \mathcal{F}} \prod_{i=1}^{n} \mathrm{P}(y_i|x_i, f),$$

where $D = (x_i, y_i)_{i=1}^{n}$ denotes the training data.

ERM vs MLE II

Proposition

Given an i.i.d. training sample $(X_i, Y_i)_{i=1}^n$*, a class of functions* $\mathcal{F}$ *and a likelihood* $p(y|x, f)$*, then the* maximum likelihood solution f_{ML} *agrees with the solution of* empirical risk minimization f_n *for the loss function* $L(y, f(x)) = -\log p(y|x, f)$*.*

Observations:

- For a given likelihood function, $p(y|x, f(x))$ we can define an associated loss function as $L(y, f(x)) = -\log p(y|x, f(x))$.
- An arbitrary loss function $L(y, f(x))$ does in general not correspond to a likelihood of the form $p(y|x, f(x)) \simeq e^{-L(y, f(x))}$.

Note:

- output space $\mathcal{Y}$ is discrete: likelihood is probability $\mathrm{P}(y|x, f)$,
- output space $\mathcal{Y}$ is continuous: likelihood is density $p(y|x, f)$.

ERM vs MLE III

Proof: By assumption we know $L(y, f(x)) = -\log \mathrm{P}(y|x, f)$, then

$$\begin{aligned}
f_{ML} &= \arg\max_{f\in\mathcal{F}} \mathrm{P}(D|f) = \arg\max_{f\in\mathcal{F}} \prod_{i=1}^{n} \mathrm{P}(Y_i|X_i, f) \\
&= \arg\max_{f\in\mathcal{F}} \quad \log\left[\prod_{i=1}^{n} \mathrm{P}(Y_i|X_i, f)\right] \\
&= \arg\max_{f\in\mathcal{F}} \quad \sum_{i=1}^{n} \log \mathrm{P}(Y_i|X_i, f) \\
&= \arg\min_{f\in\mathcal{F}} -\sum_{i=1}^{n} \log \mathrm{P}(Y_i|X_i, f) \\
&= \arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} L(Y_i, f(X_i)) = f_n,
\end{aligned}$$

Maximum A Posteriori (MAP) Estimation

Idea: integrate **prior belief** on the model parameters θ
MAP Estimation:

1. Treat the model parameters θ as a random variable, and assume a prior distribution $p(\theta)$ which accounts for prior belief.
2. Use Bayes rule to obtain the posterior of the parameters as:

$$p(\theta \mid x_1, \dots, x_n) = \frac{p(x_1, \dots, x_n \mid \theta) p(\theta)}{p(x_1, \dots, x_n)} = \frac{p(x_1, \dots, x_n \mid \theta) p(\theta)}{\int_\Theta p(x_1, \dots, x_n \mid \theta) p(\theta) d\theta}.$$

   The denominator is called the partition function (or evidence).
3. Find parameters θ by maximizing the posterior distribution, i.e.,

$$\begin{aligned}\arg\max_{\theta \in \Theta} p(\theta \mid x_1, \dots, x_n) &= \arg\max_{\theta \in \Theta} \log \Big(p(x_1, \dots, x_n \mid \theta) p(\theta) \Big) \\ &= \arg\max_{\theta \in \Theta} \left[\sum_{i=1}^{n} \log \left(p(x_i \mid \theta) \right) \right] + \log \left(p(\theta) \right).\end{aligned}$$

Example - Gaussian model

Gaussian likelihood: $p(x|\mu) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, the variance σ^2 is assumed to be known.

Gaussian prior (on the mean parameter): $p(\mu) = \frac{1}{\sqrt{2\pi\sigma_\mu^2}} e^{-\frac{(\mu-\mu_0)^2}{2\sigma_\mu^2}}$ (with known parameters).

MAP estimation of μ:

$$\begin{aligned} \arg\max_{\mu\in\mathbb{R}} p(\mu \,|\, x_1, \ldots, x_n) &= \arg\max_{\mu\in\mathbb{R}} \sum_{i=1}^{n} \log\left(p(x_i \,|\, \mu)\right) + \log\left(p(\mu)\right) \\ &= \arg\min_{\mu\in\mathbb{R}} \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - \mu)^2 + \frac{1}{2\sigma_\mu^2} (\mu - \mu_0)^2 \end{aligned}$$

The MAP estimate μ_{MAP} of the mean parameter is (note that the objective is convex in μ):

$$\mu_{\text{MAP}} = \frac{1}{1 + \frac{\sigma^2}{n\sigma_\mu^2}} \frac{1}{n} \sum_{i=1}^{n} x_i \; + \; \frac{\frac{\sigma^2}{n\sigma_\mu^2}}{1 + \frac{\sigma^2}{n\sigma_\mu^2}} \mu_0.$$

Regularized ERM and MAP estimation I

Assumption: Data $(X_i, Y_i)_{i=1}^n$ is conditionally independent given f and the inputs are independent of f.

Definition

The **maximum a posteriori** estimator for f is defined as

$$f_{MAP} = \operatorname*{arg\,max}_{f \in \mathcal{F}} \mathrm{P}(f|D) = \operatorname*{arg\,max}_{f \in \mathcal{F}} \prod_{i=1}^{n} \mathrm{P}(Y_i|X_i, f)\mathrm{P}(f),$$

where we have discarded $\mathrm{P}(D)$ since it is a constant.

Given a prior over functions $\mathrm{P}(f)$ we define the following regularization functional $\Omega(f)$,

$$\Omega(f) = -\log \mathrm{P}(f) \qquad \Longrightarrow \qquad \mathrm{P}(f) \simeq e^{-\Omega(f)},$$

Regularized ERM and MAP estimation II

Proposition

The MAP estimator f_{MAP} agrees with the minimizer of $f_{\lambda=\frac{1}{n},n}$ of the regularized empirical risk minimization if

$$\begin{array}{lrl} \text{Loss function:} & L(y, f(x)) & = -\log \mathrm{P}(y|x, f), \\ \text{Regularization functional:} & \Omega(f) & = -\log \mathrm{P}(f). \end{array}$$

Regularized ERM and MAP estimation III

Proof.

By assumption we know $L(y, f(x)) = -\log \mathrm{P}(y|x, f)$ and $\Omega(f) = -\log \mathrm{P}(f)$, then

$$
\begin{aligned}
f_{MAP} &= \arg\max_{f\in\mathcal{F}} \mathrm{P}(f|D) = \arg\max_{f\in\mathcal{F}} \prod_{i=1}^{n} \mathrm{P}(Y_i|X_i, f)\mathrm{P}(f) \\
&= \arg\max_{f\in\mathcal{F}} \sum_{i=1}^{n} \log \mathrm{P}(Y_i|X_i, f) + \log \mathrm{P}(f) \\
&= \arg\min_{f\in\mathcal{F}} \sum_{i=1}^{n} L(Y_i, f(X_i)) + \Omega(f) = f_{n,\lambda=\frac{1}{n}}.
\end{aligned}
$$

where we have used that the logarithm is a strictly increasing function. □

Full Bayesian treatment

Posterior predictive distribution:

$$p(x|D) = \int_{\Theta} p(x, \theta \mid D)\, d\theta = \int_{\Theta} p(x \mid \theta, D)\, p(\theta \mid D)\, d\theta$$

- $p(x|D)$ is **not** the true data-generating distribution!
- if the posterior $p(\theta \mid D)$ is very peaked, this is roughly the same as $p(x \mid \theta_{\mathrm{MAP}})$

Supervised learning setting:

$$p(y \mid x, D) = \int_{\Theta} p(y \mid x, \theta)\, p(\theta \mid D)\, d\theta$$

The Bayesian approach to learning considers the full distribution $p(\theta \mid D)$ against the point estimate in the MAP (and ML) estimation (advanced Lecture on Probabilistic ML).

Outline

1 Bibliograhy

2 Statistical Learning

3 Empirical Risk Minimization

4 Regularized ERM

5 Bayesian Interpretation

6 Summary

Summary

- ERM provides an approach to solve learning problems (and thus make decisions under uncertainty) when the probability measure P on the feature space $\mathcal{X}$ and the outcome (output) space $\mathcal{Y}$ is unknown, but we only observe an i.i.d sample of that measure, i.e., **taining data**.
- The key idea of ERM is to find a function that minimizes a given loss evaluated empirically (i.e., using training data). However, if the family of considered functions is too restrictive, we may end up **under-fitting** the data. Otherwise, if the family of functions is very expressive (e.g., a neural network) we may **over-fit** the training data, and thus obtain a function that generalizes poorly for new unseen data.
- **Regularized ERM** provides a framework to learn assuming a flexible function class, while mitigating overfitting via a regularization term that penalizes model (function) complexity.
- Relation between ERM and regularized ERM and, respectively, ML and MAP estimation.
- Block II is about learning regression and classification functions!