Mathematical aspects in machine learning

Lorenz Richter
BTU Cottbus-Senftenberg

October 14, 2017

Contents

1 Statistical learning theory

We consider data from an input space $\mathcal{X}$ and an output space $\mathcal{Y}$, specifically the sample $S_n = \{(X_i, Y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathcal{Y}$. The goal is to learn a (prediction) function $f : \mathcal{X} \to \mathcal{Y}$ that maps input to output data.

Example 1.1. *$\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{-1, 1\}$ (binary classification), $\mathcal{Y} = \mathbb{N}$, $\mathcal{Y} = \mathbb{R}$ (regression).*

In order to develop a proper theory we need to make some assumptions:

- There exists an unknown probability distribution $\mathbb{P}$ over $\mathcal{X} \times \mathcal{Y}$.
- The data S_n are i.i.d. from $\mathbb{P}$, i.e. $(X_i, Y_i) \sim \mathbb{P}$ for every $i = 1, \dots, n$.
- The future data (sometimes called test data) also comes from $\mathbb{P}$.

In order to measure how good we learn the prediction function f we consider a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ that indicates deviations from predictions and true values. Being stochastically more meaningful, we consider for any function f the expected loss

$$L(f) := \mathbb{E}_{(X,Y)\sim\mathbb{P}}[l(f(X), Y)],$$

which in statistics is sometimes called "(Bayes-)Risk". The goal is to come up with a learning algorithm $\mathcal{A} : S_n \mapsto \hat{f}_n$ (i.e. $\hat{f}_n := \mathcal{A}(S_n)$, indicating that the function depends on the n training data) s.t. $L(\hat{f}_n)$ is small. We therefore consider $\mathbb{E}[l(\hat{f}_n(X), Y)|S_n]$, i.e. the expectation conditioned on the samples, which is still a random quantity (as it depends on the sample data)[1].

Definition 1.2. *A predictor f^B is called Bayes-optimal if it minimizes the expected loss, i.e.*

$$L(f^B) = \inf_f L(f) =: L^B.$$

Remark 1.3. *Bayes-optimality depends on $\mathbb{P}, S_n$ and l.*

Example 1.4 (Zero noise or function learning)**.** *One could consider the case that our targets Y are deterministically prescribed by a function g, i.e. $\mathbb{P}(Y = g(X)|X = x) = 1$. However, this is a rather unrealistic case.*

Example 1.5 (Binary classification)**.** *Consider $\mathcal{Y} = \{0, 1\}$ and the loss $l(y', y) = \mathbb{1}\{y' \neq y\}$. One can show that the Bayes-optimal predictor (in this case classifier) is $f^B(x) = \mathbb{1}\{\mathbb{P}(Y = 1|X = x) > \frac{1}{2}\}$.*

Example 1.6 (Regression)**.** *We consider $\mathcal{Y} = \mathbb{R}$ and $l(y', y) = (y' - y)^2$. Then the Bayes-optimal predictor is $f^B(x) = \mathbb{E}[Y|X = x]$.*

[1] We will continuously omit the measure $\mathbb{P}$ in the expected value.

Somehow we need to work with our sample S_n. We therefore define the empirical risk to be

$$L_n(f) = \frac{1}{n}\sum_{i=1}^{n} l(f(X_i), Y_i).$$

We see that $\mathbb{E}[L_n(h)] = L(h)$ and can therefore consider empirical risk minimization (ERM), namely

$$\hat{f}_n := \arg\inf_f L_n(f)$$

as a reasonable strategy to learn the function f. However, the following example shows that this is not always a good idea.

Example 1.7. *Consider $\mathcal{X} = [0,1], \mathcal{Y} = \{0,1\}, l(y', y) = \mathbb{1}\{y' \neq y\}$ and $\mathbb{P}(Y = 1|x) = 1$ for all $x \in \mathcal{X}$. Define the predictor as*

$$\hat{f}_n(x) = \begin{cases} 1 & \textit{if } (x,y) \in S_n \textit{ for some value } y \\ 0 & \textit{otherwise} \end{cases}.$$

Note again that this predictor is not fixed, but depends on the data. We can see that it performs very badly: $L_n(\hat{f}_n) = 0$, but $L(\hat{f}_n) = 1$. This is a stereotypical example of what is described as overfitting.

2 Bibliography

[1] Hartmann, C. (2017). Wahrscheinlichkeitstheorie*. Lecture script, BTU Cottbus-Senftenberg.*

[2] Shalev-Shwartz, S. and Ben-David, S. (2014). Understanding machine learning: From theory to algorithms*. Cambridge university press.*