Guarantees for Machine Learning, Fall 2025

Lecture 1: Introduction and concentration bounds

Who is here?

Which department?

1. Computer Science
2. Mathematics/Statistics
3. Data Science
4. ITET & Robotics
5. Others

What stage of your studies are you?

1. Masters
2. PhD student
3. Bachelors

Class intro

Objective. Develop graduate students into researchers who can

- understand and criticize papers in ML theory
- conjecture and prove new theorems with high impact

Prerequisites

- Familiar with core machine learning concepts
- Should be comfortable writing rigorous mathematical proofs (for D-MATH courses)

Course structure

- First part: classical techniques for non-asymptotic risk bounds
    - Core reference: Martin Wainwright: High-dimensional statistics (available for free online via ETH)
- Second part: projects that review and extend current papers

Logistics

- Class website sml.inf.ethz.ch/gml25/syllabus.html - material
- Lecture slides will be uploaded after lectures
- TAs: Tobias Wegel, Julia Kostin (Office hours on request)
- Internet platforms to sign up for: moodle (logistical questions and on material, teammate search)
- Important announcements: in class and per email

Evaluation

- 2 homeworks (0% but pass mandatory to take exam), oral midterm (60%), project (40%)
- Homeworks (different than previous years):
  - is pass-fail
  - TAs will read your homework: to pass a homework you need to have written down attempts to solve **all** questions (except those marked optional/bonus). Explain your reasoning steps and if you're stuck, explain why.
  - You can work with others, but indicate with whom.
- Oral midterm will be 17./18.11. and we will prepare a schedule where everyone can make it
- Project work (presentation & report) and peer feedback for other presentations

Project

In groups of two:

- Pick a paper from a list (to be announced) according to your interests & background, on (**October 14**)
- You can pick your own, but double-check with Tobias and me
- Understand, discuss & follow-up on one paper with substantial theoretical content related to class
- 15-20 min presentation in last two weeks of class in December
  - present the paper and main proof ideas
  - present your discussion and extension/follow-up attempts/results
- around 10 pages of written report (due **January 12**)

Enrollment

- Current waitlist: >150. Admitted: 30. Limit for admissions: 30
- By experience, everybody who wants to take it, can
- For that: If you find after first weeks and homework it's not for you, please de-register to make space for others!
  - Final deadline to de-register: **October 8th**
  - if you fail to do that, you will be graded (if none of the above are attempted it shows as a no-show in your transcript)
- Others welcome to audit as long as there is space

Plan for today

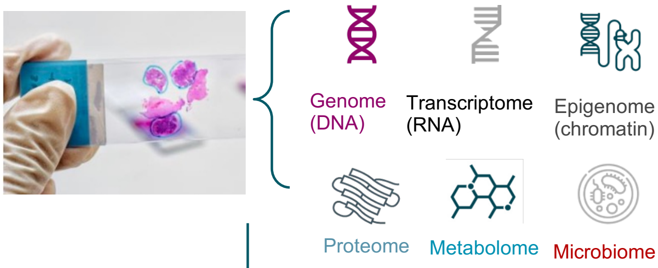
- Statistical perspective on the supervised learning pipeline
- Evaluation of an estimator using the excess risk
  and what the course is about
- Concentration bounds of empirical means

Recap: (Supervised) Machine Learning - Classification

input x

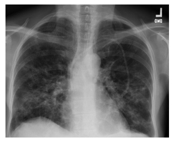
output y

x: multiomics



y: cancer survival

x: -Ray images



y: pneumonia or healthy

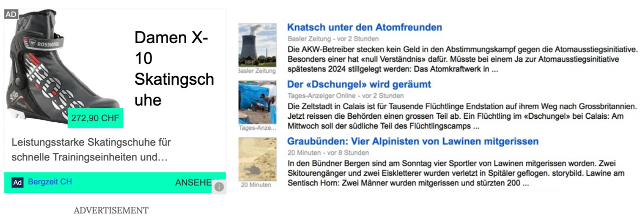
Figure 1: Classification examples

Recap: (Supervised) Machine Learning - Regression

input x

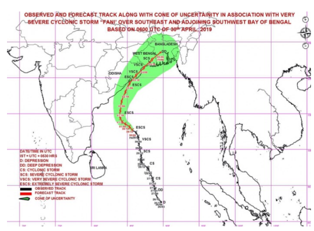
output y

x: User data & article/advertisement



y: Probability of click / purchase

x: storm track



y: storm speed

Figure 2: Regression examples

Statistical Perspective on (supervised) Machine Learning

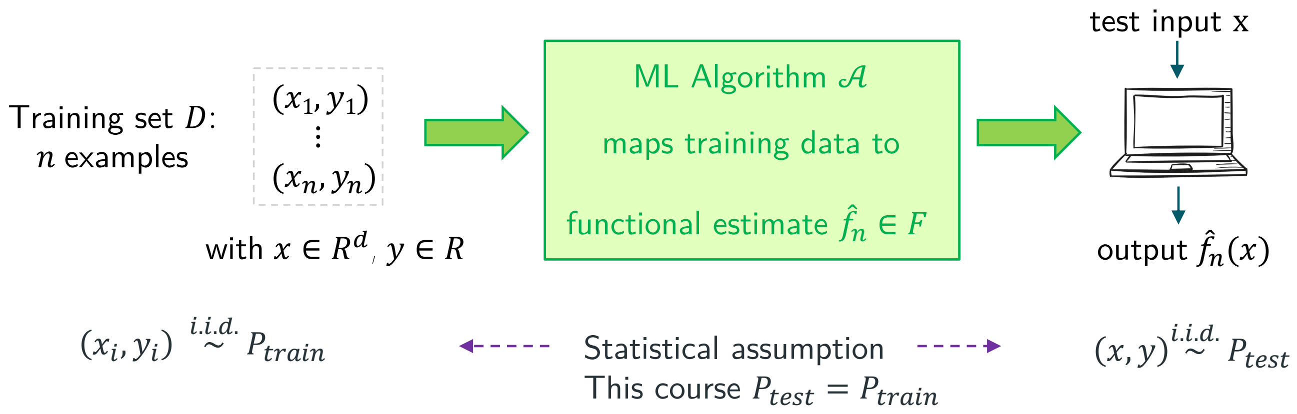


Figure 3: Supervised learning pipeline from statistical point of view

- some examples for $\mathbb{P} = \mathbb{P}_{train} = \mathbb{P}_{test}$ include
  - regression: marginal dist. over x and $y = f^\star(x) + \epsilon$ for random ϵ
  - classification: generative such as Gaussian mixture model or discriminative: marginal dist. over x and $y = \text{sign}(f^\star(x))$

Evaluation of an estimator $\widehat{f}_n$

The estimate $\widehat{f}_n \in \mathcal{F}$ depends on $\mathcal{D} = (x_i, y_i)_{i=1}^n$ (with random $(x_i, y_i) \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ for all i) and is in some function class $\mathcal{F}$ (e.g. linear, neural network etc.). Whether $\widehat{f}_n$ is "good" is decided during test time: On average over test points (x, y), we'd like the predictions $\widehat{f}_n(x)$ to be close to y.

We now formalize this:

- For a given point (x, y), we measure "closeness" via a pointwise loss ℓ, e.g. $\ell((x, y), f) = (f(x) - y)^2$ for regression or $\ell(x, y; f) = \mathbb{1}_{f(x)=y}$ for classification
- We call the average loss of any function f the *population risk* $R(f) := R(f; \mathbb{P}) = \mathbb{E}\ell((x, y); f)$ where the expectation $\mathbb{E}$ is over x, y following distribution $\mathbb{P}$. This is the "average over test points" that determines how "good" $\widehat{f}_n$ really is!

Evaluation of an estimator $\widehat{f}_n$

- We further call the training loss of any f the *empirical risk* $R_n(f) := R(f;\mathcal{D}) = \frac{1}{n}\sum_{i=1}^n \ell((x_i,y_i);f)$. Recall that we also assume $(x_i,y_i) \overset{\text{i.i.d.}}{\sim} \mathbb{P}$ for all i
- In the next lectures we'll consider the empirical risk minimization paradigm where $\widehat{f}_n$ minimizes the training loss (or a modified variant thereof, later in the course)

$$\widehat{f}_n := \arg\min_{f\in\mathcal{F}} R_n(f)$$

Q: For classification, is $R(\widehat{f}_n) = 20\%$ bad or good?

A: Depends on how hard the task is! Perhaps it's not possible to achieve perfect accuracy!

We should compare population risk of $\widehat{f}_n$ with that of the best possible function *if we knew the full distribution*, i.e. evaluate the excess risk:

$$\mathcal{E}_R(n) := R(\widehat{f}_n) - \inf_f R(f)$$

Evaluation of an estimator $\widehat{f}_n$

Grab a neighbor and discuss for 5 minutes.

1. How is the population risk of an estimator related to its test error?
2. Which parameters of the problem setting (population risk) and algorithm does the excess risk depend on? What happens to the excess risk of an estimator $\widehat{f}_n$ when we vary these parameters?
3. When we consider the *empirical risk minimizer* $\widehat{f}_n := \arg\min_{f\in\mathcal{F}} R_n(f)$ specifically, what are some tradeoffs when choosing $\mathcal{F}$?

Questions on the excess risk

1. Population risk vs. test error

- Test error on n' new samples follows $R_{n'}(\widehat{f}_n) \to R_n(\widehat{f}_n)$ by law of large numbers (LLN)

2. Excess risk depends on model class $\mathcal{F}$, dimensionality of the data d, sample size n and consists of the following factors and trends

- approximation error (if $f^\star = \arg\min_f R(f)$ is complicated): larger $\mathcal{F}$, smaller d better
- optimization error (due to optimization algorithm): Lipschitz, (strong) convex loss ℓ better
- statistical error (due to finite sample and noise): larger n (usually) better (depends on $\mathcal{F}, d$ as well) of course $\leftarrow$ this course

3. Tradeoff: Larger $\mathcal{F}$, bigger effect of noise (statistical error) but smaller approx error (variance vs. bias)

This course: Non-asymptotic take on statistical "Guarantees for Machine Learning"

We introduce general frameworks to analyze excess risk and compute concrete upper (and lower) bounds s.t. with prob. at least $1 - \delta$

$$R(\widehat{f}_n) - R(f^\star) \leq UB(n, d, \mathcal{F}, f^\star)$$

where we assume $f^\star = \arg min_f R(f)$ exists.

Questions we'd like to answer:

1. Does UB converge to 0 as n increases? (consistency)
2. If I collect double as much data, how much do I decrease my excess risk? $\to$ boils down to the exponent of n (statistical rate)

This course focuses on 2. We'll now discuss some probabilistic basics that give a sense for what to expect from course later.

Excess risk decomposition

- Recall the population risk $R(f) = \mathbb{E}\ell((X, Y); f)$
- Recall the empirical risk $R_n(f) = \frac{1}{n}\sum_{i=1}^{n} \ell((X_i, Y_i); f)$
- Remember we want to bound the excess risk

$$R(\widehat{f}_n) - R(f^\star) = R(\widehat{f}_n) - R_n(\widehat{f}_n) + \overbrace{R_n(\widehat{f}_n) - R_n(f^\star)}^{T_3 \leq 0} + R_n(f^\star) - R(f^\star)$$
$$\leq \underbrace{R(\widehat{f}_n) - R_n(\widehat{f}_n)}_{T_1} + \underbrace{R_n(f^\star) - R(f^\star)}_{T_2}$$

Question: Are T_1 and T_2 qualitatively similarly hard to bound? Is $T_3 \leq 0$ always true? Briefly discuss with your neighbor.

- $T_3 \leq 0$ is only true when $f^\star \in \mathcal{F}$!
- T_1 is harder than T_2 since it's a sum of dependent variables whereas T_2 is the difference between an emprical mean and its expectation.

Concentration bounds for single random variables (R.V.)

- *Markov* inequality: $\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}X}{t}$ for $X \geq 0$;
- Markov used on $e^{\lambda(X-\mathbb{E}X)}$ for $\lambda \geq 0$ yields the *Chernoff* bound

$$\mathbb{P}(X - \mathbb{E}X \geq t) \leq \inf_{\lambda \geq 0} \frac{\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}]}{e^{\lambda t}}$$

where where we use that $e^{\lambda x}$ is monotonically increasing for $\lambda \geq 0$ and assume the moment generating function (MGF) $\mathbb{E}e^{\lambda X}$ exists

We can use Chernoff to get tighter bounds for R.V. X with short tails

Definition (Sub-Gaussian random variables)

A random variable X with mean μ is sub-Gaussian w/ parameter σ if

$$\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\lambda^2\sigma^2/2} \quad \text{for all } \lambda \in \mathbb{R}$$

- For σ sub-Gaussians using Chernoff we obtain the tail bound

$$\mathbb{P}(X - \mathbb{E}X \geq t) \leq \inf_{\lambda \geq 0} e^{\frac{\lambda^2\sigma^2}{2} - \lambda t} = e^{-\frac{t^2}{2\sigma^2}}$$

Examples for sub-Gaussian random variables

- Gaussians $\mathcal{N}(0, \sigma^2)$ are sub-Gaussian with parameter σ
- *Rademacher variables* $\epsilon = -1, +1$ with equal probability $1/2$ are sub-Gaussian with parameter $\sigma = 1$
  - We can directly compute and bound their MGF

$$\mathbb{E}e^{\lambda\epsilon} = \frac{1}{2}(e^{-\lambda} + e^{\lambda}) \leq e^{\lambda^2/2}$$

- Almost surely bounded in $[a, b]$ (exercise)

Empirical means of independent subgaussians

Lemma (Hoeffding's inequality)

For i.i.d sub-Gaussian R.V. X_i, it holds that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}X \geq t\right) \leq e^{-\frac{nt^2}{2\sigma^2}}$$

Neighbor-Q: Prove Hoeffding's inequality

- Recall sub-Gaussian: $\mathbb{E}e^{\lambda(X-\mu)} \leq e^{\lambda^2\sigma^2/2}$ for all $\lambda \in \mathbb{R}$
- Recall Chernoff for sub-Gaussians: $\mathbb{P}(X - \mathbb{E}X \geq t) \leq e^{-\frac{t^2}{2\sigma^2}}$

Proof of Hoeffding's inequality

1. We can apply Chernoff on the mean of n independent random variables with moment generating function

$$\mathbb{E}e^{\lambda(\frac{1}{n}\sum_{i=1}^n(X_i-\mathbb{E}X_i))} = \prod_{i=1}^n \mathbb{E}e^{\frac{\lambda}{n}(X_i-\mu)} = [\mathbb{E}e^{\frac{\lambda}{n}(X_i-\mu)}]^n$$

2. Hence, the mean of n i.i.d. sub-Gaussian variables is sub-Gaussian with parameter $\frac{\sigma}{\sqrt{n}}$ since $\mathbb{E}e^{\lambda(\frac{1}{n}\sum_{i=1}^n(X_i-\mathbb{E}X_i))} \leq e^{\frac{\lambda^2\sigma^2}{2n^2}n}$
3. yielding Hoeffding's inequality for the mean of iid sub-Gaussians

$$\mathbb{P}(\frac{1}{n}\sum_{i=1}^n X_i - \mathbb{E}X \geq t) \leq e^{-\frac{nt^2}{2\sigma^2}}$$

Q: How can we now use Hoeffding's inequality to bound the term T_2?

Syllabus of course

The courses focuses on bounding T_2 using so-called (localized) uniform convergence.

We'll cover

- uniform convergence using Rademacher and Gaussian complexity
- metric entropy and chaining to bound the complexity
- application to non-parametric regression (kernel methods)
- minimax lower bounds
- recent topics in statistical machine learning

References

Concentration bounds:

- MW Chapters 2

Excess risk:

- MW Chapter 4