Statistical Foundations of Learning

Debarghya Ghoshdastidar

School of Computation, Information and Technology
Technical University of Munich

Course Organisation and Overview

Statistical Foundations of Learning

- Module: CIT4230004 (replaces IN2378)
- ECTS: 8 credits (THEO) ; 4V+2Ü SWS
- Intended audience: Master
- Pre-requisites:
  - Machine learning (IN2064)
  - Probability (IN0018) or equivalent ; Course in statistics is a bonus
  - Some knowledge of linear algebra (MA0901), analysis (MA0902)

Contact persons

- Lecturer:
  - Debarghya Ghoshdastidar
  - ghoshdas@cit.tum.de / Room: 03.11.043
- Teaching assistants
  - Maximilian Fleissner
  - maximilian.fleissner@tum.de
- Office hours: By appointment or Ask on Moodle Q&A forum

Course format

- Lectures (Debarghya Ghoshdastidar)
  - Tuesdays / 16:15–17:45 / 00.13.009A and TUMLive
  - Wednesdays / 8:15–9:45 / 00.08.038 and TUMLive
- Tutorials (Maximilian Fleissner)
  - Fridays / 14:15-15:45 / 00.13.009A (no streaming)
  - Participants encouraged to upload discussed solutions on Wiki (Moodle)
- Further Q&A and lecture materials on Moodle
  Weekly schedule also on Moodle (see changes in first 2 weeks of May)

Assessment

- Grades will be based on final examination
  - Written exam, 120 minutes on **01.08.2024** (oral only if too few registrations)
  - Must be registered in TUM to write exam / get grades (if not, then contact me)
  - There will be retake exam, possibly on **24.09.2024** (written/oral based on number of registrations)
- Bonus 0,3 or 0,4 note possible from assignments
  - Example: 2,3 + bonus = 2,0; 2,7 + bonus = 2,3
  - No bonus if final exam note is 1,0 ; 4,0 ; 4,3 ; 4,7 ; 5,0
- If you have obtained bonus in IN2378 / CIT423004 in a previous semester, then bonus from the previous course will be considered

Assignments

- There will be 6 written assignments, which will be graded (not mandatory)
  - Securing 70% of total points will lead to bonus in final note
- Submit assignments through Moodle as single PDF
  - You are welcome to discuss assignments in small groups
  - But submissions MUST be individually written / submitted
- Deadline specified on Moodle (typically 10-12 days after handed out)
  - **NO late submissions**

Purpose / expected outcomes of this course

- Understand supervised machine learning from a statistics perspective
  - Learn the fundamentals of the *statistical learning problem*
  - Learn how to theoretically analyse a ML algorithm (without experiments)
  - Glimpse of recent statistical theories in machine / deep learning
- This course is NOT about practical ML
  - No coding / Practical motivations / Examples on datasets
  - May not help in your other ML projects

What should you expect from this course?

- Good knowledge of ML (IN2064); Probability (IN0018 or equivalent) will be assumed
- Depending on your background, you may enjoy this course or find it too difficult
  - Master Informatik: Not an easy THEO course
  - Master Mathematics / Math in DS: Course partly aligned with MA4801
  - Others: Course is purely mathematical ; Not ideal for learning ML
- To get idea of math level:
  - See sample lecture videos on Moodle
  - Try first sheet of sample problems (should be able to solve at least 50% by yourself)

What should you expect from this course?

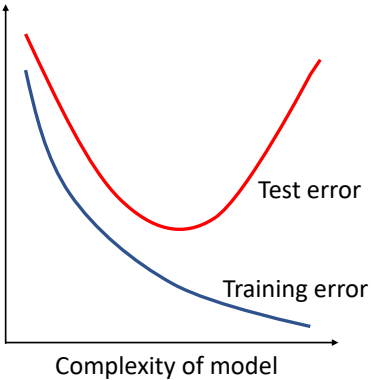
- Statistics from SoSe 2022 (IN2378)
  - 313 registered for course
  - 51 submitted 1st assignment / 33 submitted assignments till end
  - 115 registered for exams / only 36 appeared for exam / 28 passed (average note 2,3)
- If first 2-3 weeks seem too difficult / time-consuming, consider dropping / contact us

Course content

- Part - I: Classical theories (mostly for binary classification)
  - Statistical learning framework and Bayes error
  - Theories of generalisation
    Vapnik-Chervonenkis theory ; Statistical consistency ; Algorithmic stability
  - Applications of above to derive performance guarantees of
    k-nearest neighbour ; support vector machines ; 2-layer neural network ; boosting
  - Probably approximately correct (PAC) learning and No Free Lunch theorem
- Part II: Recent theories
  - Regression: Double descent phenomena, Neural tangent kernel
  - Clustering: Worst-case approximation bounds, Cost of explainability

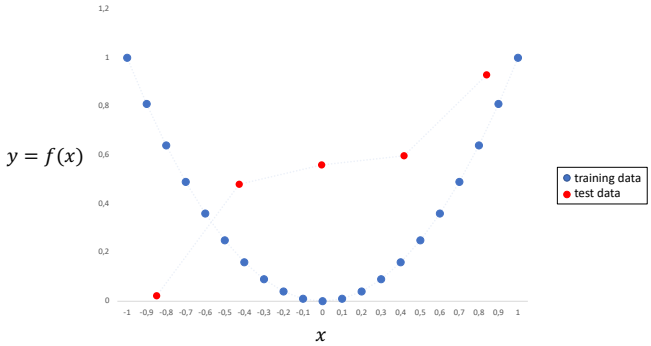
A Motivating Question

- Would you like your ML model to overfit on training data?
- Why do overfitted models show high test error?



Essence of Theory of Generalisation

- We cannot predict well if test data too different from training data

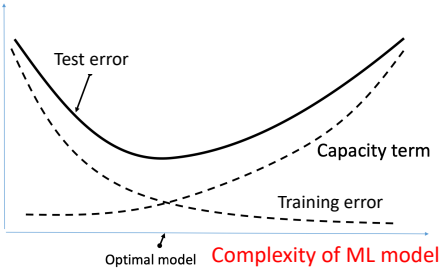


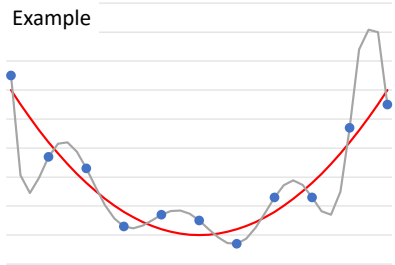
- Assume training and test data samples from same probability distribution $(x, y) \sim \mathcal{D}$
- Goal: Find model $h(\cdot)$ that minimises $\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[(y - h(x))^2\right]$

Essence of Theory of Generalisation

- Theory of Generalisation:

$$\underbrace{\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[(y-h(x))^2\right]}_{\text{expected test error}} \leq \underbrace{\frac{1}{n}\sum_{i=1}^{n}(y_i-h(x_i))^2}_{\text{training error}} + \sqrt{\frac{1}{n}\cdot \text{complexity of ML model}}$$





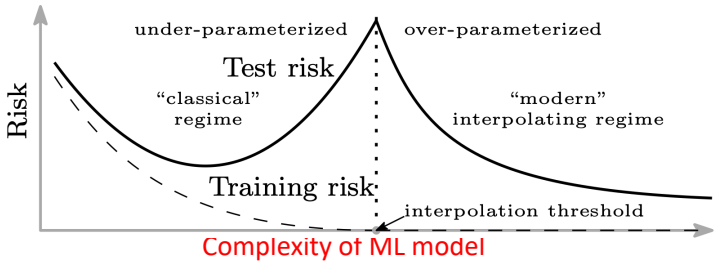
Are complex models (that can overfit) really bad?

Benchmarks on ImageNet

1.2M training images

Rank	Model	Top 1 Accuracy	Top 5 Accuracy ↑	Number of params
1	**Florence-CoSwin-H**	90.05%	99.02%	893M
2	**Meta Pseudo Labels** (EfficientNet-L2)	90.2%	98.8%	480M
3	**Meta Pseudo Labels** (EfficientNet-B6-Wide)	90%	98.7%	390M
4	**FixEfficientNet-L2**	88.5%	98.7%	480M
5	**NoisyStudent** (EfficientNet-L2)	88.4%	98.7%	480M

Double descent phenomena



- In Part II, we may see some mathematical theories on why this happens
- Is above plot "at odds" with the notion "overfitting is bad"?

Reading material and references

- Part I:
  - Shai Shalev-Shwartz & Shai Ben David.
    Understanding Machine Learning: From theory to algorithms (available online)
  - Luc Devroye, László Györfi & Gábor Lugosi.
    A probabilistic theory of pattern recognition (available online)
- Part II:
  - Lecture slides (additional references will be mentioned, but not necessary)

Feedback / emergency contacts

- For feedback / issues, talk to us or use anonymous feedback module on Moodle
- For further support:
  - If you experience / observe any discrimination or insensitive behaviour, including racism/sexism, contact
    - Fachschaft
    - TUM Compliance office
    - Gender equality officers at CIT School or CS Department