Machine Learning for Graphs and Sequential Data

Introduction

Lecturer: Prof. Dr. Stephan Günnemann

www.daml.in.tum.de

Summer Term 2023

Data Analytics and
Machine Learning

Roadmap

- Chapter: Introduction
  1. **What will you learn in this lecture?**
  2. Organizational aspects + project tasks

What is this course about?

- In short: A continuation of our intro ML lecture (IN2064) now focusing on advanced learning principles and covering more complex data domains
- Focus on algorithms and general principles, not limited to a single domain
- Project tasks will give you hands-on experience

➢ At the end you should also be able to extend existing techniques and adapt them to applied problems

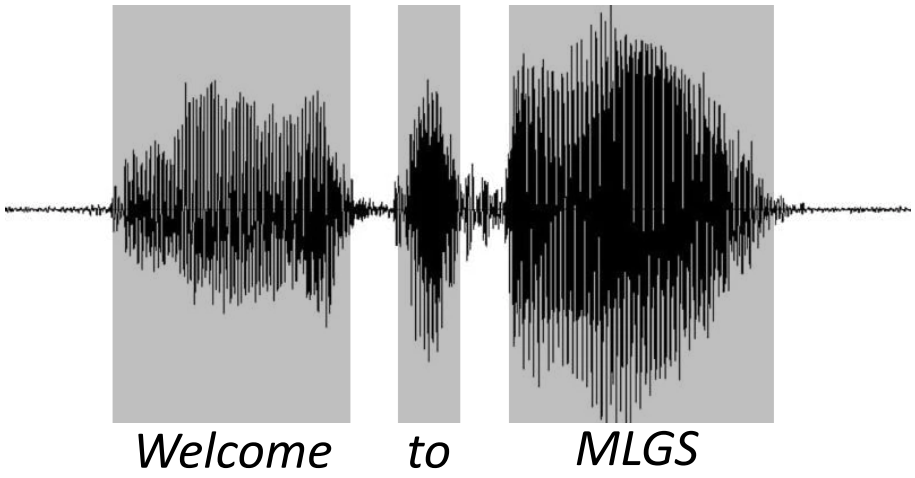
How do you learn from sequences?

- How can we understand written language?

Duchess Kate PERSON did wear an Alexander McQueen BRAND dress ITEM to the wedding OCCASION in the fall of 2017 SEASON .

[Akbik18]

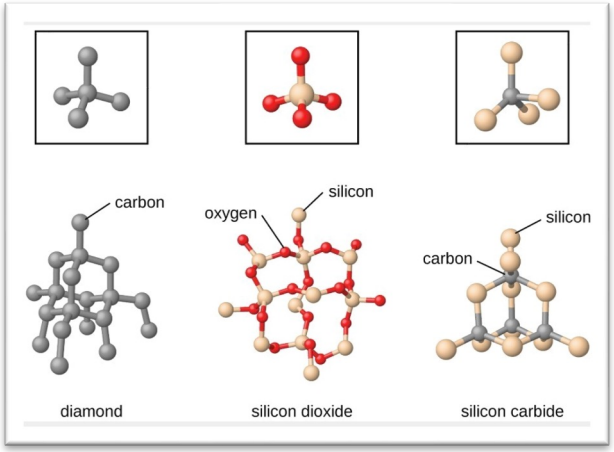
- How do you detect patterns in sound waves?



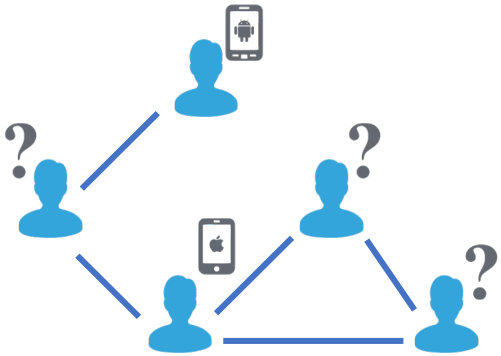
- How can we predict the future?

How to predict…?

- …the properties of molecules?
  - instances are not vectors but graphs!



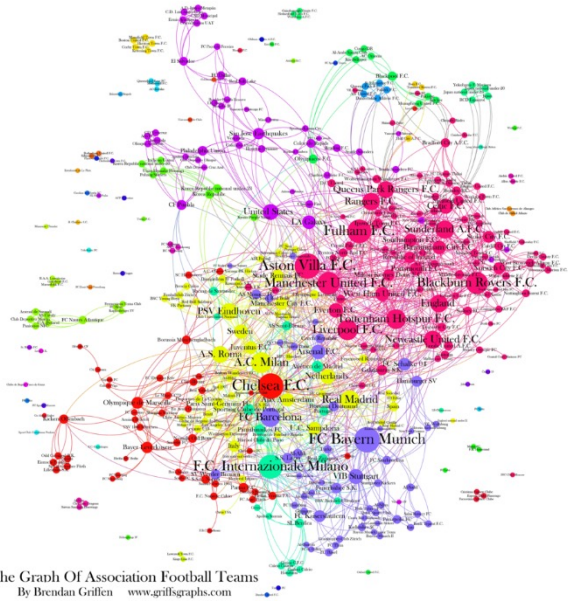
- …the preference of a user in a social network?
  - instances are connected by a graph



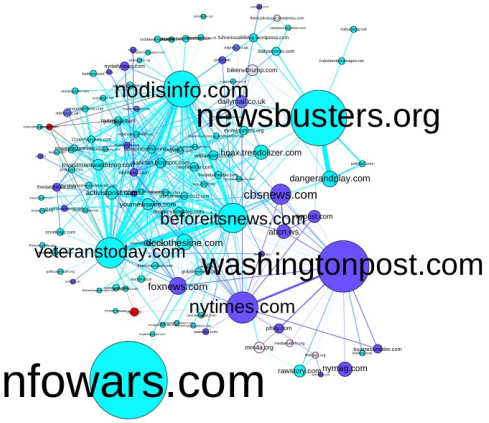
How to use/incorporate the graph structure for classification?

How to exploit graph/network data…?

- … to find communities of people?



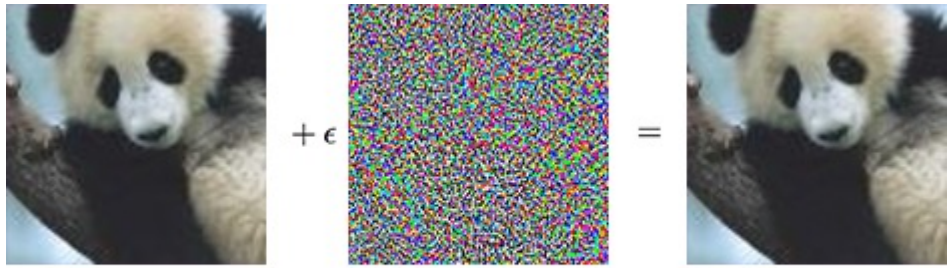
- … to detect relevant websites?



← or the most favorite football clubs ☺

Is Deep Learning fragile?

- Neural networks can be tricked by small perturbations

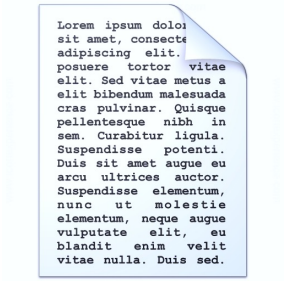


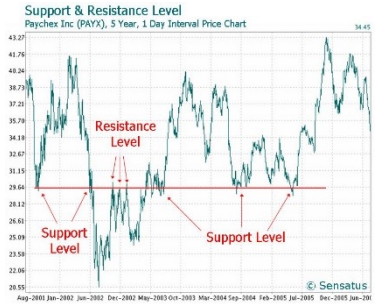
[Goodfellow14]

- How can we find such a perturbation?
- How do you measure the susceptibility of a model to these attacks?
- Can we ensure the robustness of our models against such adversarial examples, e.g. through specialized training procedures?

The Challenge of Non-I.I.D. data

- Classical Machine Learning approaches assume data instances to be i.i.d.
  - Independent, identically distributed
- However, most real-world datasets show dependencies
  - Temporal data/sequences
    - Text: sequence of words (discrete values)
    - Sensor measurements: time series (continuous values)
    - Current value depends on previous values
  - Graphs/networks
    - Already seen before: molecules, social networks, knowledge graphs, connected devices (IoT)
    - Arbitrarily complex dependency structure possible





- How to design machine learning techniques that consider these dependencies (and thus, hopefully, lead to better results)?

Contents of this Course

- Temporal & Sequential Data
  - Autoregressive models, HMM
  - Deep learning on sequences: RNN, LSTM, embeddings, Transformer
  - Temporal point processes
- Graphs
  - Laws & generative models
  - PageRank
  - Unsupervised representation learning: spectral clustering, embeddings
  - Supervised learning: label propagation, GNNs & their limitations
- Robustness
  - Adversarial examples
  - Attacks & defenses, certificates

Roadmap

- Chapter: Introduction
  1. What will you learn in this lecture?
  2. **Organizational aspects + project tasks**

Course Organization

- Lecturer
  - Prof. Dr. Stephan Günnemann
- Teaching assistants
  - Dominik Fuchsgruber, David Lüdke, Yan Scholten, Jan Schuchardt, Johanna Sommer

- 5 ECTS
- Language: English

- Ungraded exercise sheets
- Graded project tasks every 2-3 weeks

- Final exam + repeat exam
- Project tasks can grant a bonus of up to 0.3

Schedule

- In-person classroom (Timetable on Moodle)
  - Lecture and/or Exercise: Wednesday and Thursday 14:00 – 16:00
- Practice material and exercises uploaded to Moodle
  - Ungraded exercise sheets
- Project tasks via Artemis

Preliminary Timetable

Date		Topic	Projects
Wednesday	April 19, 2023	**Introduction/AR MC**	
Thursday	April 20, 2023	**AR & Markov Chains**	
Wednesday	April 26, 2023	**HMM**	**Project**
Thursday	April 27, 2023	*Exercise AR & MC HMM*	
Wednesday	May 3, 2023	**Sequence models**	
Thursday	May 4, 2023	**NNs**	**Project**
Wednesday	May 10, 2023	*Exercise Sequential*	
Thursday	May 11, 2023	**TPPs**	**Project**
Wednesday	May 17, 2023	*Exercise TPP*	
Wednesday	May 24, 2023	**Graphs - Laws, Patterns (Video)**	
Thursday	May 25, 2023	*Exercise* Graphs - *Laws, Patterns*	
Wednesday	May 31, 2023	**Graphs - Clustering**	**Project**
Thursday	June 1, 2023	**Graphs - Embedding & Ranking Pt. 1**	
Wednesday	June 7, 2023	**Graphs - Embedding & Ranking Pt. 2**	**Project**
Thursday	June 15, 2023	*Exercise* Graphs - *Embedding, Ranking, Clustering*	
Wednesday	June 21, 2023	**Graphs - Classification**	
Thursday	June 22, 2023	**GNN Pt. 1**	
Wednesday	June 28, 2023	**GNN Pt. 2**	**Project**
Thursday	June 29, 2023	**Robustness Pt. 1**	
Wednesday	July 5, 2023	*Exercise Classification and GNN*	
Thursday	July 6, 2023	**Robustness Pt. 2**	
Wednesday	July 12, 2023	**Robustness Pt. 3**	**Project**
Thursday	July 13, 2023	*Exercise Robustness*	

Prerequisites

- The course is designed for Master students of Computer Science (and specializations such as Data Engineering and Analytics, Games Engineering, etc.)

- Prerequisites:
  - Knowledge about the standard Machine Learning concepts (i.e. content of our lecture IN2064)
    - We assume the basic concepts are clear; no repetition!
    - We strongly recommend that you attend IN2064 first before taking this class
  - Knowledge about:
    - Algorithms and Data structures
    - Programming
    - Mathematics: Linear Algebra, Statistics, Optimization

Course Material + Announcements

- All course materials (slides, exercises) will be uploaded to Moodle
  - Video recordings of previous years lectures are accessible via link on Moodle
- Project submission via Artemis
- Use Piazza to ask questions! (please avoid sending e-mails)

http://piazza.com/tum.de/summer2023/in2323
Access Code: **mlgs2023**

- Please read the guidelines for using Piazza

Exercises and Project Tasks

- Exercise sheets
  - Exam preparation
  - Solutions in the tutorials
  - Due to the high number of registrations, we are unable to provide corrections to your solutions

- Project tasks
  - Get hands-on experience with advanced machine learning methods
  - Improve your final grade! (details later)

Project Format

- Format of programming tasks
  - Tasks will be published via Artemis
  - artemis.in.tum.de
- How to solve programming tasks?
  - Clone template repository from Artemis exercise
  - Solve tasks described in the repository
  - Push repository with filled-in solutions
- Bonus regulations
  - 10 points for each programming sheet
  - A Bonus of 0.1, 0.2 and 0.3 grade points will be granted upon correct completion of 25%, 50% and 75% of all project points, respectively

Project topics

- Seven project tasks on the following topics
  - Hidden Markov Chain
  - Word2vec
  - Temporal Point Processes
  - Clustering
  - Ranking
  - Graph neural networks
  - Robustness

- The specific tasks and all details will be described in the corresponding exercise sheets

Exam & Grading Scheme

- Written final exam: 90 minutes
  - Date will be announced via TUMonline
  - We currently plan with an on-site exam
  - One handwritten two-sided A4 sheet with notes

```
def final_grade(exam_grade, project_grade):
    if exam_grade > 4.0:
        return exam_grade
    else:
       return max(1.0, exam_grade - bonus)
```

→ The project is voluntary and can only improve the final grade. The project bonus applies only if you passed, and you cannot improve beyond 1.0

Our Group's Focus

Reliable Machine Learning for **Non-Independent** Data

- Interested? We offer:
  - Bachelor/Master theses, Guided Research projects, HiWi positions
- More details on specific topics closer to the end of the semester

References

Figures taken from

- Goodfellow et al. 2014, https://arxiv.org/abs/1412.6572
- Akbik et al. 2018, https://research.zalando.com/welcome/mission/research-projects/flair-nlp/
- Khan 2019, https://heartbeat.fritz.ai/stylegans-use-machine-learning-to-generate-and-customize-realistic-images-c943388dc672
- Liao et al. 2019, https://arxiv.org/abs/1910.00760