Machine Learning Lab Course

Organizational Meeting

lecturer: Prof. Dr. Stephan Günnemann

Summer Term 2018
This is a practical course (Praktikum) for **Master** students!

*Name of module: Large-Scale Machine Learning (IN2106, IN4192)*

website: [ml-lab.in.tum.de](http://ml-lab.in.tum.de)
Why attend our Machine Learning lab course?

1. Get the chance to **implement and apply** state-of-the-art ML algorithms

2. Gain **hands-on experience** working on real-world data, solving real-world tasks (e.g. by working on one of the projects by our **industry partners**).
   - Successful projects might even qualify for a subsequent master thesis.

3. Work on **large-scale problems** with the support of state-of-the-art **GPU computing resources**.
Requirements

- **Requirements for the lab course**
  - *strong programming skills* (Java, Python, C++, Java, etc.)
  - strong knowledge in data mining/machine learning
  - you should have passed relevant courses (the more, the better)
    - Mining Massive Datasets
    - Machine Learning
    - Our seminars
  - self-motivation

- **Additional selection criteria**
  - other *relevant* experience (projects in companies, experience as a HiWi)
    - you can send an overview of your experience to us *(see end of slides)*
Organization

- Groups of 3-4 students
- Each team will work on a different project, e.g. in cooperation with one of our industry partners or on a topic they have suggested themselves

- Groups are allowed (should) collaborate!
  - exchange your experience with the other groups
  - how do the other groups tackle certain problems?

- Technical aspects:
  - each group will get exclusive access to at least one high-end GPU server with
    - 4x NVIDIA GPU w/ 11GB RAM
    - 10-core CPU
    - 256 GB RAM
  - scale up your models and data!
Organization

- Weekly meetings (around 90-120 minutes)
  - each group should briefly report their progress, open problems, and next steps
- Regular documentation of your work
  - status reports and documentation (we might set up a wiki)
  - use of a central code repository
Grading

- The grade is based on the **whole semester’s performance**!
  - regular completion of **documentation**
  - **regular presentations/discussions** during semester
  - **final presentation** at the end of the semester
    - overview about what you have done, how did you implement it, what are the results, what went wrong, discussion of the framework, ...
    - each member of the team needs to present some parts
Content

- Techniques we might want to look at (if you know these, that's good!)
  - Optimization (e.g. via gradients)
  - Stochastic optimization
  - Neural networks
  - Learning with non-i.i.d. data (e.g. temporal data)

- Tasks:
  - preprocessing
  - classification
  - profiling
  - clustering/topic mining
  - recommendation
  - anomaly detection
  - ...
Projects

There are three types of projects in this lab course:

- Academic projects
- Industry projects
- Your own projects
Reproduction and improvement of a published model

- Can you spot inconsistencies in a recent publication’s experimental setup? Can you even improve their results?

- Students can choose a recent algorithm (e.g. from ICLR 2018), and aim to reproduce and improve the results in the paper.

- Given the computational resources available to the students, they can even select large-scale models and evaluate the validity of the results and claims.

- This can also be a good way to lay the foundation of a new algorithm for a master thesis.
Industry project: Oktoberfest food classification

- Industry partner: **ilass AG**, maker of software for gastronomy and party tents (e.g. Oktoberfest).

- The project will be about detecting and classifying **food items on images** to be extracted from a **video** stream.

- Representative present today: **Peter Vogel**
Industry project: Automatic anonymization of faces

- **Automatic anonymization** of faces in image and video data is important to protect the privacy of people.

- Blurring or completely graying out parts in images where faces are detected means a **loss of information** since all facial features are removed.

- **Goal**: develop a method for **face anonymization** while preserving the **most relevant facial features** to still recognize basic information like emotions.
Industry project: Siemens

- Details to be announced.
Own projects

- You can submit a **brief exposé** of your project idea provided that:
  - There is a considerable challenge from a machine learning perspective, e.g. **non-i.i.d. data** (graphs, temporal data), **very noisy data**, **new application**,
  - You have a sufficiently large and **challenging dataset** at hand (e.g. from an open data platform),
  - The project is suitable for a group of 3-4 students.
Own projects: exposé

- The exposé should contain
  - a brief description of the problem and why it is important,
  - a description of the dataset you plan to use
  - a rough outline of an approach you would like to pursue

- If you are a group of students, **only one** student should fill in the exposé and add the others’ student ID

- Max, 3,000 characters

- Submit via **online form** (see end of slides)
Registration via the matching system!

*Module name: Large-Scale Machine Learning (IN2106, IN4192)*

+ fill out the application form (see next slide)
Your Experience

- Fill out our brief online form about your experience until 14.02.2018
  - you can provide us with a list of your experience in data mining/machine learning (courses, projects, ...)
  - please send a short overview only (bullet list); not a complete CV
  - (optional) attach a brief exposé of your own project idea.

- Check ml-lab.in.tum.de for a link to the form.