Project outlines and schedule

RClass, IMDEDU, StAMA, MC-Sys

Prof. Dr. Björn Rüffer Dr. habil. Michael Schönlein
Dr. Nataliia Gorban

Winter semester 2024/2025
Bauhaus-Universität Weimar
Chair of Applied Mathematics

This document defines the topics of the special projects (NHRE), respectively, projects (DE, CS4DM). It outlines the expectations towards successful project completions, the timeline of events leading to their completion, and it also provides some basic resources that will aid with the research, programming, and report writing.

1 Project descriptions

In this semester the Chair of Applied Mathematics hosts the student projects outlined below. Each project outline contains an executive summary, defines an overall objective for project completion and the tasks that are required to reach this objective.

1.1 RClass—Classification by Rational Approximation

Executive summary

Classification—e.g., of pixel based pictures of images as can be found in the benchmark database MNIST [LCB]—can be addressed by various methods, including support vector machines, neural networks, logistic regression. This project will utilise rational approximation instead of these more established methods.

Keeping with the example, the aim of classification is, in rough terms, to find a function (the classifier) that will return the digit out of $\{0, 1, \dots, 9\}$ that is seen in the pixel image (which consists of $28 \times 28 = 784$ black and white pixels). The classifier can be thought of as a function in various different ways, e.g., as a function $f\colon \mathbb{R}^{784} \to \mathbb{R}$, or as a function $f\colon \{0,1\}^{28\times 28} \to \{0,1\}^{10}$, to just name two possibilities. Classification involves a training data set to find a function candidate that works with the training data. One then hopes that the function will also work with arbitrary data.

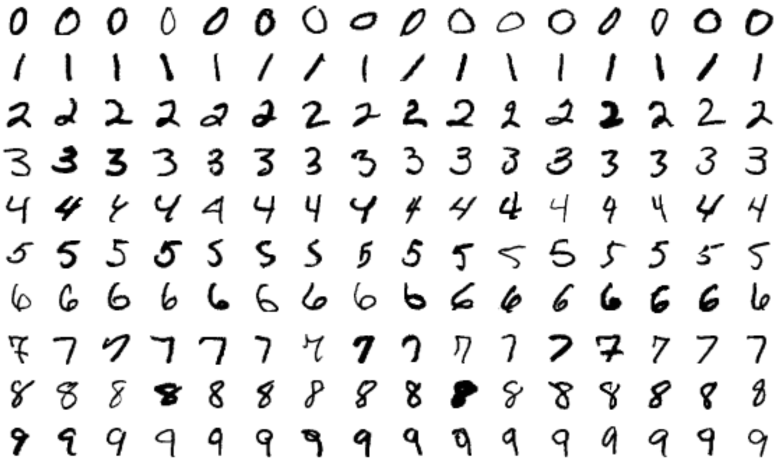

Figure 1: Example of digits encoded in the MNIST data set.

Rational approximation of one dimensional data involves fitting a *rational function*

$$r(x) := \frac{\sum n_k x^k}{\sum d_k x^k}$$

that depends on a range of parameters (the n_k and d_k in the numerator and denominator, respectively) to known pairs of data, $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$, so that the *error*

$$\sum_i \left| y_i - r(x_i) \right|$$

is minimised. This is usually achieved by optimising over the set of possible parameters.

Objective

Implement a rational classifier for MNIST images.

Tasks

- Adapt one dimensional rational classification techniques to higher dimensions
- Solve the corresponding optimisation problems
- Apply your results to the MNIST dataset and discuss the results

1.2 IMDEDU—Image Deduplication

Executive summary

Scattered on an inherited collection of hard-drives there is a substantial number of old family photos. The original owner of the photos liked to send modified versions of the photos via email to his friends and family, and—this was back in the day—he would often crop or downscale images for email transmission. Sometimes he would add fancy frames. Unfortunately, he kept all versions. To order and archive this collection, it is desirable to not keep all versions of every image, but only "the best version" of each.

Objective

Devise an algorithm (ideally in python) that groups pictures that are essentially showing the same content into groups.

Tasks

- Implement a way to gather a collection of images from a set of hard-drives, such as to automatically eliminate exact duplicates and, at the same time, keeping meta information such as original folder locations.
- Investigate ways to visually "fingerprint" an image, so that similar images result in similar finger prints. Use these fingerprints to define a "visual distance" between two images.
- Implement clustering techniques to group images that are visually close to each other together.

1.3 StAMA—Statistical Analysis in the Modern Academia

Executive summary

The academic world increasingly faces challenges related to publication misconduct [Cat24; Cha24]. Many ethically questionable practices are related to predatory journals and predatory publishers, as these do not implement and enforce sound quality control measures. A list of these "bad players", known as Beall's List [Bea], can help researchers to avoid publishing in outlets of questionable quality. The list is not free from criticism, though, and it is curated by humans. This list may also not be complete. This project aims to use statistics and machine learning techniques to automatically separate the good from the bad journals.

Objective

Devise and implement an algorithm that will identify and separate "good" and "bad" academic journals.

Tasks

- Collect citation histograms for a variety of journals, including a range of journals from Beall's List.

- Apply clustering techniques in the space of the histograms.
- Use statistical analysis to investigate whether the clusters can reveal unethical practices, e.g., by grouping predatory journals together.

1.4 MC-Sys—Modelling Complex Systems

Executive summary

The MC-Sys project is concerned with different modelling techniques to describe dynamic processes. Here we consider two populations of animals, the Canadian lynx and snow rabbits, that live in the same area in Canada. For their interaction is observed that there is a cyclic behaviour that is repeated every 9-10 years.

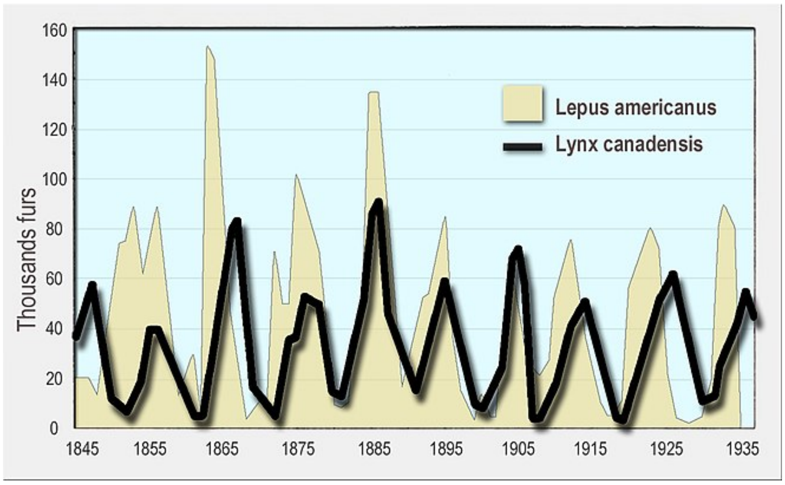

Figure 2: The evolution of the population sizes of the Canadian lynx and the snow rabbits (Source: Wikipedia)

Objective

Devise and implement different modelling techniques to describe complex systems consisting of several populations.

Tasks

- Develop different kinds of models to describe complex systems based on ordinary differential equations and agent-based modelling
- Identify the parameters so that cyclic behaviour of 9-10 years can be observed
- Extend the models to include foxes or area conditions, like creeks, mountains etc.

2 General expectations

The individual workload for a project worth 12 ECTS points is 360 hours, spread across meetings, individual and group work, preparation and presentation of reports and outcomes.

The project is the responsibility of every group member. Every student will receive an individual grade at the end.

The grade will depend on the report (weighted at 75%) and the final presentation (weighted at 25%). The report is expected to be produced by the group, but needs to contain a statement that clarifies the individual contributions and responsibilities. Similarly, every group member must contribute to the final presentation.

A group member can withdraw from the project until the end of the fourth week in the teaching period without penalty. Later withdrawal will result in a fail grade. This has the same consequences as failing an exam, e.g., there is only a finite number of attempts allowed.

The due date for the report is the day of the presentations. The day of the presentations is usually the last day of the exam period (about 4 weeks after the end of the teaching period). The date is to be fixed in the kick-off meeting.

Each group should meet at least weekly, in addition to meetings with one of the supervisors. The group should *always* bring something to show to the meeting with the supervisor and always report on the progress with respect to the project goal.

3 Resources

Here are a few pointers that may help with the projects.

- Overleaf (a website for a paid service; the basic version is free and absolutely sufficient for this project) for typesetting your report in LaTeX(your may also use a local TeX installation, which is faster and free):
  `https://www.overleaf.com/`

- A LaTeX template for reports/theses:
  `https://www.overleaf.com/read/knvycxrmhzgh#5881b1`

- A LaTeX template for slides:
  `https://www.overleaf.com/read/kvbmcdvyzvrt#c58127`

- A seminar course on academic writing: Search for "Writing for science and engineering" in BISON.

- Two highly recommended *short* books on (academic) writing: [SW99; SD24]

- SageMath for mathematics-intense programming, plotting, and analysis in Python:
  `https://www.sagemath.org/`

- Useful Python libraries for data handling and statistics

pybliometrics: Python-based API-Wrapper to access Scopus
`https://pybliometrics.readthedocs.io/en/stable/`

pandas—handle data in spreadsheets
`https://pandas.pydata.org/docs/index.html`

seaborn: statistical data visualization
`https://seaborn.pydata.org/`

colorcet: a collection of perceptually accurate colormaps
`https://colorcet.holoviz.org/`

openCV: computer vision
`https://pypi.org/project/opencv-python/#description`

4 Schedule

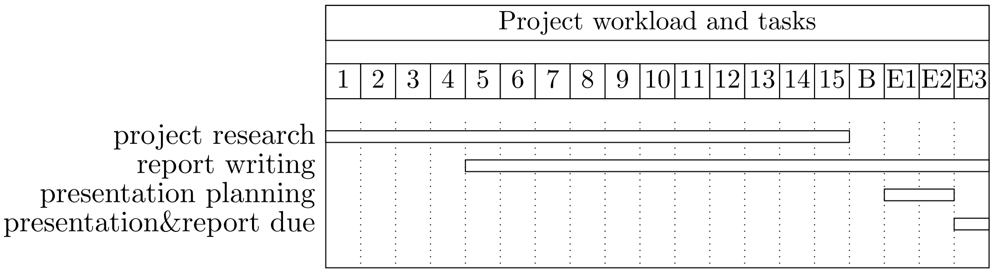

Figure 3: General project schedule showing semester weeks (make your own schedule specific to your project).

There are 360 hours of workload in a project per participant. With 15 weeks of—on average—20 h/week for project work and meetings, this leaves 30h for the preparation and delivery of the final presentation, and 30h for the final polishing of the report (you should aim to write most of it along the way). The chart shows weeks 1–15 of the semester teaching period, the one week break (B) following the teaching period but before the exam, and three weeks (E1–E3) of exams. The report and the presentation are due in the third week of exams.

References

[Bea] Jeffrey Beall. *Beall's list of potential predatory journals and publishers.* URL: `https://beallslist.net`.

[Cat24] Michele Catanzaro. "Citation manipulation found to be rife in math". In: *Science* 383.6682 (2024), pp. 470–470.

[Cha24] Dalmeet Singh Chawla. "The citation black market: schemes selling fake references alarm scientists". In: *Nature* 632.8027 (2024), pp. 966–966.

[LCB] Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. *THE MNIST DATABASE of handwritten digits*. URL: `https://yann.lecun.com/exdb/mnist/index.html`.

[SD24] Kay Smarsly and Kosmas Dragos. *Scientific Writing in Engineering*. 2nd edition. Tredition, 2024.

[SW99] William Strunk Jr. and E. B. White. *The Elements of Style*. 4th edition. Macmillan, 1999.