Advanced Data Mining

Piotr Lipiński, Klaudia Balcer, Mikołaj Słupiński, Maria Szlasa

Projects

- **GOAL:** solving a practical (real-world?) problem
  using some knowledge from our lecture or its extensions

- **KEY POINT:** creativity in data analysis
  (invention, own interpretation of the problem, proper selection of methods, etc.)

- Projects should be carried out under the supervision of the lecturer, in a group of two students (or individually in special cases)

- The final result of the project should include:
  - **original data** (a link to data, if publicly avaiable)
  - **intermediate processed data** (if obtaining them is time-consuming)
  - **software** (scripts, programs, tools)
  - **final report** (concise and short, but comprehensive)

Projects

- The report should include:
  - the description of the data and the problem definition
  - the detailed description of the methods used
    (if they are not the regular methods presented on the lecture)
  - the detailed description of the implementation of the methods used
    (if it is not obvious)
  - a detailed description of the results obtained
  - final conclusions, summary, development prospects

- The project will be assessed on a scale 0 - 30 points:
  - selection of methods and tools (5 points)
  - implementation effectiveness (5 points)
  - final and partial results obtained (5 points)
  - "insight into data analysis" (5 points)
  - final report (5 points)
  - overall solution to the problem (5 points)

Projects

- Deadlines:
  - declaration of the topic and the group composition – May 23 (email)
  - presentation of the approach to the problem – May 30 (in person)
  - discussion of the partial results – June 13 (all the time)
  - final presentation – the end of the semester (in person)

  - discussions, questions, other issues – all the time, in person, online or by email

Topics

- **Learning from Temporal Data:**
  - Learning representations for stock market data on LSE ROB (Piotr)

    Martin D. Gould, Julius Bonart, „Queue Imbalance as a One-Tick-Ahead Price Predictor in a Limit Order Book", arXiv:1512.03492
  - Cyclicity, periodicity, seasonality in time series representations for environmental engineering data (Piotr)
  - Generative AI for missing values imputation in time series (Piotr)
  - Anomaly detection in time series, based (not only) on learning representations (Piotr)
  - Interpolation of geochemical time series (Mikołaj)
  - TS2Vec and other SSL methods for time series changepoint detection (Mikołaj, Piotr)
  - Matrix Profile - universal time series representation? (Mikołaj)
  - What if? - counterfactuals (Mikołaj)
- **Advanced Recommender Systems:**
  - Temporal Graphs for recommendations (Piotr)
  - Gated Graph Neural Networks (not only) for recommendations (Piotr)
  - RecSys Challenge 2025 (Klaudia)

Topics

- **Learning from Geospatial Data:**

  (requires some self-learning – lectures on this topics will start on May 30)

  - MAE for missing data in satellite imagery – how much can we recover? (Maria)

    Investigate image reconstruction using Masked Autoencoders. Propose and implement different types of masks (e.g., block masks, random shapes, spectral-band drops). For each strategy, test which features are hardest to reconstruct—e.g. clouds, shadows, particular spectral bands, forests, rivers, etc. Determine what fraction of pixels or bands can be hidden while still achieving high-quality recovery.

    https://www.kaggle.com/datasets/apollo2506/eurosat-dataset?select=EuroSAT

  - Crop Mapping (Maria)

    Compare the performance of two architectures (for example, U-Net, Swin-U-Net, or Vision Transformer) on the task of segmenting agricultural fields. Assess which model excels under different conditions—e.g. varying crop types.

    https://www.kaggle.com/datasets/ignazio/sentinel2-crop-mapping?select=lombardia

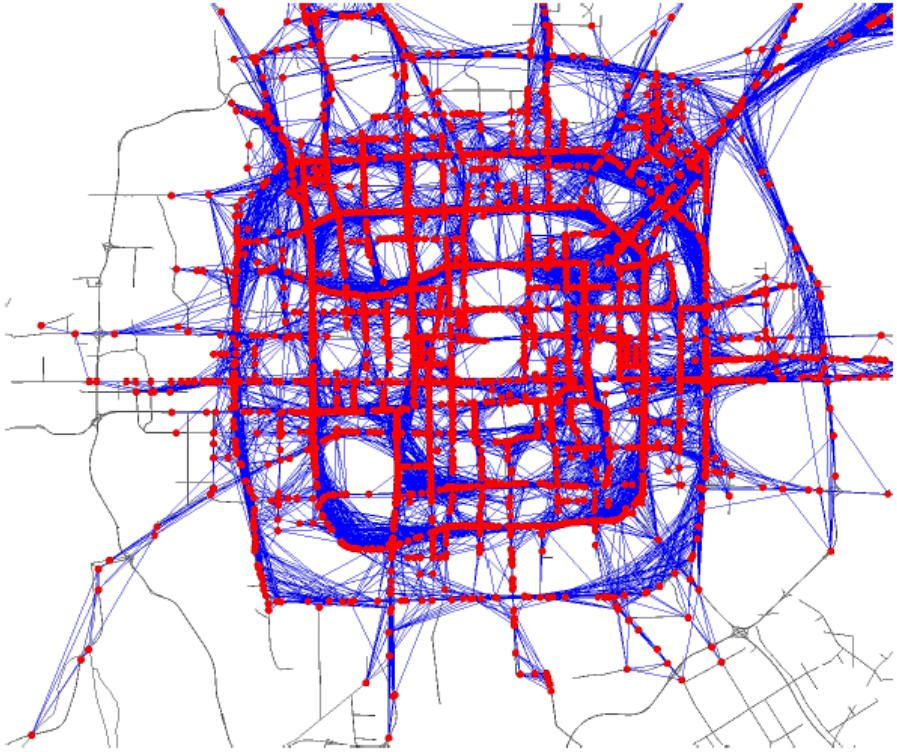
  - TRACLUS (Maria)

    Implement the TRACLUS algorithm for clustering trajectory data. Design a range of artificial trajectory sets—straight lines, curvilinear paths, noisy tracks—to rigorously test your implementation.

  - Improving GNSS with Kalman filters and other methods (Piotr, Maria, and others)

Topics

- Trajectory Data Mining



Source: Yuan, Jing, Zheng, Yu, Zhang, Chengyang, Xie, Wenlei, Xie, Xing, Sun, Guangzhong, Huang, Yan, „T-Drive: Driving Directions Based on Taxi Trajectories", [in] Proceedings of 18th ACM SIGSPATIAL Conference on Advances in Geographical Information Systems, ACM, 2010.