Machine Learning Applications in Process Mining

Kickoff meeting – 2023-04-04

Chair of Process and Data Science

RWTH AACHEN UNIVERSITY

The instructors

Marco Pegoraro

http://mpegoraro.net/
pegoraro@pads.rwth-aachen.de
🐦 @pegoraro_marco

Machine Learning applications in Process Mining

The papers in this seminar combine **machine learning** with **process mining**.

Goal: applying predictive models (supervised and unsupervised) to process mining to solve a set of problems (mainly predictive problems)

Machine learning?

How far should we go?

$$\gamma_n = \frac{|(\mathbf{x}_n - \mathbf{x}_{n-1})^T [\nabla F(\mathbf{x}_n) - \nabla F(\mathbf{x}_{n-1})]|}{||\nabla F(\mathbf{x}_n) - \nabla F(\mathbf{x}_{n-1})||^2}$$

$$\tfrac{1}{2} K_{ijk} + J_{j,ik} = \mathrm{E}_X \left[\frac{1}{2} \frac{\partial^3 \ln f_{\theta_0}(X_t)}{\partial\theta_i\,\partial\theta_j\,\partial\theta_k} + \frac{\partial \ln f_{\theta_0}(X_t)}{\partial\theta_j}\, \frac{\partial^2 \ln f_{\theta_0}(X_t)}{\partial\theta_i\,\partial\theta_k} \right]$$

$$\begin{aligned}\rho(\boldsymbol\beta, \sigma^2 | \mathbf{y}, \mathbf{X}) &\propto \rho(\mathbf{y}|\mathbf{X}, \boldsymbol\beta, \sigma^2)\rho(\boldsymbol\beta|\sigma^2)\rho(\sigma^2) \\ &\propto (\sigma^2)^{-\frac{n}{2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol\beta)^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol\beta)\right)(\sigma^2)^{-\frac{k}{2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol\beta - \boldsymbol\mu_0)^{\mathrm{T}}\boldsymbol\Lambda_0(\boldsymbol\beta - \boldsymbol\mu_0)\right)(\sigma^2)^{-(a_0+1)} \exp\left(-\frac{b_0}{\sigma^2}\right)\end{aligned}$$

$$\begin{aligned}\text{maximize } f(c_1 \ldots c_n) &= \sum_{i=1}^n c_i - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n y_i c_i (\varphi(\vec{x}_i) \cdot \varphi(\vec{x}_j)) y_j c_j \\ &= \sum_{i=1}^n c_i - \frac{1}{2}\sum_{i=1}^n \sum_{j=1}^n y_i c_i k(\vec{x}_i, \vec{x}_j) y_j c_j\end{aligned}$$

$$\text{subject to } \sum_{i-1}^n c_i y_i = 0, \text{ and } 0 \le c_i \le \frac{1}{2n\lambda} \text{ for all } i.$$

$$\mathcal{P}(s|d) \propto e^{-\mathcal{H}(d,s)} \propto e^{-\frac{1}{2}(s-m)^\dagger D^{-1}(s-m)} \propto \mathcal{G}(s-m, D)$$

MLAPM seminar

This is not a math course: the basics should be clear, but the focus is on **application**, rather than **theory**

Focus on a **running example!**
It can be the one on the paper for the outline, but you should design your own for the term paper

In a presentation, often **intuitiveness** is preferable to **formality**

MLAPM seminar

Another focus of the seminar is the **replication of the experimental results** contained in the paper

You will need to find/integrate the code of the approach and execute it

Ideally, your experiments should include the same datasets (if available), but also **new data**

Note that a **good level of independence** is expected from you!

Goals

Each one of you has to:

- Produce a **term paper outline**
- Replicate **experiments**
- Produce a **term paper**
- Produce a **deck of slides**
- Hold a **presentation**
- Participate in **all the presentations**

Outline

A total of ~4 pages, to be written in LaTex and in English (Springer 1-column format, you will receive the template)

- Presentation of the term paper: ~1 page
  - Title, abstract and/or short overview, term paper structure
- Example of application (high level): ~2-3 pages
- Things that should be clear:
  - Goal of the paper
  - Preliminary knowledge
  - Important definitions
  - Conclusions and limitations

Term Paper

A total of ~20 pages (including figures and references)
Same template of the outline

- Structure
  - Title page: Topic title, author, seminar title, date of presentation
  - Abstract/short introduction, „Assessment of (title, authors, venue)“
  - Main section
    - examination of your topic, describe in your own words
    - your own example/application of the algorithm/method
    - replication of the experiments
    - comparison with related research
  - Short summary
  - References

- Assume **your classmates** are the **target readers**

Term Paper

Note that the term paper submission should be accompanied by the **sources for the experiments**

These should be accompanied by the **dataset files** and readily executable

You will also need to write a **quickstart** and document **dependencies** (if applicable)

Any «open» format is good (es. Jupyter Notebook)

Slides

- No specific template
- Topic, context, motivation
- How does it work?
- How do you assess it?
- Were the experiments and their results replicable?
- Convince your audience that **it matters** and **it actually works** (but of course, limitations of the approach should also be clear)

Slides

- Usual advices for the slides:
  - Model the slides on the presentation, not the other way around
  - Sparse text (possibly in bullet points)
  - Use formulas sparingly (only if they really clarify)
  - Pictures/diagrams/graphs help a lot
  - Clear layout, colours, images (at a good resolution)
  - Readable and respectable sans-serif font (no Comic Sans)
  - Depending on your experience, try to present and **time yourself**

Presentation

- Roughly 30 minutes
  - Of which ~10 reserved for Q&A

- You need to manage time effectively!

- **All students need to attend all talks to pass the seminar**

- Other members of our research team might attend

Deadlines

- **Outline: 28 April, 23:59**
- **Term paper: 7 July, 23:59**
- **Slides: 19 July, 23:59**
- **Presentation: 21 July (schedule TBD)**

The deadlines are strict! No exceptions possible!

Missing a deadline implies immediate dismissal!

Deregistration and grading

Notice that **multiple submissions are possible and encouraged** (**last one counts**)

Deadline to **unregister** from the seminar:
28 April, 23:59

After the 28th of April, participants will be registered

Grading: **50% term paper, 50% presentation** (the outline is not graded)

Meetings

Meetings:

- 30 to 40 minutes
- to discuss the papers and additional literature
- to discuss how to write outline/term paper/slides

Meetings are **for you to get feedback and advice**, not for us to monitor you!

Meetings (proposed dates)

- **First: 21 April**
- **Second: 9 May**
- **Third: 30 June**
- **Fourth: 14 July**

Can be flexible
All time schedules TBD (you will get invitations)

FAQ

FAQ: do I need to implement the approach in the paper?

Yes. All papers involved have a corresponding code repository, but you will need to debug/fix/integrate.

Many of you will probably want to extend the approach and add your own ideas (that’s a bonus).

Again, ideally experiments should be replicated both on the same and new data, and with the same and new settings.

FAQ: can I (re)use content from published papers in my term paper?

You can reuse graphics and formulas (with attribution).
You should not reuse text from your assigned paper.
You may add direct citations of text from other papers (with attribution).

FAQ: do I need to actually have all the meetings?

Throughout the seminar, at least 3 of the 4 scheduled meetings need to take place. If you cannot attend, we can reschedule, but at least 3 meetings need to take place.

FAQ: how do we send the deliverables?

Given the number of people involved, submission will be via email.

Again, multiple submissions are possible and encouraged – most recent before the deadline counts.