AI Programming WS25 - Group Project

You work in groups of two students.
Register with your project partner on TUWEL.

You have to choose which type of project you want to do:

1. Applying probabilistic programming on a real world data set.
2. Implementing an advanced posterior inference algorithm.

As per the ECTS breakdown of this course, each one of you should invest about 50 hours into this project.

You complete the project by uploading a project report to TUWEL by January 9th.
Alongside the report, upload the source code of your project as a .zip file.
The report should be a summary of your project and 3-5 pages long.
Use this template: https://www.overleaf.com/1335997378dvpjzdzbzkdx#99faa1

In the oral exam, we will ask questions from a catalogue (made available after the lectures) and we will discuss your project. The report is part of this discussion, but will not be marked separately.

More details on the report and exam follow below.

If you are unsure about your project topic or scope, or if need help for a particular problem, you may reach out to Markus Böck.

1 Applying probabilistic programming on a real world data set.

You pick a real-world dataset and formulate a question that you try to answer with Bayesian inference via probabilistic programming. You may find datasets, take inspiration, or adapt models from the following resources:

- https://www.kaggle.com/datasets
- https://www.pymc.io/projects/examples/en/latest/gallery.html
- https://mc-stan.org/learn-stan/case-studies.html
- https://pyro.ai/examples/

Simple copy-paste is not allowed and will be marked with 0 points.

In the project, we want you to:

- Explore the dataset and come up with an interesting question
- Design a probabilistic model to answer this question and implement it in a PPL of your choice (recommended: PyMC, Stan, Turing.jl, Pyro, NumPyro).
- Choose and apply an inference algorithm to approximate the posterior
- Interpret the results

In the report, write a section for each of the following points:

- Describe the dataset and the question you formulated for this dataset
- Describe the probabilistic model to answer the question. What are the latent variables? What are the observed variables? How did you model the generative process from latents to observes? How did you pick the priors?
- Describe your chosen inference algorithm and justify your choice. If you had to customise/tune the inference algorithm, describe how. Analyse the inference results: was it successful or did it fail?
- Interpret the inference results with respect to your question. In particular, analyse the uncertainty of your conclusions.

In the oral exam, you have to demonstrate a solid grasp of probabilistic programming and Bayesian inference principles applied in the project. You should have a deep understanding of the dataset, your model, and the chosen inference algorithm. You should be able to competently answer questions about your project, be able to justify your modelling and inference algorithm decisions, and be able to interpret the results.

2 Implementing an advanced posterior inference algorithm.

You implement and evaluate an inference algorithm that we have not implemented in the assignments. You are free to choose an inference algorithm, below we give a few ideas.

- NUTS https://arxiv.org/abs/1111.4246
- Sequential Monte Carlo (SMC) https://arxiv.org/pdf/1903.04797
- Exact methods for discrete models like Variable Elimination / Belief Propagation / Junction Tree algorithm
- MH with programs as proposals / ADVI with programs as guides
- Exact symbolic inference alá https://github.com/eth-sri/psi
- This book is also a good resource: https://arxiv.org/abs/1809.10756

In the project, we want you to:

- Implement your chosen inference algorithm. You may build on the minimal PPL of the lecture, any existing PPL, or build your own system (some algorithms are easier to implement if you have full control of the system and data structures). Markus Böck can also help you scope your implementation and system.
- Come up with a benchmark set containing models that demonstrate the advantages and limitations of your algorithm.
- Evaluate your algorithm on the benchmark set against at least one baseline algorithm.

In the report, write a section for each of the following points:

- Describe the algorithm: How does it work? What assumptions on the model does it require? What are its advantages and limitations?

- Describe your implementation: How is the system designed to enable the application of the inference algorithm to probabilistic programs? How did you implement the algorithm in your system? Focus on the interesting parts of your implementation rather than small details.
- Describe your benchmark set and justify your choice of models.
- Report your evaluation results.

In the oral exam, you have to demonstrate a deep understanding of your implemented inference algorithm. You should be able to competently discuss your implementation, answer technical questions about it, and justify implementation and benchmark decisions. You should be able to discuss and interpret the evaluation results.