# 02460 Advanced Machine Learning LOGBOOK

Laura Perge, [s182257@student.dtu.dk](mailto:s182257@student.dtu.dk)

Daniel Horvath, [s172185@student.dtu.dk](mailto:s172185@student.dtu.dk)

Lorant Gulyas, [s172188@student.dtu.dk](mailto:s172188@student.dtu.dk)

The main purpose of the logbook is that it serves as a tool for you to organize the project. Further, it serves as a way to collecting information related to the learning objectives:

* Presentation of methods and results at meetings with project supervisor and fellow students
* Plan and carry out the course of the project in collaboration with the project supervisor
* Organize and coordinate the work in the project group

**Overall Project Goals**

**Define own learning objectives for the project**

* Derivation of EM for KDE with hold-out and leave-one-out CV - both math and coding
* Speed-up code: Replace Python’s default Multivariate Gaussian PDF - track implementation cost, visually track convergence
* Replace EM algorithm with VB to account for uncertainty in Sigma’s estimation - marginalize Sigma’s distribution
* Apply the newly derived method to impute randomly missing values in a dataset

**Carry out a well-founded delimitation of the project and formulate specific hypotheses and aims**

*Describe*

**Project Meetings**

**Week 6: 14.03.2019-17.03.2019**

**Questions**: How the structure of the algorithm should look like, EM algorithm outside and CV folds inside or the other way around? Why do we have to use a unified Sigma for every datapoint and how will we adjust this Sigma at every iteration? How this will affect the visualizations?

**Reading:** Christopher Bishop: Pattern Recognition and Machine Learning (Chapter 9, 10.2, 10.2); Tue Herlau, Mikkel N. Schmidt and Morten Mørup: Introduction to Machine Learning and Data Mining (Chapter 19). *Everybody on their own without meeting together.*

**Discussions:** Gathering our thoughts right after the supervisor meeting, summarizing our written notes into a nicer format. Clarifying our ideas how should the results look like on a 2D dataset (i.e.: faithful). Creating an implementation sketch to achieve the goals for the next supervisor meeting. *Everybody was there in person, we helped each other by spearing to make things clear.*

**Implementation:** Derivation of the equations discussed in the supervisor meeting on paper. Implementing the results in python for the hold-out method *(Laura and Daniel).*

Going through the derivation, implementing the results for hold-out CV and k-fold CV *(Lorant).*

**Results:** The log-likelihood is converging, but sometimes the code fails with singular matrix error.

**Week 7: 18.03.2019 – 24.03.2019**

**Questions:** A frequent but not always happening runtime error about singular matrices.

**Implementation**: Clearing up the code, merging different solutions and best practices *(Laura).* Perform E and M steps in separate functions, until convergence is reached for hold-out method *(Daniel)*. Create a function which plots the results in a color coded 2D graph *(Lorant).*

**Discussions**: Adjusting implementation plans to create CV folds first, do E step on all folds, do M step on all folds and repeat the last two steps until convergence.

**Results:** Visualization of results works, the algorithm is slow.

**Week 8: 25.03.2019 – 31.03.2019**

**Questions**: Why is there a difference between our custom-made multivariate function and scipy’s multivariate\_normal?

**Implementation**: Custom made multivariate function to decrease computational time *(Laura)*. Code debugging to figure out the source of the difference *(Laura, Daniel and Lorant).*

**Discussion**: Discussing further steps on how the algorithm could be speeded up, clearing up notes *(Laura and Lorant).*

**Reading**: How Choleski decomposition works and how can we apply this in our case *(Lorant)*.

**Week 9: 01.04.2019 – 07.04.2019**

Questions: The source of the difference is still unknown.

**Supervisor Meetings**

**Week 6: 14.03.2019**

*General discussion about the project*

*Gathering application areas*

*Guidelines for next week (implement EM algorithm for KDE with holdout CV first, and then apply it for k-fold CV)*

**Week 7: 21.03.2019**

*Presentation of results since last meeting and discussing common mistakes*

*Hints for making the code faster and more efficient*

*Action points for next meeting (implement a resource efficient algorithm, perform cost tracking on the finished algorithm)*

**Week 10: 11.04.2019**