The AMIDST toolbox: a Java library for scalable probabilistic machine learning

Andrés R. Masegosa\(^1\) and Ana M. Martínez\(^2\) and Darío Ramos-López\(^3\) and Thomas D. Nielsen\(^4\) and Helge Langseth\(^5\) and Antonio Salmerón\(^6\) and Anders L. Madsen\(^7\)

AMIDST is a flexible Java library for probabilistic machine learning, which provides tailored parallel and distributed implementations of Bayesian parameter learning (and probabilistic inference) for batch and streaming data. This processing is based on flexible and scalable message passing algorithms \([11]\). AMIDST handles probabilistic graphical models with latent variables and temporal dependencies \([3]\) which can be trained on large-scale data (making use of Apache Spark \(^8\) and Apache Flink \(^9\)) and provides interfaces to a number of other platforms like HUGIN, MOA, Weka and R.

In this demonstration we will show some of the main AMIDST functionalities. During the demo, the construction of customized models, possibly with latent variables and temporal dependencies, will be explained. Here is a sketch of the demo:

- First, we will define the structure of a probabilistic graphical model, by showing how to build a graph encoding the dependencies between the observed and latent variables. Alternatively, various standard models are available to use in AMIDST (Gaussian discriminative analysis, Gaussian mixtures, factor analyzer, etc). Once the structure is determined, the parameters of the model will be fit from local data using multi-core learning algorithms. Afterwards, some probabilistic queries are performed with scalable algorithms.
- Then, this example is extended to use distributed data over Flink (with only minor changes in the code), showing the flexibility of the toolbox. Scalable and distributed learning and inference algorithms provided by AMIDST will be used running on top of Flink.
- Finally, probabilistic graphical models with temporal dependencies are considered. AMIDST provides with several latent-variable dynamic models (Hidden Markov model (HMM), Factorial HMM, Kalman filter (KF), Switching KF, etc), or alternatively, customized models can be defined. Again, learning and inference algorithms are used with both local and distributed data over Flink, showing the scalability of the AMIDST algorithms.

The AMIDST toolbox is supported by a considerable number of scientific papers, both with methodological developments \([1, 2, 4, 5, 7, 8]\) and with real industrial applications \([1, 2, 6, 9, 10]\).

More information on AMIDST

The AMIDST toolbox has been developed within the AMIDST project (Analysis of MassIve Data STreams) of the European Union’s Seventh Framework Programme, under grant agreement no 619209. See more information on the AMIDST toolbox on these sites:

- The AMIDST toolbox website: https://amidst.github.io/toolbox/
- AMIDSTs Github site (demo examples in the repository ‘tutorial’): https://github.com/amidst
- AMIDST Toolbox YouTube channel, with some introductory videos: https://www.youtube.com/channel/UCBdU7xvRCVZj-c9z78n2meQ

REFERENCES


---

1. Norwegian University of Science and Technology, email: andersrn@idi.ntnu.no
2. Aalborg University, email: ana@cs.aau.dk
3. University of Almería, email: dramoslopez@ual.es
4. Aalborg University, email: tdn@cs.aau.dk
5. Norwegian University of Science and Technology, email: helgel@idi.ntnu.no
6. University of Almería, email: antonio.salmeron@ual.es
7. HUGIN Expert A/S and Aalborg University, email: anders@hugin.com


