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.

[1] Norwegian University of Science and Technology, email: andresrm@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
[8] http://spark.apache.org/
[9] http://flink.apache.org/

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] Hanen Borchani, Ana M. Martínez, Andrés Masegosa, Helge Langseth, Thomas D. Nielsen, Antonio Salmerón, Antonio Fernández, Anders L. Madsen, and Ramón Sáez. Dynamic Bayesian modeling for risk prediction in credit operations. In *Proc. of SCAI*, 2015.

[2] Hanen Borchani, Ana M. Martínez, Andrés Masegosa, Helge Langseth, Thomas D. Nielsen, Antonio Salmerón, Antonio Fernández, Anders L. Madsen, and Ramón Sáez. Modeling concept drift: A probabilistic graphical model based approach. In *Proc. of The Fourteenth Int. Symposium on IDA*, pages 72–83, 2015.

[3] D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

[4] A. L. Madsen, F. Jensen, A. Salmerón, H. Langseth, and T. D. Nielsen. Parallelisation of the PC Algorithm. In *CAEPIA 2015: Lecture Notes in Artificial Intelligence (LNAI)*, volume 9422, pages 14–24. Springer, 2015.

[5] Anders L. Madsen, Frank Jensen, Antonio Salmerón, Martin Karlsen, Helge Langseth, and Thomas D. Nielsen. A New Method for Vertical Parallelisation of TAN Learning Based on Balanced Incomplete Block Designs. In Linda C. van der Gaag and Ad J. Feelders, editors, *Probabilistic Graphical Models*, volume 8754 of *Lecture Notes in Computer Science*, pages 302–317. Springer International Publishing, 2014.

[6] A. Masegosa, A.M. Martínez, H. Borchani, D. Ramos-López, T.D. Nielsen, H. Langseth, A. Salmerón, and A.L. Madsen. AMIDST: Analysis of massive data streams. In *Proceedings of the 27th Benelux Conference on Artificial Intelligence (BNAIC 2015)*, 2015.

[7] Andres R. Masegosa, Ana M. Martinez, and Hanen Borchani. Probabilistic Graphical Models on Multi-Core CPUs Using Java 8. *IEEE Computational Intelligence Magazine*, 11(2):41–54, May 2016.

[8] A. Salmerón, D. Ramos-López, H. Borchani, A. R. Masegosa, A. Fernández, H. Langseth, A. L. Madsen, and T. D. Nielsen. Parallel importance sampling in conditional linear Gaussian networks. *CAEPIA'2015. Lecture Notes in Artificial Intelligence*, 9422:36–46, 2015.

[9] G. Weidl, A. L. Madsen, Kasper, and G. Breuel. Optimizing Bayesian Networks for Recognition of Driving Maneuvers to Meet the Automotive Requirements. In *IEEE Multi-conference on Systems and Control*, pages 1–6, 2014.

[10] G. Weidl, A. L. Madsen, V. Tereshchenko, D. Kasper, and G. Breuel. Early Recognition of Maneuvers in Highway Traffic. In *ECSQARU: Lecture Notes in Artificial Intelligence (LNAI), Springer*, pages 529–540, 2015.

[11] J.M. Winn and C.M. Bishop. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 2005.