Post-doc: From missing values to deep learning on sets

Research theme: Machine learning
Keywords: Neural networks, deep learning, dirty data, relational data, missing values
Duration: 24 months
Research teams: Soda (INRIA Saclay)
Advisers: Marine Le Morvan, Gaël Varoquaux
Contact: marine.le-morvan@inria.fr, gael.varoquaux@inria.fr

Context

Most machine learning models expect to receive tables with samples in rows and attributes in columns as input. In reality, the vast majority of structured data is stored in relational databases, where information is scattered across multiple tables. For example, a cancer patient can have information scattered in radiotherapy, chemotherapy and surgery tables depending on his treatment. This data representation, although natural, creates variable-sized inputs, with heterogeneous subsets of attributes. This poses difficulties for machine learning, as a majority of models require fixed-size vector representations. An alternative data representation can be obtained by operating a join on the tables of interest, creating a single table with [Not Applicable] missing values. This project aims at *developing appropriate and theoretically grounded neural network architectures for [Not Applicable] and relational data.*

To date, there is no standard practice for learning with [Not Applicable] missing values/small relational data. Common practice and a few lines of research address related problems with very different perspectives.

Learning with missing values Learning with missing values relates to a rich statistical literature. In practice, most of the proposed methods rely on the imputation of missing values, i.e completing the data matrix with the most probable values [1]. It reflects the fact that in the literature, it is always assumed that there exists an underlying value for a missing entry. But in many real-world cases, it is not the case. When a patient has no metastasis, it does not make sense to impute its size or mutational status. This type of missing values, which we call [Not Applicable], are very common yet imputation-based methods do not make sense for them.

Inductive Logic Programming (ILP) It is a form of machine learning (ML) based on relational logic (Cropper et al., 2022). Just like standard ML, it tries to find hypotheses that generalize to unseen data. But while standard ML typically represents samples in vector spaces, ILP uses sets of logical rules. And while standard ML learn functions, ILP learns relations. ILP provides a natural framework for handling relational data. However, this approach has been burdened with high computational costs due to its combinatorial nature.

Geometric Deep Learning Up to recently, neural networks architectures expected fixed-length vector representations for the inputs, which is problematic for relational data. However, new architectures, related to geometric deep learning (Bronstein et al., 2021), are concerned with representation learning on non-euclidean domains. Recent successes in this domain include Graph Neural Networks (GNNs) and Deep Sets (Zaheer et al., 2017) for learning on graphs and sets respectively. Relational data are also a kind of non-euclidean data, but neither GNNs nor Deep Sets are readily adapted for them. It remains thus an open challenge to leverage the power of representation learning for relational data.

Proposed work

The goal of this project is to design neural network architectures suited for learning with relational/[Not Applicable] data.

Attention-based architecture One way forward is to cast the problem as one of *learning on sets of key-value pairs*, where keys are attributes names. This idea has appeared sporadically in the missing data literature, notably for imputation purposes. Learning with key-value pairs involves two steps : first encoding keys as well as values in numerical vectors, and then aggregating them all to obtain a fixed-length vector representation of a sample. Since keys can be regarded as symbols, *entity embeddings* (Guo and Berkhahn, 2016) can be used to encode them, either taking pre-trained embeddings or embedding tailored for the task. For aggregation, simple pooling operations can typically be used such as the sum, max or mean. More complex aggregation schemes such as *self-attention* (Vaswani et al., 2017) could also be explored to learn useful concepts by aggregating together the right attributes.

Mathematical formalisation Learning with relational databases has been properly formalized in the relational machine learning literature, by relying on symbolic representations. However, this problem severely lacks a proper formalization in the context of continuous, numerical representations as used today in deep learning. We will therefore try to ground theoretically the architecture that we propose. One possibility to formalize the problem is to establish the desirable properties of an architecture for [Not Applicable] data. For example, DeepSets have been designed to enforce a permutation invariance, while Graph Neural Networks enforce an invariance to graph isomorphisms. In the case of [Not Applicable] data, a first intuition leads to a notion of equivariance. Further intuition could be drawn from how logical rules are used in the ILP literature to represent NAP data.

Team

SODA in an INRIA project team targeting applications in health and social sciences. It currently has 4 permanent researchers, 7 PhD students, 3 engineers, and also hosts the team in charge of developing and maintaining Scikit-Learn. SODA has expertise on the problem of learning with missing values. Previous works include general consistency results for learning with deterministic imputations (Josse et al., 2019; Le Morvan et al., 2021), as well as deep learning architectures (Le Morvan et al., 2020) with an appropriate inductive bias for learning with missing values. The team also has expertise in adapting modern representation learning tools –those behind the deep learning revolution– to relational data. It includes character-level embedding approaches that can make any language model robust to morphological variants (Chen et al., 2022), automatic entity matching (Cvetkov-Iliev et al., 2022), or encoding high cardinality categorical variables (Cerda and Varoquaux, 2022).

Skills

- Knowledge of machine learning or applied maths (mathematical optimization and statistics)
- Familiarity with deep learning frameworks (typically pytorch)
- Programming skills in implementing algorithms and empirical evaluation
- Good paper-writing skills, in English

References

M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *arXiv preprint arXiv:2104.13478*, 2021.

P. Cerda and G. Varoquaux. Encoding high-cardinality string categorical variables. *IEEE Transactions on Knowledge and Data Engineering*, 34(3):1164–1176, 2022.

L. Chen, G. Varoquaux, and F. M. Suchanek. Imputing out-of-vocabulary embeddings with love makes language models robust with little cost, 2022.

A. Cropper, S. Dumančić, R. Evans, and S. H. Muggleton. Inductive logic programming at 30. *Machine Learning*, 111 (1):147–172, 2022.

A. Cvetkov-Iliev, A. Allauzen, and G. Varoquaux. Analytics on non-normalized data sources: More learning, rather than more cleaning. *IEEE Access*, 10:42420–42431, 2022.

C. Guo and F. Berkhahn. Entity embeddings of categorical variables. *arXiv preprint arXiv:1604.06737*, 2016.

J. Josse, N. Prost, E. Scornet, and G. Varoquaux. On the consistency of supervised learning with missing values, 2019.

M. Le Morvan, J. Josse, T. Moreau, E. Scornet, and G. Varoquaux. Neumiss networks: differentiable programming for supervised learning with missing values. *Advances in Neural Information Processing Systems*, 33:5980–5990, 2020.

M. Le Morvan, J. Josse, E. Scornet, and G. Varoquaux. What's a good imputation to predict with missing values? *Advances in Neural Information Processing Systems*, 34, 2021.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

M. Zaheer, S. Kottur, S. Ravanbakhsh, B. Poczos, R. R. Salakhutdinov, and A. J. Smola. Deep sets. *Advances in neural information processing systems*, 30, 2017.