INRIA Saclay, Équipe Parietal http://team.inria.fr/parietal CDS Center for Data Science

Neural networks for semantics in databases

Research theme: machine learning, data science
Keywords: database, semantics, neural networks, dirty data
Duration & salary: 3 to 6 months, between 450 €and 800 € monthly
Research teams: Parietal (INRIA Saclay) and CDS (Université Paris Saclay
Adviser: Gaël Varoquaux
Contact: gael.varoquaux@inria.fr
Application: Interested candidate should send CV and motivation letter

Context: Statistics and machine learning work best on numerical vectors. On the opposite, the typical entry in a database is textual. Categorical data is usually vectorized using variants of one-hot encoding or dummy variables: the occurrence of an entity is written with a binary presence vector. Data with many categories or very rare categories make this approach brittle, as they lead to a large number of features, or nearly-empty feature vectors. Such situation, rare "clean" preprocessed data, happens often on raw data, such as with free-form text or spelling mistakes. One conceptual approach to tackle the resulting data sparsity is to find links between the categories, as with semantics.

Neural networks have been shown very efficient to jointly extract semantics and solve a supervised classification problem in natural language processing [1]. We want to explore their use in databases. One challenge is, however, that the typical database table has significantly less rows than a natural-language-processing corpus. It comes with less statistical power, and deep learning approaches cannot be readily applied.

This research is set in the DirtyData project, that develops tools for easy statistical analysis of data without prior cleaning. It is also run in the context of the Paris-Saclay CDS –center for data science– that strive to bridge between data-science research and a variety of applications.

Proposed work: The goal of this research is to find global solutions that can be applied for statistical analysis of dirty databases, and not to fine-tune a neural architecture on a specific problem. The research will be conducted starting from empirical work on 7 databases touching a variety of fields (healthcare, wages, road safety, beers). It will strive to come to general conclusions.

The first task will be to tuned the architecture of a neural network to maximize prediction accuracy on each database. We expect the network to be fairly shallow as the number of samples vary from thousands to hundreds of thousands. We also expect a bottleneck layer to capture the semantic similarity between categories.

The second task will be to strive for a compromise across the different databases: an architecture that performs well on all these datasets as well as general guiding rules to adapt an architecture to the database.

Required skills: The successful candidate will know well machine learning practice and evaluation. He should be fluent in scientific computing with Python as the work will be done in Python and involve technical challenges. In addition, he or she should understand optimization techniques in machine learning and classic machine learning model and be able to implement these methods. An interest for applications, typically in economical or social data or public health is welcomed.

[1] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537, 2011.