Thesis Proposal

MACHINE LEARNING DEPARTMENT

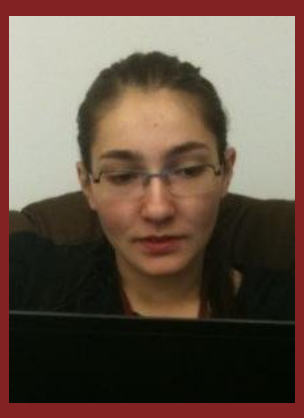
Discovering Compact and Informative Structures through Data Partitioning

Abstract:

In many practical scenarios, complex high-dimensional data contains low-dimensional structures that could be informative of the analytic problems at hand. My thesis aims at detecting such structures, if they exist, and at using them to construct compact interpretable models for different machine learning tasks that can benefit various practical applications.

We formalize the problem of Informative Projection Recovery. It is the problem of extracting a small set of low-dimensional projections of data which jointly support an accurate solution to a given learning task. Our solution to this problem is a regression-based algorithm that identifies informative projections by optimizing over a matrix of point-wise loss estimators. It generalizes to multiple types of machine learning problems, offering solutions to classification, clustering, regression, and active learning tasks. Experiments show that our method can discover and leverage low-dimensional structures in data, yielding accurate and compact models. Our method is particularly useful in applications involving multivariate numeric data in which expert assessment of the results is of the essence

The focus of our forthcoming research is on cost-sensitive feature selection put in the context of the underlying compact structures in data. We consider the process used to generate the features, as well as their reliability and interdependence to reduce the overall cost. Typically, our applications rely on a core set of features obtained through expensive measurements, enhanced using transformations derived from one or several core features. Our preliminary results show that leveraging low-dimensional structures may yield more powerful models without an increase in the cost of feature acquisition. The crux of our proposed technique is to leverage the submodular cost and the redundancy of the features by generating penalties according to the structure of their dependency graph. We will then develop online, adaptive policy-learning optimization procedures for feature selection with submodular cost constraints. We will first consider the batch mode setting and learn a model that maps samples to the appropriate feature subset, achievable by maximizing a submodular objective. The aim is to then efficiently update this mapping as more data becomes available, the main challenge being the trade-off between flexibility and robustness. We also plan to develop temporally-aware models which would dynamically adjust the selection of low-dimensional structures as the underlying data distribution varies over time.



Speaker:

Madalina Fiterau

PhD Candidate

Committee: Artur Dubrawski (Chair)
Geoff Gordon
Alex Smola
Andreas Krause (ETH Zurich)

October 27, 2014
9:00am
8102 GHC

http://www.cs.cmu.edu/~mfiterau/proposal/proposal.pdf