Internship / master project topic: A study of noise models in learning from networked data

Jan Ramon

December 2015

1 Context

The vast majority of statistical theory and machine learning theory expects input consisting of independently and identically distributed observations (i.i.d.). In large-scale data networks, objects influence each other, and it is hard to isolate independent observations. Recently, some efforts were made for developping learning theory for network-structured data (Wang, 2015), (Wang et al. 2014) Still, several challenges remain.

2 Objective

This project will make one important step forward by studying in more detail models of noise in the learning assumptions. In particular, starting from the basic methods developed in (Wang, 2015), we will study several alternative models of observable/unobservable features, the non-gaussian effects of small (sub)samples, models of noise on object features and target label features, and how this influences the optimal learning strategy learning strategy (and learning guarantees) in both SVM-like and Gaussian-process style settings.

3 More information

This project will involve:

- a literature study on existing approaches to networked statistics, concentration inequalities and learning theory.
- make a comprehensive overview of the settings to study and their parameters
- develop improved learning guarantees (and if necessary improved strategies) for several of the settings
- validate the theoretical bounds and their preciseness with a number of simple experiments on synthetic data.

4 Requirements

This project requires mathematical creativity. The student should be confident having a strong background and the required skills.

5 References

(Wang,2015) Yuyi Wang, From Graph Patterns to Networked Statistics, PhD thesis, July 2015, `https://lirias.kuleuven.be/handle/123456789/498409`

(Wang et al. 2014) Yuyi Wang, Jan Ramon and Christos Pelekis. U-statistics on network-structured data with kernels of degree larger than one, ECML/PKDD 2014 workshop on Statistically Sound Data Mining, Nancy, France, September 2014, `https://lirias.kuleuven.be/handle/123456789/469519`