Statistical learning for Markovian dependent data

The main objective of this project is to generalize some aspects of statistical learning theory from i.i.d. case into a Markovian framework.

The information explosion propelled by the advent of online social media, Internet, and global-scale communications has rendered data-driven statistical learning increasingly important. At any time around the globe, large volumes of data are generated by todays ubiquitous communication, imaging, and mobile devices such as cell phones, surveillance cameras and drones, medical and e-commerce platforms, as well as social networking sites. Learning from these large volumes of data is expected to bring significant science and engineering advances along with improvements in quality of life.

Numerous applications of statistical learning have driven both theorists as well practictioners to develop various learning methods. Most of the theory is extensively investigated only in the i.i.d. case. However, the i.i.d. assumption cannot often be strictly justified in real-world applications. Applications such as market prediction, system diagnosis, and speech recognition are inherently temporal in nature, and consequently not i.i.d. processes.

In this project we are particularly interested in investigation of learning algorithms when the data are Markovian. Such structure of data occurs naturally when one deals with storage and queuing systems, in finance, insurance or food risk assessment models and practically in any field where anomaly detection is involved. Our main goals are to investigate the performance of the learning algorithms (when dealing with Harris Markov chain samples) via empirical risk minimization. We are mainly interested in classification problems in a Markovian framework. In machine and statistical learning, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. An example would be assigning a given email into ”spam” or ”non-spam” classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of pattern recognition. Before one can use the statistical learning method in a Markovian case, one have to know if such classifying method gives reasonable responses. In this project we will investigate the behaviour and consistency of selected classification algorithms, such as k-means or $k-$nearest neighbors. Motivated by practical applications, our work is by many means theoretical and all the results will be confirmed by rigorous mathematical proofs. We will obtain most of the results for statistical learning via exponential inequalities for Harris recurrent Markov chains which we establish at the beginning of the project.

Finally, we motivate our approach and need to extend the statistical learning theory to Markovian case by so relevant words of V. Vapnik

Nothing is more practical than a good theory.

Our project is a great example of that.