Thesis Proposal

MACHINE LEARNING DEPARTMENT

Structured Sparse Regression Methods for Learning from High-Dimensional Genomic Data

Abstract:

The past several decades have witnessed an unprecedented explosion in the size and scope of genomic datasets, paving the way for statistical and computational data analysis techniques to play a critical role in driving scientific discovery in the fields of biology and medicine. However, genomic datasets suffer from a number of problems that weaken their signal to noise ratio, including small sample sizes and widespread data heterogeneity. As a result, the naive application of traditional machine learning approaches to many problems in computational biology can lead to unreliable results and spurious conclusions.

In this thesis, we propose several new techniques for extracting meaningful information from noisy genomic data. To combat the challenges posed by high-dimensional, heterogeneous datasets, we leverage prior knowledge about the underlying structure of a problem to design models with increased statistical power to distinguish signal from noise. Specifically, we rely on structured sparse regularization penalties to encode relevant information into a model without sacrificing interpretability. Our models take advantage of knowledge about the structure shared among related samples, features, or tasks, which we derive from biological insights, to boost their power to identify true patterns in the data.

Finally, we apply these methods to several widely studied problems in computational biology, including identifying genetic loci that are associated with a phenotype of interest, learning gene regulatory networks, and predicting the survival rates of cancer patients. We demonstrate that leveraging prior knowledge about the structure of a problem leads to increased statistical power to detect associations between different components of a biological system (e.g., SNPs and genes), providing greater insight into complex biological processes, and to more accurate predictions of disease phenotypes, leading to improved diagnosis or treatment of human diseases.

Speaker:

Micol Marchetti-Bowick

Thesis Committee:

Eric P. Xing, Chair
Seyoung Kim
Jian Ma
Su-In Lee (University of Washington)

Jan. 23, 2018
5:00pm
6501 GHC

Link to draft document:
http://www.cs.cmu.edu/~mmarchet/documents/thesis_proposal.pdf