CM 226: Machine Learning in Bioinformatics (Fall 2016)

Instructor: Sriram Sankararaman

Course description:

What genes cause cancer ? Have we inherited genes from Neanderthals ? How does a single genome code for so many different functions?

We can now begin to answer these fundamental questions in biology because the cost of genome sequencing has fallen faster than Moore's law. Thus, the bottleneck in answering these questions has shifted from data generation to powerful statistical models and inference algorithms that can make sense of this data. *Statistical machine learning* provides an important toolkit in this endeavor. Further, biological datasets offer new challenges to the field of machine learning.

We will learn about probabilistic models, inference and learning in these models, model assessment, and interpreting our inferences to address the biological question of interest. The course is aimed at a broad audience. It aims to introduce CS/Statistics students to this exciting source of problems and Bioinformatics/Human Genetics students to a rich set of tools.

Familiarity with probability, statistics, linear algebra and algorithms is expected. Programming experience is expected. No familiarity with biology is needed.

Learning goals:

- Students will learn about probabilistic models, efficient inference and learning in these models, model assessment, and interpreting the inferences to address the biological question at hand. The course will enable students to formulate the biological question as problems in statistical inference, to understand the assumptions and tradeoffs underlying these formulations, to find or develop efficient inference algorithms and to assess the quality of their inferences.
- A second goal is to get students started on independent research. An open-ended research project will form a major part of the course.

Textbooks:

There is no formal textbook. Readings will be posted as needed. The following texts will serve as useful references:

- Machine Learning: A Probabilistic Perspective by Kevin Murphy.
- Elements of Statistical Learning by Trevor Hastie, Robert Tibshirani and Jerome Friedman
- Biological Sequence Analysis by Richard Durbin, Sean Eddy, Anders Krogh and Tim Mitchison.
- Principles of Population Genetics by Hartl and Clark.

A tentative list of topics

1. Lecture 1: Introduction to genomics
2. Lecture 2: Introductory statistics. Multiple hypothesis testing.
3. Lecture 3: Regression. Application: association studies (GWAS).
4. Lecture 4: Multiple Regression (Sparsity and Bayesian models).
5. Lecture 5: Linear Mixed Models. Application: heritability.
6. Lecture 6: Latent Variable Models (LVMs): Clustering and Mixture models. EM algorithm
7. Lecture 7: LVMs : PCA and admixture models. Application: population structure.
8. No-class
9. Lecture 8: Directed graphical models
10. Lecture 9: Hidden Markov Models
11. Lecture 10: Tree-structured Graphical Models. Undirected graphical models
12. Lecture 11: Exact inference: sum-product algorithm. Application to admixture models
13. Lecture 12: Approximate inference: MCMC and variational inference.
14. Lecture 13: Kernel machines and gaussian processes. Application: rare-variant association
15. Lecture 14: Bayesian nonparametrics: Dirichlet process and the coalescent
16. Mid-term
17. Lecture 15: Application: Population genetics and inferring human history
18. Lecture 16: Genomic privacy: Learning with privacy constraints.
19. Lecture 17: Deep learning
20. Lecture 18: Causal inference