COM SCIM229S-2 / BIOL CHM229S-2 / HUM GENM229S-2: Machine Learning in Bioinformatics (Spring 2016)

Instructor: Sriram Sankararaman

Course description:

What genes cause cancer ? Have we inherited genes from Neanderthals ? How does a single genome code for the different cells ?

We can now begin to answer these fascinating questions in biology because the cost of genome sequencing has fallen faster than Moore's law. 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 the inferences to address the biological questions 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.

In this course, we will focus on the statistical and computational aspects of machine learning techniques and their application to key biological questions. Familiarity with probability, statistics, linear algebra and algorithms 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.

Course format:

- **Readings**: Each class will be assigned one or two readings. At the end of the class, please post a short summary or comments or critiques on the readings to CCLE.
- **Scribed lecture notes:** Each student will be assigned one lecture to scribe. The scribed lectures will be due one week after the assigned lecture. A latex template will be made available for scribing.
- **Homework**: There will be three homeworks. Questions on the homework will include programming exercises and data analyses as well as questions drawn from the assigned readings. You are free to use a programming language of your choice though R is preferred. The homeworks must be submitted in hard copy in class on the day they are due. Late submissions will not be accepted.
- **Project**: A major component of this course will be an open-ended project. The project can be the development of a statistical model/algorithm to a biological problem or application of an existing technique. I will post a list of potential projects on CCLE. You are welcome to propose any project that is relevant to the course, including rotation projects.

  Each group should decided on their project by the third week. The group will be expected to present their project in class near the end of the quarter and submit a project report.

Grading:

1. Project: 50% (30% paper, 20% presentation)
2. Homeworks: 30%
3. Scribing: 10%
4. Readings: 10%

A tentative list of topics

1. Lecture 1: Introduction to genomics
2. Lecture 2: Introductory statistics. Multiple hypothesis testing.
3. Lecture 3: Association studies. Regression (Linear, logistic)
4. Lecture 4: Association studies. Regression (sparse and Bayesian linear models)
5. Lecture 5: Heritability and its missingness. Linear Mixed Models.
6. Lecture 6: Guest lecture
7. Lecture 7: Gene expression: Latent Variable Models: Clustering and Mixture models. EM algorithm
8. Lecture 8: Inference in LVMs : MCMC and variational inference
9. Lecture 9: Correcting for confounding. LVMs: PCA and admixture models
10. Lecture 10: Hidden Markov Models
11. Lecture 11: Phylogenetics: Tree-structured Graphical Models
12. Lecture 12: Cancer genomics: Loopy Graphical Models
13. Lecture 13: Rare variant association: Kernel machines

14. Lecture 14: Population genomics: coalescent theory
15. Lecture 15: Bayesian nonparametrics: Poisson, Gaussian process
16. Lecture 16: Bayesian nonparametrics: Dirichlet process
17. Lecture 17: Deep learning
18. Lecture 18: Genomic privacy: Learning while preserving privacy
19. Lecture 19: Project presentations
20. Lecture 20: Project presentations