CM 226: Machine Learning in Bioinformatics (Fall 2021)

Instructor: Sriram Sankararaman

Course description:

What genes cause disease? How does a single genome code for different biological functions? Have we inherited genes from Neanderthals?

We can now begin to answer these fundamental 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 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.

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:

- **Homework**: There will be five homeworks. Questions on the homework will include programming exercises and data analyses as well as questions drawn from the assigned readings. You are strongly encouraged to use R (a free statistical programming language) although you could use other languages as well. 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 typically involves the development of a statistical model/algorithm to a biological problem or application of an existing technique to a biological dataset. We will post a list of potential projects. You are welcome to propose any project that is relevant to the course, including rotation projects. You can work on the project in groups of 1-3.

  Each group should decide on their project by the third week. The group will be expected to prepare a video presentation near the end of the quarter and submit a project report. Barring exceptional circumstances, all members of a group will be awarded the same score for the project.

- **Mid-term**: 24-hour take-home exam (open book/notes).

Grading:

1. Project: 30% (5% for proposal, 10% for video presentation, 15% for final report).
2. Homeworks: 50%
3. Mid-term exam: 20%

A tentative list of topics

1. 9/27 Introduction to genomics
2. 9/29 Introductory statistics. Multiple testing.
3. 10/4 Regression. Application: association studies for quantitative traits.
4. 10/6 Logistic regression. Application: Association studies for binary traits.
5. 10/11 Ridge and penalized regression. Application: The mystery of missing heritability
6. 10/13 Clustering and Mixture models.
7. 10/18 The EM algorithm.
8. 10/20 PCA. Application: Inferring population structure
9. 10/25 Admixture models. Application: Ancestry inference.
10. 10/27 Graphical models. Application: confounding and population stratification.
11. 11/1 No class. Midterm
12. 11/3 Graphical models and conditional independence.
13. 11/8 Hidden Markov Models.
14. 11/10 Phylogenetic trees.
15. 11/15 Kernel methods.
16. 11/17 Deep learning.
17. 11/22 Interpretable ML

18. 11/24 Genomic privacy
19. 11/29 Project presentations
20. 12/1 Project presentations