CSI 5180. Topics in Artificial Intelligence
Machine Learning for Bioinformatics Applications

Marcel Turcotte (marcel.turcotte@uottawa.ca)

School of Electrical Engineering and Computer Science
University of Ottawa

“In the not so distant past, data generation was the bottleneck, now it is data mining, or extracting useful biological insights from large, complicated datasets.”

Xu, C. & Jackson, S. A.  
Machine learning and complex biological data. 
Genome Biology 20, (2019).

Preamble

At the fall 2019, I will be lecturing the first edition of Machine Learning for Bioinformatics Applications. Its emphasis will primarily be on the analysis of complex biological data using modern machine learning methods. At this point, it is difficult to say what the prerequisite for the course will be since the course is still under development. Given that I want the course to be self-contained, I will not assume prior machine learning knowledge. However, a basic understanding of probability and statistics is needed, as well as, calculus and linear algebra. Likewise, I will not assume prior knowledge of bioinformatics. For several years now, I have been lecturing a course entitled Algorithms in Bioinformatics (CSI 5126) focusing on the data structures and algorithms behind the main bioinformatics applications. Here, rather than using the traditional algorithmic approaches, we will be using machine learning, so no knowledge of bioinformatics is assumed. Nevertheless, I am expecting that you can write programs in a high-level programming language, specifically Python. Now, what about biology? Biology is important as bioinformatics strives to solve “real-world” problems. There will be at least two lectures introducing essential concepts of the molecular biology of the cell. Inevitably, we will revisit these concepts each time that a new problem will be introduced. At the very least, I am expecting a desire to learn more about biology. Finally, I would like to stress that the breadth and depth of the course will most likely not be what I would like it to be in this first edition. Continue reading for further information and resources.

September 5, 2019

1 Web sites

- https://www.eecs.uottawa.ca/~turcotte/teaching/CSI-5180/
- https://piazza.com/uottawa.ca/fall2019/CSI5180
- https://uottawa.brightspace.com

2 Schedule

The first lecture will be held on September 5, 2019.

- Lectures: Tuesday, 13:00 to 14:30, and Thursday, 11:30 to 13:00, at 125 University (MNT) 103
- Office hours: Tuesday from 14:30 to 16:00 at STE 5-106
- Official schedule: www.uottawa.ca/course-timetable
3 Description

Machine learning theories and methods with applications to biological sequence data, gene expression, genomics and proteomics.

4 Learning Outcomes

Upon completion of the course, you will be able to:

- **Encode** and **clean** biological data for machine learning applications
- **Apply** modern machine learning methods to solve bioinformatics problems
- **Find** optimal values for the hyperparameters a given machine learning algorithm and data set
- **Use** a sound methodology for your machine learning projects
- **Critically review** scientific publications in this field
- **Locate** and **critically evaluate** scientific information
- **Present** scientific content to a small technical audience

5 Outline

Here is a tentative and ambitious course outline.

1. Overview
2. Essential Cell Biology (Part 1)
3. Essential Cell Biology (Part 2)
4. Essential Bioinformatics Skills (Databases, APIs, Frameworks)
5. Fundamentals of Machine Learning
6. Feature Engineering
7. Data Imputation
8. Dimensionality Reduction
9. Unsupervised Learning
10. Linear and Logistic Regression
11. Decision Trees, Random Forests and eXtreme Gradient Boosting
12. Extreme Learning Machines
13. Hidden Markov Models
14. Kernel Methods
15. Support Vector Machines
17. Deep Learning: Embeddings
18. Deep Learning: Architectures
19. Concept and Rule-based
20. Learning Graphs
21. Ensemble
22. Semi-supervised Learning
23. Data Integration
24. Automated Scientific Discovery

6 Evaluation

The evaluation has four components: assignments (3), a presentation, a project, and examinations (2).

- 30% — assignments (3)
- 10% — presentation
- 20% — project
- 40% — examinations (2)
6.1 Deadlines

- Assignments
  - A1 - October 10, 2019, 18:00
  - A2 - October 31, 2019, 18:00
  - A3 - November 21, 2019, 18:00

- Presentation
  - Schedule will be published on September 19, 2019
  - Presentations between October 1, 2019 and December 3, 2019

- Project
  - Outline - October 1, 2019
  - Report - December 3, 2019

- Examinations
  - Midterm - October 24, 2019
  - Final - December 5 to 18, 2019

6.2 Assignments

The assignments are done individually and there will be three of them. They are programming assignments with specific learning objectives in mind. For example, the learning objectives for the first assignment include: encode biological data for a specific machine learning task, implement two metrics to compare sequence data, apply an unsupervised learning algorithm to summarize some data. Python will be used for our assignments, along with popular machine learning libraries, including Scikit-Learn, Keras, and TensorFlow.

6.3 Presentation

Papers in (refereed) journals and conference proceedings are the main vehicle for communicating scientific information. You must select a publication that presents either a specialized application or a more efficient algorithm on a topic that has been presented in class. Starting October 1, 2019, there will be one student presentation per lecture. Each presentation will be related to the topic of the lecture. Students are randomly assigned a date.

6.3.1 Learning Objectives

- Thoroughly study of a specific topic in bioinformatics
- Familiarity with the modes of communicating research
- Develop your presentation skills

6.3.2 Deliverable

- 15–20 minutes presentation

6.4 Project

6.4.1 Learning Objectives

- Thorough study of a specific bioinformatics problem using two machine learning approaches
- Learning to study autonomously
6.4.2 Deliverable

You will replicate the results from a recent scientific publication. You must create a suitable data set and apply at least two distinct algorithms. You will apply the methodology proposed in class to select the values of the hyperparameters and evaluate the result. You must hand in your data, your source code, as well as a short report (5-10 pages).

6.5 Examinations

There will be a midterm and final examination. Students that have 90% or above for their combined grade for the assignments, midterm, and presentation can be exempted from writing the final examination.

7 Material and Resources

- Lecture notes (slides) and complementary resources will be posted on the course Web site: http://www.eecs.uottawa.ca/~turcotte/teaching/csi-5180/lectures

Below, you will find a number of references to Springer Link, which provides our community with access to journals, books, series, protocols and reference documents, access is restricted to the University of Ottawa, based on your IP address.

7.1 Monographs

- Sumeet Dua and Pradeep Chowriappa, *Data Mining for Bioinformatics*. CRC Press, 2012. (On Amazon.ca)

7.2 Essential Cell Biology

  - NCBI Bookshelf
8 Relationship to CSI 5126: Algorithms in Bioinformatics

**Algorithms in bioinformatics** focuses on the data structures and algorithms for solving classical bioinformatics problems. For instance, it introduces suffix trees and suffix arrays, which allow to solve a number of string problems in linear time. Dynamic programming finds many applications in bioinformatics, to align biological sequences or to reconstruct the ancestral states in a phylogeny, to name but a few examples. The calendar description is as follows:

> Fundamental mathematical and algorithmic concepts underlying computational molecular biology; physical and genetic mapping, sequence analysis (including alignment and probabilistic models), genomic rearrangements, phylogenetic inference, computational proteomics and systemic modelling of the whole cell.

No prior knowledge of bioinformatics should be needed to succeed with **Machine Learning for Bioinformatics Applications** and the intersection between the content of two courses should be minimum.

- [http://www.site.uottawa.ca/~turcotte/teaching/csi-5126/lectures/](http://www.site.uottawa.ca/~turcotte/teaching/csi-5126/lectures/)

### 8.1 Bioinformatics Resources

For a high-level, first encounter with bioinformatics, I am suggesting the following textbook. Beware, this is the fifth edition. If you can, I recommend accessing the latest edition since high-throughput technologies are fast evolving. Here is what the editor had to say regarding the latest edition: “A host of new material includes new content on next generation sequencing, function prediction, sequence assembly, epigenomics, the bioinformatics of gene editing, and the effects of single nucleotide variants.”


For a practical and economical \(^1\) introduction, you might have a look at **The Biostar Handbook: Bioinformatics data analysis guide, 2019**, which also gives you access to a number of online courses. Namely, the guide provides an introduction to Unix and Conda, which are both of importance for Bioinformatics, but also for Machine Learning!

- [https://biostar.myshopify.com](https://biostar.myshopify.com)

For those who would like to explore the subject even further, the following monographs are part of my short list of essential Bioinformatics books.


### 8.2 On the Web

- **MIT 7.91 J**: Foundations of Computational and Systems Biology
  - [https://www.youtube.com/playlist?list=PLUL4u3cNGP63uK-oWiLg07LLJV62CWXac](https://www.youtube.com/playlist?list=PLUL4u3cNGP63uK-oWiLg07LLJV62CWXac)
- **Johns Hopkins University**: Algorithms for DNA Sequencing
  - [https://www.coursera.org/learn/dna-sequencing](https://www.coursera.org/learn/dna-sequencing)

\(^1\) $25 USD.
University of California San Diego: Bioinformatics Specialization

- https://www.coursera.org/learn/dna-analysis?specialization=bioinformatics
- https://www.coursera.org/learn/genome-sequencing?specialization=bioinformatics
- https://www.coursera.org/learn/comparing-genomes?specialization=bioinformatics
- https://www.coursera.org/learn/genomic-data
- https://www.coursera.org/learn/dna-mutations
- https://www.coursera.org/learn/bioinformatics-project

Learning bioinformatics through problem solving:

- http://rosalind.info/

9 Plagiarism

Academic fraud is an act by a student that may result in a false evaluation (including papers, tests, examinations, etc.). It is not tolerated by the University. Any person found guilty of academic fraud will be subject to severe sanctions. Here are some examples of academic fraud:

- Plagiarism or cheating of any kind;
- Present research data that has been falsified;
- Submit a work for which you are not the author, in whole or part;
- Submit the same piece of work for more than one course without the written consent of the professors concerned.
- Please consult this webpage: it contains regulations and tools to help you avoid plagiarism.

An individual who commits or attempts to commit academic fraud, or who is an accomplice, will be penalized. Here are some examples of possible sanctions:

- Receive an “F” for the work or in the course in question;
- Imposition of additional requirements (from 3 to 30 credits) to the program of study;
- Suspension or expulsion from the Faculty.
- You can refer to the regulations on this webpage.

10 Student Services

10.1 Academic Writing Help Centre

At the AWHC you will learn how to identify, correct and ultimately avoid errors in your writing and become an autonomous writer. In working with our Writing Advisors, you will be able to acquire the abilities, strategies and writing tools that will enable you to:

- Master the written language of your choice
- Expand your critical thinking abilities
- Develop your argumentation skills
- Learn what the expectations are for academic writing

Further information is available here:

- http://www.sass.uottawa.ca/writing/
10.2 Career Services

Career Services offers various services and a career development program to enable you to recognize and enhance the employability skills you need in today's world of work.

- http://www.sass.uottawa.ca/careers/

10.3 Counselling Service

There are many reasons to take advantage of the Counselling Service. They offer:

- Personal counselling
- Career counselling
- Study skills counselling

Further information is available here:

- http://www.sass.uottawa.ca/personal/

10.4 Access Service

The Access Service acts as an intermediary between students, their faculty and other University offices to ensure that the special needs of these students are addressed and that the best possible learning conditions are being offered.

Note that the University of Ottawa is affiliated with AERO and ACE services for the adaptation of accessible academic materials for students with perceptual disabilities. If you have any questions, please contact the Accessibility Librarian or the Access services for textbooks.

- http://www.sass.uottawa.ca/access/

10.5 Policy – Prevention of Sexual Violence

The University of Ottawa will not tolerate any act of sexual violence. This includes acts such as rape and sexual harassment, as well as misconduct that take place without consent, which includes cyberbullying. The University, as well as various employees and student groups, offers a variety of services and resources to ensure that all uOttawa community members have access to confidential support and information, and to procedures for reporting an incident or filing a complaint. For more information, please visit www.uOttawa.ca/sexual-violence-support-and-prevention.

11 Information Sharing and Copyright

All documents prepared by the course instructor, including assignments, course notes, and exams, are protected by copyright. Copying, digitizing, or publishing on a Web site is therefore a violation of copyright and is illegal.

12 Regarding the Instructor

I have been conducting research in bioinformatics since 1989. I am particularly interested in the inference of structural motifs, mostly for ribonucleic acids.

A References

Reviews and Comparative Analyses


Evaluating Learning Algorithms


Dimensionality Reduction, Feature Selection, and Feature Engineering


**Data Imputation**


**Unsupervised Learning**


**Supervised Learning**


**Linear, Logistic Regression, naïve Bayes**


**Decision Trees, Random forests and eXtreme Gradient Boosting**


**Hidden Markov Models (HMM)**


**Extreme Learning Machines**


**Support Vector Machine (SVM)**


**Kernel Methods**


**Artificial Neural Networks (ANN)**


**Deep Learning**


Deep Learning: Word Embeddings


Concept and Rule-based Learning


Learning Graphs


[6] Meng Li, Jianmei Zhao, Xuecang Li, Yang Chen, Chenchen Feng, Fengcui Qian, Yuejuan Liu, Jian Zhang, Jianzhong He, Bo Ai, Ziyu Ning, Wei Liu, Xuefeng Bai, Xiaole Han, Zhiyong Wu, Xiue Xu, Zhidong Tang, Qi Pan, Liyan Xu, Chunquan Li, Qiuyu Wang, and Enmin Li. HiFreSP: A novel high-frequency sub-pathway mining approach to identify robust prognostic gene signatures. *Brief Bioinform*, Jul 2019.


Ensemble Learning


**Semi-Supervised Learning**


**Data Integration**


Reinforcement Learning


Statistics


Tools and Data Sets


Puzzles or puzzling


Automated Scientific Discovery


*Version of September 5, 2019*