SS857 DATA ANALYTICS FOR BIOINFORMATICS: MACHINE LEARNING 3-0-0 3

PhD Course

AMRITA VISHWA VIDYAPEETHAM

COURSE OBJECTIVE

The course is a machine learning course, but uses bioinformatic uses cases to teach the subject. Each module of the course starts with introduction to the machine learning (ML) concepts, followed by a bioinformatics use case. The student may pick a problem that is important to her/ his thesis in each module, or one will be assigned by the instructor.
Specific learning objectives for this course are:

1. Learn the theory and implementation of machine learning algorithms
2. Think about how machine learning algorithms assist cutting edge bioinformatics data
3. Relevance of machine learning to build life sciences hypotheses
4. Get students to start formulating biological problems in terms of machine learning parameters across various types of ML domains.

LEARNING OUTCOMES

This course seeks to teach machine learning to help the students learn to adapt solutions to various problems in bioinformatics use cases. ML provides powerful data analytic tools, but it is equally important to be able to recognize what tools are to be applied in what context. It is here, that this course is specific to big data bioinformatics problems and challenges. ML uses a combination of mathematics, computational theory. Biological data is vast and complex, thus, manual processing in its entirely is virtually impossible and will be largely driven by chance if not for the powerful tools machine learning provides. This course will teach the students how to clearly think about biological problems of interest to be parameterized so ML algorithms can be used to solve them, or whatever the objective is (hypothesis generation, classification, prediction, etc.). As the course proceeds, the students will aim to publish one journal paper based on learning from this course and their on field of interest.

COURSE RELEVANCE

Machine learning in bioinformatics is the exploration of biological data using tools provided by machine learning. With the high throughput data generation in the field of biology, it provides a strongly needed training to advanced graduate students to get up and running on data analysis of biological data. Computer Science students taking this course will have programming background, but will come out with clear understanding of machine learning concepts as applied to biological data. This analysis will be highly relevant in present day's state of research in biology where

biological experimentation and innovation is largely driven by inferences drawn from large data sets. So much, that biologists also look to inferences produced by ML and driven by biochemistry, biophysical concepts, and systems biology (like NGS and interaction maps) using various modeling techniques.

TEACHING METHODS

Teaching will done via A-view. the second half of the course will be graded based on hands-on lab exams. Apart from that, all other modes of testing, like T1, T2, etc. will be conducted.

REQUIRED COURSE MATERIALS AND READINGS

1. *Machine Learning: a Probabilistic Perspective* , Kevin Murphy.

OPTIONAL COURSE MATERIALS & READINGS (CASES, ARTICLES, REPORTS ETC)

Coursera course by Emily Fox/ Carlos Guestrin

EVALUATION CRITERIA

Students are evaluated based on their performance in personal learning paper; class participation and quiz; mid-term and end-term examination; and group project and presentations. The weightage for various components will be as follow:

First Internals	15%
Second Internal	15%
End-term examination	20%
Project on Bioinformatics	50%

DETAILS OF SESSION

Serial No.	TOPICS TO BE COVERED
	Unit I
1.	Regression, an introduction
2.	Simple Linear Regression
3.	Multiple Linear Regression
4.	Selected topics in Regression
5.	Bioinformatics use case (exercise will be decided by instructor or picked by student with instructor's approval)
	Unit II
6.	Classification
7.	Linear Classifiers
8.	Logistic Regression
9.	Decision Trees
10.	Selected Topics in Statistical Classification Techniques
11.	*Students problem statement presentation (X 2)*
12.	Bioinformatics use case (exercise will be decided by instructor or picked by student with instructor's approval)
	Unit III
13.	Clustering
14.	Nearest Neighbours
15.	k-means
16.	Hierarchical clustering
17.	Selected Topics in Clustering
18.	Bioinformatics use case (exercise will be decided by instructor or picked by student with instructor's approval)
	Unit IV
19.	*Students presentation start (multiple sessions)*