INFO B-429/B-529 Machine Learning in Bioinformatics

Department of BioHealth Informatics
Indiana University School of Informatics and Computing, Indianapolis
Spring 2022

Credit Hours: 3
Time: Thursday 3:00 pm – 5:40 pm
Location: IT 355, Informatics & Communications Technology Complex
535 West Michigan Street, Indianapolis, IN 46202
Zoom meeting ID: 730 117 6460 Password: 135791
https://iu.zoom.us/j/7301176460?pwd=MDZYV3NxVWN0UmRBWVR2bWRFZVdIQT09
First Class: January 13, 2021

Instructor: Amir Manzour, PhD
Office Hours: Monday 4:00 pm – 5:00 pm by appointment
Office: IT 479
535 W. Michigan Street, Indianapolis, IN 46202
Zoom meeting ID: 730 117 6460 Password: 135791
https://iu.zoom.us/j/7301176460?pwd=MDZYV3NxVWN0UmRBWVR2bWRFZVdIQT09
Phone: (317) 278-4108 (Office)
Email: amanzour@iu.edu
Website: https://soic.iupui.edu/people/amir-manzour/

Prerequisites: B519 Introduction to Bioinformatics

COURSE DESCRIPTION

Machine learning is a key technology in bioinformatics, especially in the analysis of "big data" in bioinformatics. This course gives an overview of basic concepts, algorithms, and applications in machine learning, including topics such as classification, linear regression, support vector machines, hidden Markov models, and probably approximately correct (PAC) learning. It describes the basic intuition behind machine learning methods to help students understand how the methods work. In addition, applications of machine learning in bioinformatics are discussed.

Reference Texts:

Title:	Pattern Recognition and Machine Learning
Author:	Christopher M. Bishop
Edition:	1st edition (October, 2007)
Publisher:	Springer
ISBN:	ISBN-10: 0-387-31073-8

Title:	Machine Learning
Author:	Tom M. Mitchell
Edition:	1st edition (March, 1997)
Publisher:	McGraw-Hill
ISBN:	ISBN 0-07-042807-7

Title:	The Elements of Statistical Learning: Data Mining, Inference, and Prediction
Authors:	T. Hastie, R. Tibshirani, and J. Friedman
Edition:	2nd edition (2009)
Publisher:	Springer
ISBN:	ISBN 0387848576

Title:	Learn from Data
Authors:	Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin
Edition:	1st edition (March 2012)
Publisher:	AMLBook
ISBN:	ISBN 1600490069

Course Outcomes:

The learning objectives of this course include the following:

1. To explain some of the basic concepts and theories in machine learning.
2. To explore with machine learning algorithms and tools used in biological data analysis.
3. To understand the pros and cons of various machine learning algorithms in applications.
4. To obtain the ability to choose appropriate machine learning algorithms and apply them on various applications in bioinformatics.

Core Competencies:

The core competencies of this course include the following:

1. The ability to model biological data analysis problems as machine learning problems.
2. The ability to analyze biological data sets with state-of-the-art machine learning algorithms and tools.
3. The ability to select correct machine learning tools and parameters in biological data analysis and to find valuable information from the data.

EXPECTATIONS, GUIDELINES, AND POLICIES

Incomplete:

The instructor may assign an Incomplete (I) grade only if at least 75% of the required coursework has been completed at passing quality and holding you to previously established time limits would result in unjust hardship to you. All unfinished work must be completed

by the date set by the instructor. Left unchanged, an Incomplete automatically becomes an F after one year. http://registrar.iupui.edu/incomp.html

Deliverables:

You are responsible for completing each deliverable (e.g., assignment, quiz) by its deadline and submitting it by the specified method. Deadlines are outlined in the syllabus or in supplementary documents accessible through Canvas. Should you miss a class, you are still responsible for completing the deliverable and for finding out what was covered in class, including any new or modified deliverable. You are allowed one excused or unexcused late submission in 24 hours after the deadline for homework assignments. For the second late submission, a grade on a deliverable shall be zero.

Exams/quizzes:

Mid-Term Exam	March 3	3:00 – 5:00 pm
Final Exam	April 28	3:00 – 5:00 pm

Class assignments:

Four home assignments will be given during the course.

Project:

Students will complete projects about analysis of biological data. Students will work in a group of 2 – 4 students. Each group will find biological data, design analysis methods, and report analysis reports. Each group will submit a methodology report and a final report, as well as present analysis results.

Grading Information:

- Homework 25%
- Exams 50%
  - Mid-term exam 20%
  - Final exam 30%
- Project 25%
  - Methodology report 8%
  - Final report 17%

WEEKLY SCHEDULE

Date for each class meeting

Date	Class topic	Percentage
January 13	Lecture 1 Introduction to machine learning	8%
January 20	Lecture 2 Linear models	8%
January 27	Lecture 3 Decision theory	8%
February 3	Lecture 4 VC dimensions and parameter estimation	8%
February 10	Lecture 5 Logistic regression and neural networks	8%
February 17	Lecture 6 Clustering	8%
February 24	Lecture 7 Motif finding and bi-clustering	8%
March 10	Lecture 8 Decision trees and motif finding	8%
March 24	Lecture 9 Support vector machines	8%
March 31	Lecture 10 Hidden Markov models	8%
April 7	Lecture 11 Dimensionality reduction	8%
April 14	Lecture 12 Ensemble learning	8%
April 21	Deep Learning & Review	

Lecture 1 Introduction to machine learning

Readings

- Abu-Mostafa YS, Magdon-Ismail M, and Lin HT (2012). *Learn from Data.* AMLBook
    - Chapter 1 The Learning Problem.

Lecture 2 Linear models

Readings

- Bishop CM (2007*). Pattern Recognition and Machine Learning*. McGraw-Hill.
    - Chapter 4.1 Linear Model for Classification: Discriminant Functions
- Golub *et al*. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science* 286: 531-7, 1999.

Lecture 3 Decision theory

- Bishop CM (2007*). Pattern Recognition and Machine Learning*. McGraw-Hill.
    - Chapter 1.5 Decision theory
- Abu-Mostafa YS, Magdon-Ismail M, and Lin HT (2012). *Learn from Data.* AMLBook.

- Chapter 2.1 Theory of Generalization

Lecture 4 VC dimensions and parameter estimation

- Abu-Mostafa YS, Magdon-Ismail M, and Lin HT (2012). *Learn from Data.* AMLBook.
  - Chapter 2.1 Theory of Generalization
- Mitchell TM (1997). *Machine Learning*, McGraw-Hill
  - Chapter 6 Bayesian Learning.

Lecture 5 Logistic regression and neural networks

- Bishop CM (2007*). Pattern Recognition and Machine Learning*. McGraw-Hill
  - Chapter 5 Neural networks (5.1 and 5.2).

Lecture 6 Clustering

- Bishop CM (2007*). Pattern Recognition and Machine Learning*. McGraw-Hill
  - Chapter 9 Mixture models and EM.

Lecture 7 Motif finding and biclustering

- Bailey TM and Elkan C, "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", *ISMB* 1994, pp. 28-36.
- Buhler J and Tompa M, "Finding Motifs Using Random Projections", *Journal of Computational Biology*, 9: 225-42, 2002.
- Cheng Y and Church GM. Biclustering of expression data. *ISMB* 2000, 8:93-103.

Lecture 8 Decision trees and motif finding

- Hastie T, Tibshirani R, and Friedman J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer.
  - Chapter 14.4 Self-organization maps.
- Mitchell TM (1997). *Machine Learning*, McGraw-Hill
  - Chapter 3 Decision tree learning.
- Lee NK and Wang D. SOMEA: self-organizing map based extraction algorithm for DNA motif identification with heterogeneous model, *BMC Bioinformatics*, 2011; 12(Suppl 1): S16.

Lecture 9 Support vector machines

- Hastie T, Tibshirani R, and Friedman J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer.
  - Chapter 12 Support vector machines.

Lecture 10 Hidden Markov models

- Rabiner LR, A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77, 257-283, 1989.

Lecture 11 Dimensionality reductions

- Fodor I, A survey of dimension reduction techniques. Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, Technical Report.

Lecture 12 Ensemble learning

- Hastie T, Tibshirani R, and Friedman J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer.
    - Chapter 15 random forest
    - Chapter 16 ensemble learning

Grading Scale:

Grading is based on the ranks of students.

Grade Category	Grade	Relative Class Rank Range
Exceptional	A+	Top 10%
Outstanding	A	Top 25%
Good	A-	Top 45%
Satisfactory	B+	Median%
Reasonable (Pass)	B	Lower 35%
Marginal (Pass)	B-	Lower 15%
Fail	C+, C, C-	-

MISSION STATEMENT

The Mission of IUPUI is to provide for its constituents' excellence in

- Teaching and Learning;
- Research, Scholarship, and Creative Activity; and
- Civic Engagement.

With each of these core activities characterized by

- Collaboration within and across disciplines and with the community;
- A commitment to ensuring diversity; and
- Pursuit of best practices.

IUPUI's mission is derived from and aligned with the principal components—Communities of Learning, Responsibilities of Excellence, Accountability and Best Practices—of Indiana University's Strategic Directions Charter.

STATEMENT OF VALUES

IUPUI values the commitment of students to learning; of faculty to the highest standards of teaching, scholarship, and service; and of staff to the highest standards of service. IUPUI recognizes students as partners in learning. IUPUI values the opportunities afforded by its location in Indiana's capital city and is committed to serving the needs of its community. Thus, IUPUI students, faculty, and staff are involved in the community, both to provide educational programs and patient care and to apply learning to community needs through service. As a leader in fostering collaborative relationships, IUPUI values collegiality, cooperation, creativity, innovation, and entrepreneurship as well as honesty, integrity, and support for open inquiry and dissemination of findings. IUPUI is committed to the personal and professional development of its students, faculty, and staff and to continuous improvement of its programs and services.

CODE OF CONDUCT

All students should aspire to the highest standards of academic integrity. Using another student's work on an assignment, cheating on a test, not quoting or citing references correctly, or any other form of dishonesty or plagiarism shall result in a grade of zero on the item and possibly an F in the course. Incidences of academic misconduct shall be referred to the Department Chair and repeated violations shall result in dismissal from the program.

All students are responsible for reading, understanding, and applying the *Code of Student Rights, Responsibilities and Conduct* and in particular the section on academic misconduct. Refer to *The Code* > *Responsibilities* > *Academic Misconduct* at http://www.indiana.edu/~code/. All students must also successfully complete the Indiana University Department of Education "How to Recognize Plagiarism" Tutorial and Test. https://www.indiana.edu/~istd You must document the difference between your writing and that of others. Use quotation marks in addition to a citation, page number, and reference whenever writing someone else's words (e.g., following the *Publication Manual of the American Psychological Association*). To detect plagiarism instructors apply a range of methods, including Turnitin.com. http://www.ulib.iupui.edu/libinfo/turnitin

Academic Misconduct:

1. **Cheating:** Cheating is considered to be an attempt to use or provide unauthorized assistance, materials, information, or study aids in any form and in any academic exercise or environment.
    a. A student must not use external assistance on any "in-class" or "take-home" examination, unless the instructor specifically has authorized external assistance. This prohibition includes, but is not limited to, the use of tutors, books, notes, calculators, computers, and wireless communication devices.
    b. A student must not use another person as a substitute in the taking of an examination or quiz, nor allow other persons to conduct research or to prepare work, without advanced authorization from the instructor to whom the work is being submitted.
    c. A student must not use materials from a commercial term paper company, files of papers prepared by other persons, or submit documents found on the Internet.

d. A student must not collaborate with other persons on a particular project and submit a copy of a written report that is represented explicitly or implicitly as the student's individual work.
e. A student must not use any unauthorized assistance in a laboratory, at a computer terminal, or on fieldwork.
f. A student must not steal examinations or other course materials, including but not limited to, physical copies and photographic or electronic images.
g. A student must not submit substantial portions of the same academic work for credit or honors more than once without permission of the instructor or program to whom the work is being submitted.
h. A student must not, without authorization, alter a grade or score in any way, nor alter answers on a returned exam or assignment for credit.

2. **Fabrication:** A student must not falsify or invent any information or data in an academic exercise including, but not limited to, records or reports, laboratory results, and citation to the sources of information.
3. **Plagiarism:** Plagiarism is defined as presenting someone else's work, including the work of other students, as one's own. Any ideas or materials taken from another source for either written or oral use must be fully acknowledged, unless the information is common knowledge. What is considered "common knowledge" may differ from course to course.
   a. A student must not adopt or reproduce ideas, opinions, theories, formulas, graphics, or pictures of another person without acknowledgment.
   b. A student must give credit to the originality of others and acknowledge indebtedness whenever:
      1. directly quoting another person's actual words, whether oral or written;
      2. using another person's ideas, opinions, or theories;
      3. paraphrasing the words, ideas, opinions, or theories of others, whether oral or written;
      4. borrowing facts, statistics, or illustrative material; or
      5. offering materials assembled or collected by others in the form of projects or collections without acknowledgment
4. **Interference:** A student must not steal, change, destroy, or impede another student's work, nor should the student unjustly attempt, through a bribe, a promise of favors or threats, to affect any student's grade or the evaluation of academic performance. Impeding another student's work includes, but is not limited to, the theft, defacement, or mutilation of resources so as to deprive others of the information they contain.
5. **Violation of Course Rules:** A student must not violate course rules established by a department, the course syllabus, verbal or written instructions, or the course materials that are rationally related to the content of the course or to the enhancement of the learning process in the course.
6. **Facilitating Academic Dishonesty:** A student must not intentionally or knowingly help

or attempt to help another student to commit an act of academic misconduct, nor allow another student to use his or her work or resources to commit an act of misconduct.

OTHER POLICIES

1. **IUPUI course policies:** A number of campus policies governing IUPUI courses may be found at the following link: http://registrar.iupui.edu/course_policies.html
2. **Classroom civility**: To maintain an effective and inclusive learning environment, it is important to be an attentive and respectful participant in lectures, discussions, group work, and other classroom exercises. Thus, unnecessary disruptions should be avoided, such as ringing cell phones engagement in private conversations and other unrelated activities. Texting, surfing the Internet, and posting to Facebook or Twitter during class are generally not permitted. IUPUI nurtures and promotes "a campus climate that seeks, values, and cultivates diversity in all of its forms and that provides conditions necessary for all campus community members to feel welcomed, supported, included, and valued" (IUPUI Strategic Initiative 9). IUPUI prohibits "discrimination against anyone for reasons of race, color, religion, national origin, sex, sexual orientation, marital status, age, disability, or [veteran] status" (Office of Equal Opportunity). Profanity or derogatory comments about the instructor, fellow students, invited speakers or other classroom visitors, or any members of the campus community shall not be tolerated. A violation of this rule shall result in a warning and, if the offense continues, possible disciplinary action.
3. **Bringing children to class:** To ensure an effective learning environment, children are not permitted to attend class with their parents, guardians, or childcare providers.
4. **Disabilities Policy:** In compliance with the Americans with Disabilities Act (ADA), all qualified students enrolled in this course are entitled to reasonable accommodations. Please notify the instructor during the first week of class of accommodations needed for the course. Students requiring accommodations because of a disability must register with Adaptive Educational Services (AES) and complete the appropriate AES-issued before receiving accommodations. The AES office is located at UC 100, Taylor Hall (Email: aes@iupui.edu, Tel. 317 274-3241). Visit http://aes.iupui.edu for more information.
5. **Administrative Withdrawal:** A basic requirement of this course is that students participate in all class discussions and conscientiously complete all required course activities and/or assignments. If a student is unable to attend, participate in, or complete an assignment on time, it is the student's responsibility to inform the instructor. If a student misses more than half of the required activities within the first 25% of the course without contacting the instructor, the student may be administratively withdrawn from this course. Administrative withdrawal may have academic, financial, and financial aid implications. Administrative withdrawal will take place after the full refund period, and a student who has been administratively withdrawn from a course is ineligible for a tuition refund. Contact the instructor with questions concerning administrative withdrawal.