QBIO 460 Introduction to Machine Learning in Biology

Units: 4

Fall 2024 Semester

Lecture: Mondays and Wednesdays, 2:00-3:20 pm

Discussion: Fridays, 11:00-11:50 am

Location: RRI 301

Instructor: Tsu-Pei Chiu, PhD
Office: RRI 413J
Office Hours: TBD
Contact Info: tsupeich@usc.edu

Teaching Assistant: TBD
Office: TBD
Office Hours: TBD
Contact Info: TBD

Short Description

Fundamentals of ML; regression analysis, classification methods, clustering techniques; dimensionality reduction methods, ensemble learning, and deep learning technologies; applications in various biological contexts; project-based.

Course Description

This course presents an extensive range of Machine Learning (ML) techniques, with a primary focus on traditional methodologies, complemented by an exploration of deep learning approaches. It guides students through the process of applying these technologies to a variety of biological challenges. Emphasizing a thorough and accessible teaching method, the course engages students in hands-on experiences, working intimately with a diverse array of biological datasets.

Learning Objectives

The course covers a wide range of topics, beginning with the fundamentals of ML and its applications in various biological contexts. Students will learn Python programming for ML, delve into regression analysis, and explore classification methods such as logistic regression, KNN, and SVM, along with clustering techniques. They will also examine dimensionality reduction methods like PCA and t-SNE, understand ensemble learning with Random Forest and AdaBoost, and delve into deep learning technologies. The primary programming language used will be Python, which will be introduced in lectures tailored to ML applications. Throughout the course, students will engage with various biological datasets, enhancing their skills in data analysis and model implementation. These skills will be applied in their weekly computing assignments and a comprehensive end-of-semester project. After taking this course, students will be able to apply machine learning techniques to real-world biological problems, analyze complex datasets, and develop models to gain insights in the field of biology.

Prerequisite(s): none

Recommended Preparation: There are no prerequisites or co-requisites for this course. Programming experience in Python is recommended.

Course Notes
This course is taken for a letter grade. Lecture slides will be posted on Brightspace.

Technological Proficiency and Hardware/Software Required
Students will need access to a computer. It will be helpful (but not required) if students have a laptop that they can bring it to class.

Suggested Readings and Supplementary Materials
There is no textbook for this course. However, we provide suggested readings and supplementary materials below. Additional supplemental readings will be posted on Brightspace before or after each lecture.

Reading materials:
[KM] Probabilistic Machine Learning: An Introduction. Kevin Murphy.
[AG] Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems, 2nd Edition. Aurélien Géron. O'Reilly Press.
[JV] Python Data Science Handbook, Jake VanderPlas. O'Reilly Press.

Supplementary materials:
[S1] Sun, B.B., Kurki, M.I., Foley, C.N. et al. (2022). Genetic associations of protein-coding variants in human disease. Nature, 603, 95–102. https://doi.org/10.1038/s41586-022-04394-w
[S2] Mathelier, A., Xin, B., Chiu, T.P., Yang, L., Rohs, R., & Wasserman, W.W. (2016). DNA shape features improve transcription factor binding site predictions in vivo. Cell Systems, 3, 278-286.
[S3] Gligorijević, V., Renfrew, P.D., Kosciolek, T. et al. (2021). Structure-based protein function prediction using graph convolutional networks. Nature Communications, 12, 3168. https://doi.org/10.1038/s41467-021-23303-9
[S4] Melms, J.C., Biermann, J., Huang, H. et al. (2021). A molecular single-cell lung atlas of lethal COVID-19. Nature, 595, 114–119. https://doi.org/10.1038/s41586-021-03569-1
[S5] Alipanahi, B., Delong, A., Weirauch, M. et al. (2015). Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology, 33(8), 831–838. https://doi.org/10.1038/nbt.3300
[S6] Zhou, J., Troyanskaya, O. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods, 12(10), 931–934. https://doi.org/10.1038/nmeth.3547

Assignments
The course includes weekly computing assignments and an end-of-the-semester project. Each week, students will receive computer-based problem sets in Python, aimed at reinforcing the principles discussed in lectures and providing practical computing experience. The semester will culminate in a research project, where students are required to design, execute, and write a 3 to 5-page report on their findings, due at semester's end. Students have the liberty to choose their project topics, which should be related to the lecture content. Topic suggestions will be provided if needed, and graduate students are welcome to use their thesis data. A one-page project proposal should be discussed with me by week 10.

Sample assignment questions:

In this assignment, you are tasked with exploring the use of PCA (Principal Component Analysis) and tSNE (t-Distributed Stochastic Neighbor Embedding) to analyze scRNA-Seq data. These methods are instrumental in reducing dimensionality and visualizing the complex data, thereby enabling the identification of patterns and clusters that signify different cell types or states. Your analysis will delve into the effectiveness of these

techniques in mitigating the inherent noise and sparsity of scRNA-Seq data, facilitating a deeper understanding of cellular diversity and function.

1. *Import the RNA-Seq dataset 'counts.npy' using the NumPy library's np.load() function and store it in a variable called 'dataset'. Next, load the 'labels.txt' file, containing data labels, with np.loadtxt() and assign it to 'labels'. This organization facilitates your data and labels being ready for subsequent analysis. [1pt]*

2. *In this task, you are to perform PCA on the dataset you previously loaded. First, import the PCA class from sklearn.decomposition. Set up PCA with n_components=2, indicating that we wish to reduce our dataset to two principal components. Apply the fit_transform method to your dataset to obtain the PCA results named pca_results and visualize the results using matplotlib. [2pt]*

Grading Breakdown

Assessment	Points	% of Grade
In-class quizzes	0.3-0.5 each, 8 in total	8
Weekly computing assignments	7 each, 70 in total	70
Final project proposal	2	2
End-of-semester project	20	20

Assignment Submission Policy

Most weeks there will be a computing assignment. Assignments will both be posted and submitted on Brightspace.

Course Specific Policies

Late assignments will not be accepted without prior approval. Every student must submit their own assignment.

The professor reserves the right to make changes to the syllabus; these changes will be announced as early as possible so that students can adjust their schedules.

Academic Integrity

The University of Southern California is foremost a learning community committed to fostering successful scholars and researchers dedicated to the pursuit of knowledge and the transmission of ideas. Academic misconduct is in contrast to the university's mission to educate students through a broad array of first-rank academic, professional, and extracurricular programs and includes any act of dishonesty in the submission of academic work (either in draft or final form).

This course will follow the expectations for academic integrity as stated in the USC Student Handbook. All students are expected to submit assignments that are original work and prepared specifically for the course/section in this academic term. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s). Students suspected of engaging in academic misconduct will be reported to the Office of Academic Integrity.

Other violations of academic misconduct include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

The impact of academic dishonesty is far-reaching and is considered a serious offense against the university and could result in outcomes such as failure on the assignment, failure in the course, suspension, or even expulsion from the university.

For more information about academic integrity see the student handbook or the Office of Academic Integrity's website, and university policies on Research and Scholarship Misconduct.

Course Content Distribution and Synchronous Session Recordings Policies

USC has policies that prohibit recording and distribution of any synchronous and asynchronous course content outside of the learning environment.

Recording a university class without the express permission of the instructor and announcement to the class, or unless conducted pursuant to an Office of Student Accessibility Services (OSAS) accommodation. Recording can inhibit free discussion in the future, and thus infringe on the academic freedom of other students as well as the instructor. (Living our Unifying Values: The USC Student Handbook, page 13).

Distribution or use of notes, recordings, exams, or other intellectual property, based on university classes or lectures without the express permission of the instructor for purposes other than individual or group study. This includes but is not limited to providing materials for distribution by services publishing course materials. This restriction on unauthorized use also applies to all information, which had been distributed to students or in any way had been displayed for use in relationship to the class, whether obtained in class, via email, on the internet, or via any other media. (Living our Unifying Values: The USC Student Handbook, page 13).

Course Schedule

	Topics/Daily Activities	Readings/ Preparation	Deliverables
Week 1	**Introduction Machine Learning (ML) in Biology** Lecture 1: Overview of ML and its applications in biology Lecture 2: Introduction to biological data for ML and its challenges	Posted on Brightspace	(none)
Week 2	**Introduction to Python** Lecture 3: Fundamentals of Python Lecture 4: Python in data analysis	[JV] Ch.1	(none)
Week 3	**Regression** Lecture 5: Regression and its application in predicting protein variant effects Lecture 6: Python in ML modeling	[KM] Ch.1, Ch.11 [S1] [AG] Ch.1	Assignment #1
Week 4	**Advanced Regression** Lecture 7: Regularization techniques and additional biological applications Lecture 8: Optimization algorithms in ML	[KM] Ch.4.5, Ch.8.1	Assignment #2
Week 5	**Classification I** Lecture 9: Logistic regression and its application in predicting protein-DNA interactions and mechanisms. Lecture 10: Naive Bayes classification and its application in genetic variants and disease	[KM] Ch.10, [S2]	Assignment #3
Week 6	**Classification II** Lecture 11: sK-Nearest Neighbors (KNN) and its application to the aforementioned question Lecture 12: Decision Trees and its application to the aforementioned question	[KM] Ch.16.1, Ch.18.1	Assignment #4
Week 7	**Advanced Classification** Lecture 13: Support Vector Machines (SVM) and its application in protein function prediction Lecture 14: Review and more biological applications	[KM] 17.3 [S3]	Assignment #5

Week 8	**Clustering** Lecture 15: K-means Clustering and its application in analyzing differential gene expression in RNA-seq data Lecture 16: Hierarchical Clustering and its application to the aforementioned question	[KM] 21.3 [S4]	Assignment #6
Week 9	**Dimensionality Reduction** Lecture 17: Principal Component Analysis (PCA) and its application in analyzing single-cell transcriptomics Lecture 18: Other high dimensionality reduction techniques	[KM] Ch.20.1 [S4]	Assignment #7
Week 10	**Ensemble Learning** Lecture 19: Bagging and Boosting, Random Forest and its application in gene expression profiling Lecture 20: AdaBoost and its application in biomarker discovery	[KM] 18.2-5	Assignment #8
Week 11	**Deep Learning I** Lecture 21: Neural Networks and its application in predicting protein secondary structure Lecture 22: Implementing deep learning in Python	[KM] Ch.13.1-2 [S5]	Assignment #9
Week 12	**Deep Learning II** Lecture 23: Convolutional Neural Network (CNN) and its application in predicting protein residue-residue contact and distance Lecture 24: Recurrent Neural Network (RNN) in sequence data	[KM] Ch.14.1-2, Ch.15.1	Assignment #10
Week 13	**Special Topics** Lecture 25: Review of ML methods and introduction to Ethics in ML Lecture 26: Applications of ML in personalized medicine and ethical considerations	Posted on Brightspace	Final project
Week 14	**Emerging Technology I** Lecture 27: Introduction to Transformers in structural biology research Lecture 28: Guest speaker	Posted on Brightspace	Final project
Week 15	**Emerging Technology II** Lecture 29: Guest speaker Lecture 30: Introduction to Quantum Computing in biological research	Posted on Brightspace	Final project
Final	**Final project due in university assigned final exam period**		

Statement on Academic Conduct and Support Systems

Academic Integrity:
The University of Southern California is a learning community committed to developing successful scholars and researchers dedicated to the pursuit of knowledge and the dissemination of ideas. Academic misconduct, which includes any act of dishonesty in the production or submission of academic work, compromises the integrity of the person who commits the act and can impugn the perceived integrity of the entire university community. It stands in opposition to the university's mission to research, educate, and contribute productively to our community and the world.

All students are expected to submit assignments that represent their own original work, and that have been prepared specifically for the course or section for which they have been submitted. You may not submit work written by others or "recycle" work prepared for other courses without obtaining written permission from the instructor(s).

Other violations of academic integrity include, but are not limited to, cheating, plagiarism, fabrication (e.g., falsifying data), collusion, knowingly assisting others in acts of academic dishonesty, and any act that gains or is intended to gain an unfair academic advantage.

The impact of academic dishonesty is far-reaching and is considered a serious offense against the university. All incidences of academic misconduct will be reported to the Office of Academic Integrity and could result in outcomes such as failure on the assignment, failure in the course, suspension, or even expulsion from the university.

For more information about academic integrity see the student handbook or the Office of Academic Integrity's website, and university policies on Research and Scholarship Misconduct.

Please ask your instructor if you are unsure what constitutes unauthorized assistance on an exam or assignment, or what information requires citation and/or attribution.

Students and Disability Accommodations:

USC welcomes students with disabilities into all of the University's educational programs. The Office of Student Accessibility Services (OSAS) is responsible for the determination of appropriate accommodations for students who encounter disability-related barriers. Once a student has completed the OSAS process (registration, initial appointment, and submitted documentation) and accommodations are determined to be reasonable and appropriate, a Letter of Accommodation (LOA) will be available to generate for each course. The LOA must be given to each course instructor by the student and followed up with a discussion. This should be done as early in the semester as possible as accommodations are not retroactive. More information can be found at osas.usc.edu. You may contact OSAS at (213) 740-0776 or via email at osasfrontdesk@usc.edu.

Support Systems:

Counseling and Mental Health - (213) 740-9355 – 24/7 on call
Free and confidential mental health treatment for students, including short-term psychotherapy, group counseling, stress fitness workshops, and crisis intervention.

988 Suicide and Crisis Lifeline - 988 for both calls and text messages – 24/7 on call
The 988 Suicide and Crisis Lifeline (formerly known as the National Suicide Prevention Lifeline) provides free and confidential emotional support to people in suicidal crisis or emotional distress 24 hours a day, 7 days a week, across the United States. The Lifeline is comprised of a national network of over 200 local crisis centers, combining custom local care and resources with national standards and best practices. The new, shorter phone number makes it easier for people to remember and access mental health crisis services

(though the previous 1 (800) 273-8255 number will continue to function indefinitely) and represents a continued commitment to those in crisis.

Relationship and Sexual Violence Prevention Services (RSVP) - (213) 740-9355(WELL) – 24/7 on call
Free and confidential therapy services, workshops, and training for situations related to gender- and power-based harm (including sexual assault, intimate partner violence, and stalking).

Office for Equity, Equal Opportunity, and Title IX (EEO-TIX) - (213) 740-5086
Information about how to get help or help someone affected by harassment or discrimination, rights of protected classes, reporting options, and additional resources for students, faculty, staff, visitors, and applicants.

Reporting Incidents of Bias or Harassment - (213) 740-5086 or (213) 821-8298
Avenue to report incidents of bias, hate crimes, and microaggressions to the Office for Equity, Equal Opportunity, and Title for appropriate investigation, supportive measures, and response.

The Office of Student Accessibility Services (OSAS) - (213) 740-0776
OSAS ensures equal access for students with disabilities through providing academic accommodations and auxiliary aids in accordance with federal laws and university policy.

USC Campus Support and Intervention - (213) 740-0411
Assists students and families in resolving complex personal, financial, and academic issues adversely affecting their success as a student.

Diversity, Equity and Inclusion - (213) 740-2101
Information on events, programs and training, the Provost's Diversity and Inclusion Council, Diversity Liaisons for each academic school, chronology, participation, and various resources for students.

USC Emergency - UPC: (213) 740-4321, HSC: (323) 442-1000 – 24/7 on call
Emergency assistance and avenue to report a crime. Latest updates regarding safety, including ways in which instruction will be continued if an officially declared emergency makes travel to campus infeasible.

USC Department of Public Safety - UPC: (213) 740-6000, HSC: (323) 442-1200 – 24/7 on call
Non-emergency assistance or information.

Office of the Ombuds - (213) 821-9556 (UPC) / (323-442-0382 (HSC)
A safe and confidential place to share your USC-related issues with a University Ombuds who will work with you to explore options or paths to manage your concern.

Occupational Therapy Faculty Practice - (323) 442-2850 or otfp@med.usc.edu
Confidential Lifestyle Redesign services for USC students to support health promoting habits and routines that enhance quality of life and academic performance.