San José State University
Computer Science Department

CS286: Topics in Sequence-based Machine Learning for Bioinformatics, Spring 2021

Course and Contact Information

Instructor:	William “Bill” Andreopoulos
Office Location:	Online (former MacQuarrie Hall 416)
Email:	william.andreopoulos@sjsu.edu Please use Canvas Messaging and the Discussion Forum
Class Days/Time:	Tuesday and Thursday 16:30-17:45pm
Classroom:	Online via Zoom
Office Hours:	F 3:00-5:00 pm

Faculty Web Page and Canvas Messaging

Course materials such as syllabus, handouts, notes, assignment instructions, etc. can be found on Canvas Learning Management System course login website at http://sjsu.instructure.com. You are responsible for regularly checking with the Canvas messaging system to learn of any updates. You should modify the Canvas settings for notifications of announcements and discussion forum postings to be sent to you.

Course Description

Machine learning and deep learning applications to solving sequence analysis problems in molecular and cell biology.

Prerequisites

This course is offered to students enrolled in the MS Bioinformatics or MS Computer Science program at San Jose State University. MS Bioinformatics students must have completed BIOL 123B and MATH 162 with a grade of C- or better. MS Computer Science students must have completed either CS156 or CS171. All students should have some knowledge of Python programming.

Course Learning Outcomes (CLO)

Upon successful completion of this course, students will be able to:

1. Use machine learning and deep learning in bioinformatics sequence analysis to answer biological questions and to generate biological hypotheses.

2. Comprehend the nature, scope and limits of using machine learning and deep learning in the field of bioinformatics.
3. Develop machine learning and deep learning solutions for sequence data.
4. Compare different machine learning algorithms and choose a solution based on suitability for the particular data set.
5. Contrast how biomolecular information analysis with machine learning compares with use of classical bioinformatics tools.
6. Appreciate some of the most challenging problems in life sciences that use machine learning methods, possess insight into how to solve those problems.

Texts/Readings

We don't use a specific textbook in this class as there is lots of relevant material on bioinformatics dispersed. The reading material will be the slides, references and handouts.

A copy of my slides will be available to the students enrolled in the class.

Additional handouts will be provided through Canvas.

Major references:

- Walsh, Ian; Pollastri, Gianluca; Tosatto, Silvio C. E. (September 2016). "Correct machine learning on protein sequences: a peer-reviewing perspective". *Briefings in Bioinformatics*. 17(5): 831–840.
- Krallinger, Martin; Erhardt, Ramon Alonso-Allende; Valencia, Alfonso (March 2005). "Text-mining approaches in molecular biology and biomedicine". *Drug Discovery Today*. 10 (6): 439–445.
- Chicco, D (December 2017). "Ten quick tips for machine learning in computational biology". *BioData Mining*. 10 (35): 35.
- Yang, Yuedong; Gao, Jianzhao; Wang, Jihua; Heffernan, Rhys; Hanson, Jack; Paliwal, Kuldip; Zhou, Yaoqi (May 2018). "Sixty-five years of the long march in protein secondary structure prediction: the final stretch?". *Briefings in Bioinformatics*. 19 (3): 482–494.
- Larrañaga, Pedro; Calvo, Borja; Santana, Roberto; Bielza, Concha; Galdiano, Josu; Inza, Iñaki; Lozano, José A.; Armañanzas, Rubén; Santafé, Guzmán (March 2006). "Machine learning in bioinformatics". *Briefings in Bioinformatics*. 7 (1): 86–112.
- Mathé, Catherine; Sagot, Marie-France; Schiex, Thomas; Rouzé, Pierre (October 2002). "Current methods of gene prediction, their strengths and weaknesses". *Nucleic Acids Research*. 30 (19): 4103–4117.
- Wang, Sheng; Peng, Jian; Ma, Jianzhu; Xu, Jinbo (December 2015). "Protein secondary structure prediction using deep convolutional neural fields". *Scientific Reports*. 6: 18962.
- Pirooznia, Mehdi; Yang, Jack Y.; Yang, Mary Qu; Deng, Youping (2008). "A comparative study of different machine learning methods on microarray gene expression data". *BMC Genomics*. 9 (1): S13.
- d'Alché-Buc, Florence; Wehenkel, Louis (2008). "Machine Learning in Systems Biology". *BMC Proceedings*. 2 (4): S1.

Other technology requirements / equipment / material

Students will use colab.research.google.com and create Jupyter notebooks in Python to ensure their work is shareable and reproducible.

Course Requirements

SJSU classes are designed such that in order to be successful, it is expected that students will spend a minimum of forty-five hours for each unit of credit (normally three hours per unit per week), including preparing for class, participating in course activities, completing assignments, and so on.

Reading assignments: Readings will regularly be assigned for the next class (see schedule). Slides will be posted under the Canvas modules before the next class.

Hands-On Worksheets:
We will have a number of hands-on worksheets. The purpose of the hands-on exercises is to develop your understanding of the material and skills in using the tools. Many of the hands-on worksheets will involve use of bioinformatics tools.

The Hands-On worksheets will involve learning how to use machine learning and deep learning tools with the Python programming language for performing bioinformatics analysis. Students will use colab.research.google.com and create Jupyter notebooks in Python to ensure their work is shareable and reproducible.

Term Project and In-Class Presentation:
There will be a term project. This is a group project. Each group consists of two or three students. Team Formation is due on Thursday, February 11, 2021. A list of potential projects will be provided to you by the instructor.
A Progress Report is due on Thursday, March 11, 2021.
The final project is due on Tuesday, May 11, 2021.
The in-class presentation will also take place on May 11-13, 2021.
A grading rubric will be provided.
All homework should be submitted on Canvas, not by e-mail.

Exams:
Midterm Exam One: Thursday, March 11, 2021.
Midterm Exam Two: Thursday, April 22, 2021.
Final Exam: Friday, May 21, 2021.
The midterm exams are each one hour and fifteen minutes long. The final exam is two hours and fifteen minutes long.

The exams will contain multiple choice questions, true/false and short answer questions. Exams are *open book, open notes*, and comprehensive. The exams should be done individually and are not group work. No make-up exams except in case of verifiable emergency circumstances.

Presentation of a research paper:
Each student should present an influential bioinformatics research paper of his/her choice to one of the classes. Students should sign up in the given spreadsheet for a date to present a paper. The paper, chosen by the student, should either use machine learning/deep learning towards making a biological discovery or introduce a novel bioinformatics tool. The presentation should last for no more than 10 minutes followed by Q&A. A grading rubric will be provided.

Discussion Forum on Canvas

Students should use the Canvas Discussion Forum for all issues about the course. Regular participation is recommended. The instructor will open a graded thread for each module. Students must ask at least 1 original question and provide at least 1 original answer in each graded thread. Discussion Forum participation counts for 5% of the course grade.

Determination of Grades

The course grade is based on:

Hands-On Worksheets: 20%
Midterms: 20%
Final: 20%
Project: 25%
Participation in the discussion forum: 5%
Presentation of a research paper: 10%

Grade	*Points*	*Percentage*
A plus	*960 to 1000*	*96 to 100%*
A	*930 to 959*	*93 to 95%*
A minus	*900 to 929*	*90 to 92%*
B plus	*860 to 899*	*86 to 89 %*
B	*830 to 859*	*83 to 85%*
B minus	*800 to 829*	*80 to 82%*
C plus	*760 to 799*	*76 to 79%*
C	*730 to 759*	*73 to 75%*
C minus	*700 to 729*	*70 to 72%*
D plus	*660 to 699*	*66 to 69%*
D	*630 to 659*	*63 to 65%*
D minus	*600 to 629*	*60 to 62%*

Communication with the instructor

Questions for the instructor may be asked during Zoom class meetings or office hours, or at any time via the Canvas Discussion Forum or Canvas messaging. Announcements of general interest will be posted under Announcements on Canvas. Questions about worksheet-specific code should be asked during class meeting time, not by email.

Class Attendance

Class attendance (via Zoom) is highly recommended. Classes will be recorded as Zoom screencasts and posted on Canvas. Students are responsible for all material presented in all classes.

Regrading Procedure

Grades assigned are final, unless there was an error in the grading. Students may request regrading by filling out a Regrade Request form on Canvas. Regrading may result in a lower grade.

Classroom Protocol

Students should be muted when not speaking, and must be dressed appropriately when their camera is on.

Add/Drop Policy

For those wishing to add this course, the deadline is January 26, 2021. The last day to drop a course without a “W” grade is February 8, 2021. To drop after this date, a Late Drop petition will be required. According to University and Department guidelines, dropping after February 8, 2021, requires a serious and compelling reason to drop a course. Grades alone do not constitute a reason to drop a course. Students who stop attending without officially dropping will be issued a “U” at the end of the semester which is counted as an F in calculations of GPA.

Students are responsible for understanding the policies and procedures about add/drop, grade forgiveness, etc. Refer to the current semester's Catalog Policies section at http://info.sjsu.edu/static/catalog/policies.html. Add/drop deadlines can be found on the current academic year calendars document on the Academic Calendars webpage at http://www.sjsu.edu/provost/services/academic_calendars/. The Late Drop Policy is available at http://www.sjsu.edu/aars/policies/latedrops/policy/. Students should be aware of the current deadlines and penalties for dropping classes. Information about the latest changes and news is available at the Advising Hub at http://www.sjsu.edu/advising/.

Consent for Recording of Class and Public Sharing of Instructor Material

University Policy S12-7, http://www.sjsu.edu/senate/docs/S12-7.pdf, requires students to obtain instructor's permission to record the course. Common courtesy and professional behavior dictate that you notify someone when you are recording him/her. You must obtain the instructor's permission to make audio or video recordings in this class. Such permission allows the recordings to be used for your private, study

purposes only. The recordings are the intellectual property of the instructor; you have not been given any rights to reproduce or distribute the material.

Course material developed by the instructor is the intellectual property of the instructor and cannot be shared publicly without his/her approval. You may not publicly share or upload instructor generated material for this course such as exam questions, lecture notes, hands-on exercises or homework solutions without instructor consent.

Academic Integrity

Your commitment as a student to learning is evidenced by your enrollment at San Jose State University. The University Academic Integrity Policy S07-2 at http://www.sjsu.edu/senate/docs/S07-2.pdf requires you to be honest in all your academic course work. Faculty members are required to report all infractions to the office of Student Conduct and Ethical Development. The Student Conduct and Ethical Development website is available at http://www.sjsu.edu/studentconduct/. Instances of academic dishonesty will not be tolerated. Cheating on exams or plagiarism (presenting the work of another as your own, or the use of another person's ideas without giving proper credit) will result in a failing grade and sanctions by the University. For this class, all assignments are to be completed by the individual student unless otherwise specified. If you would like to include your assignment or any material you have submitted, or plan to submit for another class, please note that SJSU's Academic Integrity Policy S07-2 requires approval of instructors.

- Anyone caught cheating (including sharing answers with others during exams) in the class will receive a failing grade on the exam or assignment, in addition to other sanctions that are permitted by the University, including but not limited to the filing of a report with the Dean of Student Services and expulsion from the University.

Campus Policy in Compliance with the American Disabilities Act

If you need course adaptations or accommodations because of a disability, or if you need to make special arrangements in case the building must be evacuated, please make an appointment with me as soon as possible, or see me during office hours. Presidential Directive 97-03 at http://www.sjsu.edu/president/docs/directives/PD_1997-03.pdf requires that students with disabilities requesting accommodations must register with the Accessible Education Center (AEC) at http://www.sjsu.edu/aec to establish a record of their disability.

In 2013, the Disability Resource Center changed its name to be known as the Accessible Education Center, to incorporate a philosophy of accessible education for students with disabilities. The new name change reflects the broad scope of attention and support to SJSU students with disabilities and the University's continued advocacy and commitment to increasing accessibility and inclusivity on campus.

University Policies

Per University Policy S16-9, university-wide policy information relevant to all courses, such as academic integrity, accommodations, etc. will be available on Office of

Graduate and Undergraduate Programs' Syllabus Information web page at http://www.sjsu.edu/gup/syllabusinfo/

CS286: Topics in Sequence-based Machine Learning for Bioinformatics, Spring 2021

The schedule is subject to change with fair notice.

Course Schedule

Week	Topic
01/26	Overview of unsupervised and supervised ML in bioinformatics
02/02	Sequence classification with Logistic Regression
02/09	Sequence classification with Naïve Bayes
02/16	Language models using k-mers
02/23	Error correction in DNA sequences
03/02	Vector space representations: clustering & visualization with PCA, t-SNE, UMAP
03/09	Review for midterm with problem-solving exercises / *Midterm 1*
03/16	Hidden Markov Models and Markov chains
03/23	Efficient sequence searching
03/29-04/02	Spring recess
04/06	Word embeddings with neural networks
04/13	Recurrent Neural Networks (RNNs) for sequence modelling
04/20	Review for midterm with problem-solving exercises / *Midterm 2*
04/27	Long Short Term Memory (LSTM) neural networks for sequence prediction
05/04	Case studies using deep learning / Project discussion
05/11	Project presentations
05/21	**Final exam – Friday, May 21, 14:45-17:00pm**