San José State University
Computer Engineering Department
CMPE/SE 188, Machine Learning for Big Data, Section 01, Fall 2019

Course and Contact Information

Instructor:	Magdalini Eirinaki
Office Location:	ENG 283F
Telephone:	(408) 924 - 3828
Email:	magdalini.eirinaki@sjsu.edu
Office Hours:	Tuesday, 1-3 pm Always check with the CMPE website for the most up to date office hours at http://cmpe.sjsu.edu/content/office-hours.
Class Days/Time:	Tuesday & Thursday, 10:30am – 11:45am
Classroom:	ENG 325
Prerequisites:	CMPE 126 (for BS CMPE students) or CS 146 (for BS SE students)

Course Format

Technology Intensive, Hybrid, and Online Courses

This course requires the student to have a personal laptop that is installed with a modern operating system. The lectures will be delivered in the classroom, however the students might be asked to use their laptops or smart devices during the class, or offline in order to participate in the class assignments.

Faculty Web Page and MYSJSU Messaging

Course materials such as syllabus, handouts, notes, assignment instructions, etc. can be found on Canvas Leaning Management System course login website at http://sjsu.instructure.com. You are responsible for daily checking with the messaging system through Canvas and MySJSU at http://my.sjsu.edu to learn of any updates.

Course Description

Introduction to machine learning and pattern recognition for big data analytics; machine learning concepts, theories, approaches, algorithms, and big data analytic applications; supervised learning, unsupervised learning, and learning theory.

Course Goals

This course focuses on machine learning algorithms and methodologies to support large-scale data analysis. The course covers fundamental machine learning algorithms and techniques, such as regression, classification, and clustering models, as well as more contemporary ones, including collaborative filtering and social network analysis techniques. The course will also review techniques that allow for scalability and processing of large amounts of data, such as parallelization models, hashing, and dimensionality reduction techniques.

This course involves a group-based term project to provide students with the opportunity to build a simplified data or web mining application, and to enhance their professional engineering skills including practical application of state-of-the-art big data and machine learning tools and frameworks, teamwork, technical leadership, and effective communication skills (both written and verbal).
The course also includes a set of individual assignments and survey projects to enable students to deepen their knowledge on the material.

Course Learning Outcomes (CLO) (Required)

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

1. CLO 1- Describe the fundamental concepts of several machine learning algorithms and techniques.
2. CLO 2 – Demonstrate an understanding of and ability to use emergent big data technologies.
3. CLO 3 – Explain how appropriate machine learning approaches and techniques can be applied to solve given problems.
4. CLO 4 – Use machine learning models, methods, and big data technology and tools to complete a given big data analytics project

Required Texts/Readings

This class does not have a single textbook. Instead, the students have to study material coming from various books, papers and other resources, all of which are free to download (for academic use). It is each student's responsibility to consult with the updated syllabus on Canvas in order to identify which readings cover the concepts that are taught each week.

A list of reference textbooks is also provided for those who'd like to get some background knowledge or seek more details on the topics covered in class.

Textbooks

[HKP] *Data Mining: Concepts and Techniques,* by Jiawei Han, Micheline Kamber and Jian Pei
Morgan Kaufmann, Elsevier Inc. (2011), ISBN: 9780123814791
(available as ebook from the SJSU Library)

[ISLR] *An Introduction to Statistical Learning with Applications In R*, by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer Texts in Statistics, 2013 (download from http://www-bcf.usc.edu/~gareth/ISL/)

Other Readings

- Papers, tutorial slides, articles and all other material that will be made available via Canvas
- Lecture slides (available via Canvas)

Reference textbooks

Data Mining, The Textbook, by Charu C. Aggarwal
Springer (2015), ISBN: 9783319381169

Recommender Systems: The textbook, by Charu Aggrawal
Springer, 2016, ISBN 978-3-319-29659-3
(available as ebook from the SJSU Library)

Machine Learning, by Tom M. Michell,
McGraw Hill (1997), ISBN: 0070428077

Mining of Massive Datasets, by Jure Leskovec, Anand Rajaraman and Jeffrey Ullman, 2nd edition, Cambridge University Press, December 2014 (download from http://infolab.stanford.edu/~ullman/mmds/book.pdf)

Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, by Bing Liu
Springer (2011), ISBN: 3540378812

Social Media Mining, An Introduction, by Reza Zafarni, Mohammad Ali Abasi, Huan Liu
Cambridge University Press, 2014 (available to download at: http://dmml.asu.edu/smm/SMM.pdf)

Other technology requirements / equipment / material

Most programming assignments will be performed using Python (sklearn and networkx libraries). Other programming languages, platforms, software applications and tools, such as Spark, Mahout, Gephi, Tableau, etc. that will be required for this class are either free to download, or the instructor will provide the students with academic licenses. Students will be informed in class and via Canvas ahead of time in order to install all required software.

Course Requirements and Assignments

Success in this course is based on the expectation that students will spend, for each unit of credit, a minimum of 45 hours over the length of the course (normally three hours per unit per week) for instruction, preparation/studying, or course related activities, including but not limited to internships, labs, and clinical practica. Other course structures will have equivalent workload expectations as described in the syllabus.

Student Assessment

In-class activities	5%
Individual homework assignments	5%
Quizzes	10%
Programming assignment/Competition	10%
Term Project	20%
Midterm Exam	20%
Final Exam (comprehensive)	30%

Descriptions of Assignments/Exams

In-class activities: Students will be evaluated based on their participation in in-class assignments. All students are required to write their names on the submitted work and/or submit their answers online using their unique IDs, shared with the instructors. Failing to do so, even if the student was indeed present in the class, will result in zero credit as the instructor is unable to verify the student's claims. Moreover, students whose names appear on submitted work, but were not in class, as well as the students who submitted their name on their behalf are violating the academic integrity policy and will be reported immediately to the office of Student Conduct and Ethical Development. As the name implies, credit will be given only to those present when the activity took place in the classroom. No make-ups or remote participation is allowed.

Individual Written/Programming Assignments and Quizzes: Students will be provided with details describing the assignments and how they will be graded every week. These assignments will be in-class or take-home written assignments, in-class or take-home lab assignments, and/or presentation assignments for research papers or articles. Students will also have to answer to quizzes that will be based on the homework assignment that is

due that day. The worst quiz grade will not be counted towards the final pop quiz grade of each student ("worst-one out policy").

Programming Assignment/Competition: Students will participate in one or more competition-like programming assignments related to the contents of the class. The students will have to implement a solution to a given problem and will be evaluated against a baseline with a given related metric. An accompanying report will be submitted. Credit will be given to those who successfully beat the baseline. Extra credit might be given to top-performing submissions.

Term Project: Groups of 3 students will be formed to work on a term-long group project related to data or web mining. The project has deliverables throughout the semester. The quality and completeness of all the deliverables will be considered in grading the projects. All projects will be demonstrated in class. The project details will be announced by the instructor and posted on the course's web site well before the deadlines.
Each group member is expected to participate in every phase of the project. The final grade of each member will be proportional to his/her participation in the group, as assessed by the instructor and the student's peers. Each member should be able to answer questions regarding the project, present some part of the project demo, and participate in the system implementation and the writing of the technical reports. The term project will be graded on the basis of the following three components: a) project implementation, b) project report, c) project demonstration. Grading will be rubric-based.

Exams: Exams will be a combination of multiple choice and short answer questions and will be based on the individual assignments and course material covered in class.

Final Examination or Evaluation

The class has a final examination. The final exam is comprehensive and the date is determined by the University's Final Examination Schedule.

Grading Information

Determination of Grades

The final grades will be calculated based on the following:

A+(plus) = 98% to 100%
A = 94% to 97.9%
A-(minus) = 90% to 93.9%
B+ = 85% to 89.9%
B = 75% to 84.9%
B- = 70% to 74.9%
C+ = 68% to 69.9%
C = 64% to 67.9%
C- = 60% to 63.9%
D = 50% to 59.9%
F = less than 50%

- No late assignments will be accepted. An extension will be granted only if a student has serious and compelling reasons that can be proven by an independent authority (e.g. doctor's note if the student has been sick).
- The exam dates are final.

All students have the right, within a reasonable time, to know their academic scores, to review their grade-dependent work, and to be provided with explanations for the determination of their course grades.

Classroom Protocol

You are expected to arrive in time for class. While in class you need to turn off your cellphone unless directed otherwise by your instructor. Laptop/tablet/smart phone use is allowed only for activities related to the class. Please be considerate of your fellow students.

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/". Make sure to visit this page, review and be familiar with these university policies and resources.

Department Policies

- Students who do not provide documentation of having satisfied the class prerequisite or co-requisite requirements (if any) by the second class meeting will be dropped from the class.
- All non-proctored report (or similarly sized) assignments in courses where some of the final grade depends on prose writing will be submitted to Turnitin.com.
- Major exams in this class may be video recorded to ensure academic integrity. The recordings will only be viewed if there is an issue to be addressed. Under no circumstances will the recordings be publicly released.

CMPE/SE 188 / Machine Learning for Big Data, Fall 2018, Course Schedule

The schedule (and related dates/readings/assignments) is tentative and subject to change with fair notice. In case of guest lectures the syllabus will be updated accordingly. Any changes will be announced in due time in class and on the course's web site (Canvas). The students are obliged to consult the most updated and detailed version of the reading material and syllabus, which will be posted on Canvas.

Course Schedule

Week	Date	Topics, Readings, Assignments, Deadlines
1	8/22	Introduction to CMPE 188
2	8/27	Introduction to Machine Learning
2 - 3	8/29	Data preparation
	9/3	
	9/5	
4 - 7	9/10	Supervised learning: Prediction & Classification – Simple Linear Regression, Decision Trees, K-NN, Bayesian, Issues
	9/12	
	9/17	
	9/19	
	9/24	
	9/26	
	10/1	
7	10/3	MIDTERM
8 - 9	10/8	Supervised learning: Prediction & Classification (cont'd) – Random Forests, Ensemble methods, Evaluation Scaling for big data Case study: Bot or not?
	10/10	
	10/15	
9 - 10	10/17	Unsupervised Learning: Clustering – K-Means, Evaluation Association Rules Mining (if time allows)
	10/22	
	10/24	
11-13	10/29	Recommendation systems – Content-based Collaborative Filtering, User- and Item-based Collaborative Filtering, Scaling for big data: Latent factor Collaborative Filtering/Matrix Factorization, Evaluation methods Case study: Netflix
	10/31	
	11/5	
	11/7	
	11/12	
13 -14	11/14	Advanced topics (Social Network Analysis / Deep Learning – as time allows)
	11/19	

Week	Date	Topics, Readings, Assignments, Deadlines
14-16	11/21	Project Presentations (11/28: Thanksgiving holiday)
	11/26	
	12/3	
	12/5	
Finals	Friday, 12/13	FINAL EXAM 9:45am – 12noon