San José State University
Computer Engineering Department
CMPE 255, Data Mining, Section 1, Fall 2017

Course and Contact Information

Instructor: David C. Anastasiu

Office Location: ENG 179

Telephone: (408) 924-2938

Email: david.anastasiu@sjsu.edu

Office Hours: TBA

Always check with the CMPE website for the most up to date office hours, at https://cmpe.sjsu.edu/content/office-hours

Student Assistant: TBA

Class Days/Time: TBA

Classroom: TBA

Prerequisites:

Basic understanding of Linear Algebra.
Familiarity with Python or R, and a lower-level programming language (C, C++, or Java).
Classified graduate standing or instructor consent.

Course Format

This course requires the student to have a personal computer 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 at http://sjsu.instructure.com. You are responsible for regularly checking with the messaging system through MySJSU at http://my.sjsu.edu and Canvas to learn of any updates.

Course Description

Data representation and preprocessing, proximity, finding nearest neighbors, dimensionality reduction, exploratory analysis, association analysis and sequential patterns, supervised inference and prediction, classification, regression, model selection and evaluation, overfitting, clustering, advanced topics.

Learning Outcomes

Upon successful completion of this course, students will:

1. Be able to demonstrate an understanding of advanced knowledge of the practice of computer/software engineering, from vision to analysis, design, validation and deployment.
2. Be able to tackle complex engineering problems and tasks, using contemporary engineering principles, methodologies and tools.
3. Be able to demonstrate leadership and the ability to participate in teamwork in an environment with different disciplines of engineering, science and business.
4. Be aware of ethical, economic and environmental implications of their work, as appropriate.
5. Be able to advance successfully in the engineering profession, and sustain a process of life-long learning in engineering or other professional areas.
6. Be able to communicate effectively, in both oral and written forms.

Course Learning Outcomes (CLO)

The main focus of this course is on data mining and its applications. More specifically, we will cover a broad range of data mining algorithms and techniques, focusing on how they are used in diverse applications such as predicting future income, detecting and preventing spam, finding genes with similar functions, etc. The lectures will revolve around the fundamental concepts of the areas covered as well as case studies, providing students the opportunity to gain a deep understanding and apply these concepts in real-life scenarios. Multiple in-class and homework assignments required throughout the class will test the students' ability to effectively harness the power of data mining in different scenarios. Although some data analytics and mining tools will be used for demonstration purposes, mastering specific tools is not the primary objective of the lectures. Instead, the students will be able to gain hands-on experience on such technologies via individual and group-based projects, where they will be expected to perform in-depth analysis using real-world data, and to enhance their professional engineering skills, including teamwork, technical leadership, and effective communication skills (both written and verbal).

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

1. Identify the proper techniques and algorithms needed to prepare, preprocess, analyze and mine highly unstructured datasets, such as large text collections.
2. Understand advantages and disadvantages of dimensionality reduction as a preprocessing tool for data analysis.
3. Identify the correct class of methods that can be used to efficiently build a nearest neighbor graph, given properties of the input data.

4. Discuss and apply fundamental data mining concepts and techniques, such as classification, clustering, or frequent pattern mining.
5. Quickly get accustomed to any data mining/data analysis software application and be able to use it.
6. Gain hands-on experience by performing an extensive analysis using data mining techniques, in individual and group projects.
7. Effectively present and communicate the knowledge they have acquired in the course.

Required Texts/Readings

This class does not have a required textbook. Lecture slides and selected readings from online materials will cover most of the topics. A list of reference textbooks is also provided for those who would like to get some background knowledge or seek more details on the topics covered in class.

It is each student's responsibility to consult with the updated syllabus on Canvas to identify which readings cover the concepts that are taught each week.

Online Materials

[A] *Data Mining: The Textbook*, by Charu C. Aggarwal, Springer, May 2015 (Springer: http://www.springer.com/us/book/9783319141411, Amazon: https://www.amazon.com/dp/3319141414/, Amazon other: https://www.amazon.com/gp/offer-listing/3319141414/condition=all)

[TSK] *Introduction to Data Mining*, by Pang-Ning Tan, Michael Steinbach, Vipin Kumar, ISBN-13: 9780321335661 (download Ch. 4, 6, and 8 from http://www-users.cs.umn.edu/~kumar/dmbook/index.php)

[JWHT] *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 other reading materials will be made available via Canvas.

Reference textbooks (not required)

[TSK] *Introduction to Data Mining*, by Pang-Ning Tan, Michael Steinbach, Vipin Kumar, ISBN-13: 9780321335661 (remaining chapters)

[LRU] *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)

[ZM] *Data Mining and Analysis: Fundamental Concepts and Algorithms*, by Mohammed J. Zaki, Wagner Meira, Jr., Cambridge University Press, May 2014 (download from http://www.dataminingbook.info/pmwiki.php/Main/BookDownload)

[HKPK] *Data Mining: Concepts and Techniques*, by Jiawei Han, Micheline Kamber and Jian Pei, Morgan Kaufmann, Elsevier Inc. (2011) (2nd edition also acceptable)

[DBR] *Open Intro Statistics*, by David M. Diez, Christopher D. Barr, and Mine Cetinkaya-Rundel, 3rd edition, 2015 (https://www.openintro.org/stat/textbook.php)

Other technology requirements / equipment / material

Programming languages, platforms, as well as software applications and tools, such as Spark, Mahout, R/RStudio, Python, etc. that will be required for this class are either free to download. Specifically, the class will use Python and Jupyter notebooks to allow for more interactivity in the class. Installation instructions will be available on Canvas. Python is used in most in-class activities, homework assignments and projects. Students more comfortable with R instead of Python may request special permission to use R in homework assignments.

Course Requirements and Assignments

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. More details about student workload can be found in University Policy S12-3 at http://www.sjsu.edu/senate/docs/S12-3.pdf.

Descriptions of Assignments/Exams

In-class & social network assignments: Students will be evaluated based on their participation in in-class and social network 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 name appears 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.

Short story assignment: Each student will be assigned a week during which they must find, and present in class, a news story (e.g. about a new algorithm/technology etc.) that is related to the topics covered in class. This story must have some technical background and the student will be asked to elaborate and analyze the elements pertaining to the class as well as carry a discussion with the rest of the group based on it. The student's peers will be able to provide feedback, which will be taken into account for the grade.

Homework assignments and pop quizzes: Students will be provided instructions describing the assignments and how they will be graded. These assignments will be in-class or take-home written assignments, or presentation assignments for research papers or articles. Students will also have to answer to pop quizzes that will be based on the homework assignment that is due that day. The worst pop quiz grade will not be counted towards the final pop quiz grade of each student ("worst-one out policy").

Programming assignments: Students will be provided instructions describing programming assignments related to specific data mining problems. Generally, there are many ways to solve a problem. As such, students are encouraged to try multiple ways and compete against their peers to achieve the best result. Students submitting the top 3 scoring programs and most interesting solution to the problem will be awarded extra credit points.

Term project: Groups of 3 students will be formed to work on a term-long group project related to data 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 based on the following three components: a) project implementation, b) project report, c) project demonstration.

Extra-credit project: Students may choose to work on an **optional** individual extra-credit project designed to compare the efficiency of data mining algorithms. The project will require programming in at least one low-level programming language (C, C++, or Java) and one scripting/high-level programming language (Python, Pearl, R, Matlab, etc). Students will be required to work on their own and will not be allowed to share their solutions with others. The extra-credit project will be graded based on the following three components: a) project implementation, b) solution efficiency, c) project report.

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.

NOTE that University policy F69-24 at http://www.sjsu.edu/senate/docs/F69-24.pdf states that "Students should attend all meetings of their classes, not only because they are responsible for material discussed therein, but because active participation is frequently essential to insure maximum benefit for all members of the class. Attendance per se shall not be used as a criterion for grading."

Students cannot take this class without fulfilling its prerequisite or obtaining instructor approval. Please note that, according to department policy, "*students who do not provide documentation of having satisfied the class prerequisite and co-requisite requirements (if any) by the second class meeting will be dropped from the class.*"

Final Examination or Evaluation

This course has a comprehensive final exam.

Grading Information

The final grades will be calculated based on the following:

(A+)	>= 98
(A)	>= 94 and <98
(A-)	>= 90 and <94
(B+)	>= 85 and <90
(B)	>= 75 and <85
(B-)	>= 70 and <75
(C+)	>= 68 and <70
(C)	>= 64 and <78
(C-)	>= 60 and <64
(D)	>= 50 and <60
(F)	< 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.

Determination of Grades

The percentage weight assigned to class assignments is listed below. Detailed grading rubrics for the short-story assignment, team and extra-credit projects, and exact due dates for each assignment will be posted on Canvas. Students will have at least one week to complete each homework assignment.

In-class & online activities	5%
Short story assignment	5%
Homework assignments	5%
Pop quizzes	10%
Programming assignments	10%
Term project	20%
Extra-credit project (*optional*)	5%
Midterm exam	15%
Final exam (comprehensive)	30%

Classroom Protocol

Students are expected to arrive in time for class. While in class, they need to turn off cellphones unless directed otherwise by the instructor. Laptop/tablet/smart phone use is allowed only for activities related to the class.

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/"

CMPE 255-01 / Data Mining, Fall 2017, 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

Wk	Date	Topics	Readings	Deadlines
1	8/24	Introduction to Data Mining	HKP:1	
1	8/29	Data & Proximity	HKP:2-3	
2	8/31	---''---		
2	9/5	---''---		9/6 Drop Deadline
3	9/7	---''---		HW 1
3	9/12	Finding Nearest Neighbors	LRU:3, various papers	9/13 Add deadline
4	9/14	---''---		
4	9/19	---''---		
5	9/21	Dimensionality Reduction	various papers	
5	9/26	---''---		PR 1
6	9/28	---''---		
6	10/3	Association Analysis	HKP:6-7, TSK:6	Project proposal
7	10/5	---''---		
7	10/10	---''---		HW 2
8	10/12	*Midterm exam*		
8	10/17	Classification & Regression	HKP:8-9, TSK:4, skim JWHT:3-4&9	
9	10/19	---''---		
9	10/24	---''---		
10	10/26	---''---		
10	10/31	---''---		
11	11/2	Clustering	HKP:10-11, TSK:8	PR 2
11	11/7	---''---		
12	11/9	---''---		
12	11/14	---''---		

Wk	Date	Topics	Readings	Deadlines
13	11/16	---’’---		HW 3
13	11/21	Text Mining	skim MRS:1-2,6&19	
14	11/28	Graph Mining Overview	HKP:13, MRS:21	Project report, code & slides
14	11/30	Advanced topics		PR 3
15	12/1	Scaling Data Mining Methods		
15	12/5	Group Project Presentations		
16	12/7	Group Project Presentations		
	TBA	*Final Exam*, TBA		