DSC 232R: BIG DATA ANALYSIS USING SPARK

Winter 2026

Course Syllabus

Instructional Team Information

Instructor:: **Edwin Solares**
Email: **esolares@ucsd.edu**
Office hours: **TBD**

Course Description

This course has two main goals: The first is providing an introduction to using large scale data analysis frameworks (Spark, XGBoost and others). This includes the underlying computer architecture and the programming abstractions. The second is to combine methods from statistics and machine learning to perform large scale analysis, identify statistically significant patterns, and visualize statistical summaries.

Course Objectives

At the end of the course, students will be able to:

- Program Spark using Psyspark
- Identify computational bottlenecks in big data analysis
- Apply principal component analysis (PCA) to weather data
- Understand different machine learning methodologies
- Study the stability of learning algorithms

Course Format

This is a 10-week course that is asynchronous in that you are not required to attend any sessions during the week. All your core lectures are pre-recorded for you to view, by Professor Yoav Freund (UCSD professor and designer of this course). Additional recorded resources will be provided by your current instructor, both ahead of time and throughout the quarter. There will also be lots of opportunities to connect with your instructional team, both live over Zoom for office hours and weekly live discussions, and in different asynchronous discussion (Piazza and Discord).

Curriculum

Topic	Content
Engineering Big Data	• I/O limited computation and the memory bottlenecks • Data parallel computation in the cloud • Map-reduce, NFS, Spark • DataFrames and SQL in Spark
Analyzing Big Data	• Unsupervised Learning: PCA • Analysis of NOAA historical weather data using PCA • Visualizing and understanding PCA eigen-decomposition • Mapping weather information using iPyLeaflet
Classification	• Decision trees, bagging, and random forests • Boosting
Machine Learning Paradigms	• RMS Methodology • The strange behaviors of high dimensional data • Generative vs. discriminative learning • Learning by elimination
Quantifying Stability and Overfitting	• k-fold cross validation • The bootstrap • Margins • Easy and hard examples

Learning Materials

Required: Lecture Videos

Each week includes a series of lecture videos divided into two classes, each covering different topics. Each class's lectures are required, and it is recommended you view them in order. All information needed to complete the class can be found in these lectures.

Resource Library

You can access the course's Resource Library from the home page of this course. The Resource Library contains all the course lecture videos organized by module, as well as resources to help reinforce understanding foundational concepts which students need to understand to be successful in the course.

Assessments, Evaluation, and Grading

Quizzes

There will be 15 quizzes, all offered through Canvas and graded by the course TA. Not every week/class has a quiz. This means modules will contain up to 1 quiz per week. Quizzes for each week will become available on Mondays at 12:00 a.m. PST and will be due the following Sunday at 11:59 p.m. PST.

Students will have three chances to take the quiz and the highest of your three scores will be recorded. Quiz answers will be available at midnight the day after they are due, so no late quizzes will be accepted.

The quizzes together will represent 10% of the final grade. However, each quiz is graded out of a different number of points reflecting the content of the quiz.

Assignments

There will be 4 graded assignments. Assignments will be in the form of Jupyter Notebooks / NBGrader assignments offered through a tool called Vocareum.

Group Project

You will work in groups of up to 5 students to complete an independent project using a public dataset of your choosing that you will post to a website called Kaggle for review. This will be a competition, where the top 3 students (voted on by students) will receive extra credit.

Final

A final exam will be offered the last week of the class. This is a timed exam, and you will have 3 hours to complete it. It will be offered through Vocareum, similarly to your assignments.

Grading

Activity	Percent
Programming Assignments	30%
Quizzes	10%
Group Project	30%
Final Exam	30%

Grade	Range
A+	100
A	95-99
A-	90-94
B+	87-89
B	83-86
B-	80-82
C+	77-79
C	73-76
C-	70-72
F	<70

Course and UCSD Policies

UCSD Code of Conduct

All participants in the course are bound by the **University of California Code of Conduct** (https://aisc.uci.edu/students/index.php)

Netiquette

Be respectful. Be sensitive. Be aware. Effective written communication and open academic dialogue are crucial for sustaining a learning community that is respectful, considerate, relevant, creative, and thought-provoking. In an online classroom, expressions, meaning, and tone can quickly be taken out of context, making it imperative that online learners adhere to the communication guidelines below:

- Treat your classmates with respect.
- Be thoughtful and open in a discussion.
- Be aware and sensitive to different perspectives.
- Build one another up and encourage one another to succeed.

The following behavior should be avoided:

- Using insulting, condescending, or abusive words.
- Using all capital letters, which comes across as SHOUTING.
- Contacting learners or posting advertisements and solicitations.
- Posting copyrighted material.

Academic Integrity

Academic Integrity is expected of everyone at UC San Diego. This means you must be honest, fair, responsible, respectful, and trustworthy in your actions.

Lying, cheating, or other forms of dishonesty will not be tolerated because they undermine learning and the University's ability to certify students' knowledge and abilities. Thus, any attempt to get, or help another get, a grade by cheating, lying, or dishonesty will be reported to the Academic Integrity Office and result in sanctions. Sanctions can include an F in the class and suspension or dismissal from the University.

So, think carefully before you act. Before you act, ask yourself the following questions: a: is my action honest, fair, respectful, responsible, and trustworthy, and b) is my action authorized by the instructor? If you are unsure, don't ask a friend; ask your instructor, instructional assistant, or the Academic Integrity Office. You can learn more about academic integrity at academicintegrity.ucsd.edu.

Accessibility

Students requesting accommodations for this course due to a disability must provide a current Authorization for Accommodation (AFA) letter issued by the UC San Diego Office for Students with Disabilities (OSD) which is located in University Center 202 behind Center Hall. Students are required to present their AFA letters to Faculty (please make arrangements to

contact me privately) and to the OSD Liaison in the department in advance so that accommodations may be arranged. Contact the OSD for further information: https://disabilities.ucsd.edu/.

Religious Accommodation

See: EPC Policies on Religious Accommodation, Final Exams, Midterm Exams

It is the policy of the university to make reasonable efforts to accommodate students having bona fide religious conflicts with scheduled examinations by providing alternative times or methods to take such examinations. If a student anticipates that a scheduled examination will occur at a time at which his or her religious beliefs prohibit participation in the examination, the student must submit to the instructor a statement describing the nature of the religious conflict and specifying the days and times of conflict.

For final examinations, the statement must be submitted no later than the end of the second week of instruction of the quarter.

For all other examinations, the statement must be submitted to the instructor as soon as possible after a particular examination date is scheduled.

If a conflict with the student's religious beliefs does exist, the instructor will attempt to provide an alternative, equitable examination that does not create undue hardship for the instructor or for the other students in the class.

Accessibility

See: Nondiscrimination Policy Statement

CARE at the Sexual Assault Resource Center
858.534.5793 | sarc@ucsd.edu | https://care.ucsd.edu
Counseling and Psychological Services (CAPS)
858.534.3755 |https://caps.ucsd.edu

Subject to Change Policy

Note that the information contained in the course syllabus, other than the grade and absence policies, may be – under certain circumstances such as a modification to enhance student learning – subject to change with reasonable advance notice, as deemed appropriate by the instructor.