CSC 466: Knowledge Discovery from Data (KDD)  
Fall 2023  
Course Syllabus  

September 21, 2023

Instructor: Alex Dekhtyar  
email: dekhtyar@calpoly.edu  
office: 14-212 (new office location)

<table>
<thead>
<tr>
<th>What</th>
<th>Who</th>
<th>When</th>
<th>Where</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture</td>
<td>Section 01</td>
<td>MWF</td>
<td>2:10 pm – 3:00pm</td>
</tr>
<tr>
<td>Lab</td>
<td>Section 01</td>
<td>MWF</td>
<td>3:10pm – 4:00pm</td>
</tr>
<tr>
<td>Final Exam</td>
<td>Section 01</td>
<td>December 15 (Friday)</td>
<td>1:10 – 4:00pm</td>
</tr>
</tbody>
</table>

Note: the class will not have a written final exam. There may be final project presentations scheduled for the finals week.

**Office Hours**

<table>
<thead>
<tr>
<th>When</th>
<th>Where</th>
<th>Zoom</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monday</td>
<td>11:10am - 12:00am</td>
<td>14-212</td>
</tr>
<tr>
<td>Tuesday</td>
<td>1:10pm - 3:00pm</td>
<td>14-212</td>
</tr>
<tr>
<td>Wednesday</td>
<td>10:10am - 11:00am</td>
<td>14-212</td>
</tr>
</tbody>
</table>

Additional appointments can be scheduled by emailing the instructor at dekhtyar@calpoly.edu.

**Description**

This class is an overview of the field of knowledge discovery from data (KDD, also often referred to as "Data Mining" or "Machine Learning") and related technologies. The course is intended for junior/senior students in Computer Science, Software Engineering and Computer Engineering majors, as well as for students completing the Cross-Disciplinary Studies Minor in Data Science. The course gives a broad overview of data mining (association rules mining, classification, clustering), information filtering and recommender systems, information retrieval and web search, and web mining.

**Learning Objectives**

After taking the course the students are expected to be able to

1. **recognize** different types of KDD procedures and **identify** their uses;
2. **implement** algorithms/methods/techniques for KDD tasks to **solve** KDD problems;
3. **interpret** and **analyze** the **results** of KDD processes;
4. **recognize** and **evaluate societal impact** of KDD technology, **make informed choices** about use of KDD technology.
Texbook


This book is mandatory. It contains almost all material studied in the course (and then some). While we rely on instructor’s lecture notes as much as we rely on the content of the book, the book is extremely useful.

In addition, some other books may be of use. If you want an alternative take on most of the material covered in class, I recommend (as an optional book)


This book is the textbook in Professor Khosmood’s version of CSC 466.

Topics

<table>
<thead>
<tr>
<th>No.</th>
<th>Topic</th>
<th>Duration</th>
<th>Liu</th>
<th>Zaki, Meira</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Association Rules</td>
<td>1</td>
<td>Chapter 2</td>
<td>Chapters 8, 9</td>
</tr>
<tr>
<td>2.</td>
<td>Supervised Learning (Classification)</td>
<td>2</td>
<td>Chapter 3</td>
<td>Chapters 18, 19, 22</td>
</tr>
<tr>
<td>3.</td>
<td>Unsupervised Learning (Clustering)</td>
<td>2</td>
<td>Chapter 4</td>
<td>Chapters 13, 14, 15, 17</td>
</tr>
<tr>
<td>4.</td>
<td>Collaborative Filtering</td>
<td>1</td>
<td>Chapter 11, 12</td>
<td></td>
</tr>
<tr>
<td>5.</td>
<td>Information Retrieval</td>
<td>2</td>
<td>Chapter 6</td>
<td></td>
</tr>
<tr>
<td>6.</td>
<td>Link Analysis</td>
<td>2</td>
<td>Chapter 7</td>
<td></td>
</tr>
</tbody>
</table>

Please note that the order in which these topics are covered may be different than the order in which they are presented above.

Grading

- Labs 55-60%
- Lab Exam 10-15%
- Projects 25-30%

Course Policies

Prerequisites

The official prerequisite for this course is CSC 349 (Algorithms). However, in recent years it is becoming clear that the real prerequisite for the course needs to be DATA 301. However, as DATA 301 is not, in fact, an official prerequisite, we will make sure that any DATA 301 material needed for the course is covered in the class itself. Primarily, this affects what I call data-centric Python programming (i.e., Python programming that uses the numeric and data management Python packages such as NumPy, pandas, SciPy, matplotlib). CSC 466 will use Jupyter notebooks as the main software development platform for the course assignments, and proper use of the abovementioned packages (including proper use of matrix operation broadcasting and similar programming techniques) are important components of the course.

Those of you who took DATA 301 may see some content repetition (e.g., we will cover again some of the algorithms you saw in DATA 301). Please bear with me on this.

Exams

The course will have no written exams. In their stead you will be offered three things:

1. A take-home group project. The project assignment will come out after Week 5-6 of the class, once the finish covering the core topics of the course (classification and clustering). The group assignment will ask you to apply the methods studied in the course to exploratory analysis of some real-life data.

2. A take-home individual assignment. This assignment will come out in the last 2-3 weeks of the class.

---

1 This will be put in place when we switch to semesters; unfortunately, we could make this change to the quarter version of the course.
3. **An in-class lab exam.** The lab exam will most likely take place during the lab periods of the last week of the course. You will be asked to write some code to implement some straightforward KDD tasks. We will be using Jupyter Notebooks for the exam, and you will be given some sample problems at least a week or two ahead of the time.

Most of lab assignments and other coursework in this course is done in either small teams or in pairs. The individual assignment and the lab exam were introduced into the class to allow me to properly assess how individual students in the course attain the learning objectives - something that otherwise may have been obscured due to teamwork.

The reserved final examination time will be used for group project presentations.

## Labs

Hands-on KDD-related activities are the core part of the course. Each lab assignment will involve some data analysis task, that may involve using existing software, software provided by the instructor as well as (and mostly) the software developed by you. The course concentrates on **basic algorithms for performing standard KDD tasks**:

- the labs/assignments give you an opportunity to cement the knowledge of the algorithms covered in class and gain experience running them and investigating their properties.
- Most of the lab assignments are pair programming assignments, although some exceptions may be made either in favor of small teams, or in favor of individual assignments. This will be announced ahead of each lab assignment.

**Note:** Machine learning, data mining and other KDD algorithms that we are going to study in the class are widely available, both as open source code and, in some cases, as methods/functions in popular KDD libraries. **One of the goals of this course is to have you implement these algorithms from scratch!** The assignments will specify when you can, and when you cannot third-party code/libraries to achieve the goals of the assignment.

**Note 2:** At the same time, in a lot of the assignments, the main deliverable will be not the code you write, but rather, the insight you obtain by running your code on the data provided to you. Please be aware of that, as this shift in the main deliverable, is perhaps one of the key unique features of my sections of CSC 466.

## Use of Generative AI in this course.

Use of generative AI for producing code for the assignments in the course contradicts the learning objectives of the course, and thus it is not allowed and will be treated as a violation of University Academic Integrity policies. No code submitted as part of your lab deliverables can be generated by any generative AI service. This is regardless of whether this code is correct or not, and whether you understand what it does.

Use of generative AI for producing any written portions of lab reports, project report, or other written assignments also contradicts the learning objectives of the course, and thus is not allowed. All text submitted by you as part of course assignment deliverables shall contain no portions produced by generative AI services. Failure to comply will be treated as a violation of University Academic Integrity policies.

Use of generative AI as part of your project is permitted in some circumstances. For example, you course project may pursue as a goal the study of generative AI. Your project can use text, or any other outputs of generative AI methods as data for analysis. Your project may involve aspects of prompt engineering as part of your work, if it fits the overall objectives of the project. Your project may compare outputs of generative AI with output of your own code. In general, generative AI as **an object of study** is a valid use in the course, while use of generative AI to create (portions of) assignment deliverables is not. (Please note, the course project assignment involves submitting a project proposal, which needs to be approved by the instructor - any use of generative AI shall be disclosed in the proposal, and will be negotiated and approved at that time).

### Late Submissions

**Lab assignments.** Lab assignment due dates in times will be explicitly stated on each assignment. Most lab assignments (including reports) will be submitted through handin. Usually, there is an about 24 hour grace period for assignment submissions - this extension depends on when the assignments are due and when I am planning to collect them from the handin directory.

Late lab assignment submissions can be turned in before or at the beginning of the next lab period for a 10-30% penalty (the exact amount will depend on the submission time and the specific circumstances. No lab assignments can be submitted for credit after that.

\[\text{Or we will find another way.}\]
Extensions. If you have circumstances that affect your ability to submit on time, first and foremost, please inform me of them. I will work with you on a case-by-case basis to find an accommodation that helps you, while allowing us to grade your work promptly.

Project and end-of-quarter assignment. These will be graded during the finals week, and will have as late deadlines as I can afford and still have a chance to grade them. As a result, these final deliverables have hard deadlines.

Course Calendar
Please note the following changes to our schedule:

- **September 22, Friday** (first day of the class): double lecture. No lab.
- **October 16, Monday**: I am out of town. There will be a replacement lecture.
- **October 18, Wednesday**: I am out of town. Double lab for you (proctored by the course ISA).
- **November 10, Friday**: Veterans Day. No class.

Communication
The class will have an official mailing list. The email address for the mailing list is csc-466-01-2238@calpoly.edu. All students enrolled in the class are automatically subscribed to the mailing list (using the email addresses that the CS department has on file).

I encourage questions during classtime and questions via email. My answers to email questions may be broadcast to the entire class via the mailing list, if the answer may be relevant to everyone (e.g. a correction in a text of a handout, or a clarification of a homework problem), and may also appear on the web page. The questions can also be posted to the mailing list directly. The mailing list will also be used for all announcements related to the course. It is your responsibility to read your class-related email. Failure to read email posted to the mailing list cannot be used as an excuse in the class.

Office Hours
I expect that most office hours will be in person. However, I reserve the option of holding office hours over zoom if circumstances make it so that I am at home during the scheduled office hour time (this pertains mostly to Tuesday office hours). My zoom room can be reached via:

https://calpoly.zoom.us/my/dekhtyar

or

https://calpoly.zoom.us/j/2118158081

Each time office hours will be held over zoom, I'll post the announcement and the zoom link to Slack.

Web Page
Class web page can be found at

http://www.csc.calpoly.edu/~dekhtyar/466-Fall2023

Through this page you will be able to access all class handouts including homeworks, lab assignments, project information, lab/project data and lecture notes.

Links to additional information, and notes and announcements will also be posted.

Please note, this class does not use Canvas.

Slack
In addition to the course mailing list and course web page, we will also have Slack workspace. You should have received invitations sent to your Cal Poly email. I have Slack open on all my computing devices, which means that if you need to discuss something with me, this may be a better medium than email, whether this is a one-on-one conversation, or a conversation that involves and entire section or a group of students.
Students with Disabilities

If you have a disability for which you are or may be requesting an accommodation, you are encouraged to contact both your instructor and the Disability Resource Center, Building 124, Room 119, at (805) 756-1395, as early as possible in the term.

Academic Integrity

University Policies

Cal Poly’s Academic Integrity policies are found at

http://www.academicprograms.calpoly.edu/content/academicpolicies/Cheating

In particular, these policies define cheating as

“...obtaining or attempting to obtain, or aiding another to obtain credit for work, or any improvement in evaluation of performance, by any dishonest or deceptive means. Cheating includes, but is not limited to: lying; copying from another’s test or examination; discussion of answers or questions on an examination or test, unless such discussion is specifically authorized by the instructor; taking or receiving copies of an exam without the permission of the instructor; using or displaying notes, ”cheat sheets,” or other information devices inappropriate to the prescribed test conditions; allowing someone other than the officially enrolled student to represent same.”

Plagiarism, per University policies is defined as

“... the act of using the ideas or work of another person or persons as if they were one’s own without giving proper credit to the source. Such an act is not plagiarism if it is ascertained that the ideas were arrived through independent reasoning or logic or where the thought or idea is common knowledge. Acknowledgement of an original author or source must be made through appropriate references; i.e., quotation marks, footnotes, or commentary.”

University policies state: ”Cheating requires, at a minimum, an ”F” assigned to the assignment, exam, or task, and this ”F” must be reflected in the course grade. The instructor may assign an ”F” course grade for an incidence of cheating.”

Additionally, the procedure for addressing cheating specifies: ”Irrespective of whether an appeal is made, the instructor is obligated to submit to the OSRR director a Confidential Faculty Report of Academic Dishonesty. Physical evidence, circumstantial evidence, and testimony of observation may be attached.”

University policies for addressing plagiarism give instructors more leeway in dealing with it: ”Plagiarism may be considered a form of cheating and therefore subject to the same procedure which requires notification to the OSRR director and, at a minimum, an ”F” assigned to the assignment, exam, or task. However, plagiarism may be the result of poor learning or poor attention to format, and may occur without any intent to deceive; consequently, some instructor discretion is appropriate. Provided that there was no obvious intent to deceive; consequently, some instructor discretion is appropriate.”

Course Policies

Violations of Academic Integrity occur where there is pressure on the student and an opportunity for some form of cheating. CSC 349 assignments are out of necessity such that there is plenty of opportunity. If you feel the pressure (that may be coming in the form of a looming deadline, issues with coursework in other classes, or issues outside the classroom), I would much rather that you contacted me and discussed the assignment and your situation with me before you commit an act of academic cheating.

Having said that, the other side of the coin is that any academic cheating that is revealed as part of the course will be dealt with according to the University policies, which means reporting cases to the Department and Office of Student Rights and Responsibilities (OSSR), with OSSR making the final determination in any cases referred to them.

To that end, here are the specific policies for the course, in addition to the no generative AI policy. Lab assignments are to be completed by the appropriate units (individual, pair, group), and no code/solution-sharing between units is permitted. Students are encouraged to discuss class content among themselves but NOT in a manner that constitutes plagiarism and cheating as defined above. Work on individual assignments needs to be performed by each student individually. Work on project may involve more open consultations, however the final deliverables shall be fully attributable solely to the project team.