Data Mining Capstone

Course Description

The Data Mining Capstone course provides an opportunity for those students who have already taken multiple topic courses in the general area of data mining to further extend their knowledge and skills of data mining through both reading recent research papers and working on an open-ended real-world data mining project. (Note: You should complete Introduction to Data Mining and Text Information Systems before beginning this course.)

The course contains two synergistic components to be offered in parallel, an online seminar and a capstone project. In the online seminar, the students would collaboratively learn about frontier data mining research topics by reading, discussing, and presenting recent research papers in data mining and related areas. In the capstone project, the students would learn how to apply data mining techniques to solve real-world data mining challenges by working on an open-ended exploratory project that involves the use of learned algorithms and techniques in previous topic courses in data mining to analyze and mine real world data sets for discovery of useful knowledge. The two components are synergistic in that the new algorithms learned from the research papers through the seminar component can be potentially applied to the data mining project, which would further help students solidify the understanding of the new algorithms; students are encouraged to pursue such opportunities to integrate the two course components.

The first component, online seminar, will provide the students with an opportunity to collaboratively learn about frontier research topics in data mining and related areas. The format is as follows:

1. At the beginning of the semester, a set of recent research papers broadly relevant to data mining would be selected by the instructor. The number of papers may depend on the number of students enrolled in the course, but it would be typically no more than 15 papers. Each student is expected to 1) read, digest, and present one paper randomly assigned to him/her, and 2) review a minimum three presentations by other students in the class (of three different papers than the student himself/herself is assigned to present). Thus every student is guaranteed to know very well about the paper to present and know reasonably well about three other papers via peer-reviewing the presentations of those papers by others. With this format, students would be able to digest multiple papers efficiently via watching the presentations of those papers by others, while also helping others learn by producing their own presentations. Students are encouraged to watch all presentations including those that may have not been assigned to them for reviewing.

2. The seminar proceeds in three phases: The first phase (about 3 weeks) is for students to read, discuss (on Piazza) and digest the assigned papers. The second phase (about 2 weeks) is for

students to produce their presentations. . Once a student creates a presentation, the student will submit it for peer-review on Coursera. The third phase (about 4 weeks) is for students to review the presentations of their peers and grade them. Peer grading offers the opportunity for a student to understand what others have read, learn from them, and offer his or her feedback, thoughts and suggestions.

3. To facilitate discussion of papers and peer-grading, the students will be randomly divided into three or more small study groups each with a minimum of four students (the actual size would depend on the number of students enrolled in the class). Those study groups would read the same set of papers in each session in parallel. The students in the same study group would each be assigned a different paper to read, discuss, and present, and they would peer-review three presentations by others in the group. For example, suppose a study group has 5 students, and we have three study groups. The three study groups would be reading the same set of 5 papers, which would be assigned randomly to the 5 students in every study group with each student receiving one different paper. The three study groups would then read, discuss, and present the 5 papers in parallel. Every student is responsible for reading the assigned paper in detail and producing a presentation for the paper, which would be reviewed by three peers in the same study group. In the end, every student would have fully understood one paper and reviewed the presentations of three other papers.

4. All the student presentations would be available to all the students so that they can also (optionally) learn more about other papers if they want.

5. The multiple study groups reading the same set of papers would result in multiple students assigned the same paper to present, creating an opportunity for collaborative learning via discussion of the same paper using Piazza. That is students who are assigned the same paper to read and present can help each other digest the content of the paper accurately by using Piazza. Specifically, a separate post will be made for every paper on Piazza so that the students can discuss every paper by posting comments on the Piazza post corresponding to the paper. The discussion of a paper is not limited to only those students who are assigned the paper, so every student is welcome to join the discussion of any paper. Indeed, if you happen to be very familiar with any of the papers not assigned to you, we would appreciate your participation in the discussion of that paper as it would be beneficial to students who are assigned the paper.

6. The grade for the online seminar component will be based on peer grading of the presentation.

The second component of the course, capstone project, will ask the students to solve a real-world challenge. Specifically, students will work on a restaurant review data set from Yelp and use all the knowledge and skills they have learned from the previous courses to mine this data set to discover interesting and useful knowledge. The design of the Project emphasizes:

1. Simulating the workflow of a data miner in a real job setting;
2. Integrating different mining techniques covered in multiple individual courses;
3. Experimenting with different ways to solve a problem to deepen your understanding of techniques;
4. Allowing students to propose and explore their own ideas creatively.

In this semester, the Capstone Project is to analyze and mine a large Yelp review data set with reviews of restaurants to discover useful knowledge to help people make decisions in dining. The project will include the following outputs:

• Opinion visualization: explore and visualize the review content to understand what people have said in those reviews

• Cuisine map construction: mine the data set to understand the landscape of different types of cuisines and their similarities

• Discovery of popular dishes for a cuisine: mine the data set to discover the common/popular dishes of a particular cuisine

• Recommendation of restaurants to help people decide where to dine: mine the data set to rank restaurants for a specific dish and predict the hygiene condition of a restaurant

From the perspective of users, a cuisine map can help them understand what cuisines are there and see the big picture of all kinds of cuisines and their relations. Once they decide what cuisine to try, they would be interested in knowing what the popular dishes of that cuisine are and decide what dishes to have. Finally, they will need to choose a restaurant. Thus, recommending restaurants based on a particular dish would be useful. Moreover, predicting the hygiene condition of a restaurant would also be helpful.

By working on these tasks, students will gain experience with a typical workflow in data mining that includes data preprocessing, data exploration, data analysis, improvement of analysis methods, and presentation of results. The students will have an opportunity to combine multiple algorithms from different courses to complete a relatively complicated mining task and experiment with different ways to solve a problem to understand the best way to solve it. We will suggest specific approaches, but you are highly encouraged to explore your own ideas since open exploration is, by design, a goal of the Project.

Students are required to submit a brief report for each of the tasks for peer grading. A final consolidated report is also required, which will be peer-graded. The grade for the Capstone Project component is based on a weighted combination of the grades of individual tasks and the grade of the final project report.

The final grade for the course is based on a weighted combination of the grades for the two components. Peer grading is emphasized in this capstone course to facilitate collaborative learning. Clear rubrics will be provided to ensure peer grading to be as fair as possible. In the unlikely case when a student feels that his/her work has not been fairly graded in the peer grading process, the student may request for regrading of the work and a TA will regrade it.

Course Prerequisite

- Introduction to Data Mining
- Text Information Systems

Course Expectation

This course is designed to be studied on your own. No further lecture content will be provided. Collaborative learning is strongly encouraged and facilitated via peer grading and discussions on Piazza.

Course Goals

By the end of the course, you will be able to:

- Read, discuss and digest a research paper in data mining or related areas
- Produce a technical presentation of a research paper
- Explain the basic process of generating useful knowledge from a raw data set
- Analyze a large data set with both structured attributes and unstructured text
- Combine different data mining algorithms to accomplish a complex data mining task or improve the effectiveness of data mining
- Evaluate data mining algorithms quantitatively with a test collection
- Choose appropriate techniques to effectively visualize results from data mining algorithms
- Write a report to summarize a data mining task and results, including explanation of the techniques used and discussion of the results

Textbook

Please note: There are no required textbooks or required readings for this course. Relevant readings may be posted on Course Piazza as needed.

Course Schedule

Week	Duration (MM/DD-MM/DD)	Online Seminar	Capstone Project
1	1/15 – 1/21	Paper Reading & Discussion	Task 1: Exploration of Data Set
2	1/22 - 1/28	Paper Reading & Discussion	Task 2: Cuisine Clustering and Map Construction
3	1/29 - 2/4	Paper Reading & Discussion	Task 2: Cuisine Clustering and Map Construction
4	2/5 - 2/11	Paper Presentation Creation	Task 3: Dish Recognition
5	2/12 - 2/18	Paper Presentation Creation	Task 3: Dish Recognition
6	2/19 - 2/25	Presentation Peer Review	Task 4: Popular Dishes
7	2/26 - 3/4	Presentation Peer Review	Task 4: Popular Dishes
8	3/5 - 3/11	Presentation Peer Review	Pattern discovery; Clustering; Text retrieval;

			Text mining and analysis; Visualization
9	3/12 - 3/18	Presentation Peer Review	Task 5: Restaurant Recommendation
10	3/19 - 3/25	Spring Break	Task 5: Restaurant Recommendation
11	3/26 - 4/1		Task 6: Hygiene Prediction
12	4/2 - 4/8		Final Report

Elements of This Course

The course comprises the following elements:

- **Lecture Video.** A short video gives an overview of the Capstone project and tasks. No further lecture videos will be provided in this course.
- **Paper Presentation.** You will read, discuss and present one paper. You will be assigned a paper to read in detail and digest its content accurately. You will create a video presentation to talk about the paper that is assigned to you. In the presentation, you are not only required to demonstrate your presentation skills, but also required to share insight and takeaway from the paper. You will submit your presentation for your peers to review. You are required to peer-review three other presentations to gain in-depth understanding of what others have read and share/offer constructive feedback.
- **Capstone Project.** A comprehensive project with multiple tasks is offered. Finishing these tasks and writing a consolidated project report are your main goals. These tasks are scattered through the whole semester, each with a specific deadline as described above. For each task, you are required to submit a brief report describing your work and also to review the submissions of some of your peers. At the end of the course, you are required to submit a consolidated project report to summarize all your work in the course. All the reports will be peer graded.
- **Requested Regrading.** In the unlikely case when you feel that your work has not been fairly graded in the peer grading process, you may request regrading of any of your work, and a TA will regrade it.

Assignment Deadlines

For all assignment deadlines, please refer to the Course Deadlines, Late Policy, and Academic Calendar page.

Grading Distribution and Scale

Grading Distribution

Your final grade will be calculated based on the activities listed in the table below. Your official final course grade will be listed in Enterprise. The course grade you see displayed in Coursera may not match your official final course grade.

Assignment	Percentage
Paper Presentation	33%
Capstone Project	67%
Total	100%

Student Code and Policies

A student at the University of Illinois at the Urbana-Champaign campus is a member of a University community of which all members have at least the rights and responsibilities common to all citizens, free from institutional censorship; affiliation with the University as a student does not diminish the rights or responsibilities held by a student or any other community member as a citizen of larger communities of the state, the nation, and the world. See the University of Illinois Student Code for more information.

Academic Integrity

All students are expected to abide by the campus regulations on academic integrity found in the Student Code of Conduct. These standards will be enforced and infractions of these rules will not be tolerated in this course. Sharing, copying, or providing any part of a homework solution or code is an infraction of the University's rules on academic integrity. We will be actively looking for violations of this policy in homework and project submissions. Any violation will be punished as severely as

possible with sanctions and penalties typically ranging from a failing grade on this assignment up to a failing grade in the course, including a letter of the offending infraction kept in the student's permanent university record.

Again, a good rule of thumb: *Keep every typed word and piece of code your own*. If you think you are operating in a gray area, you probably are. If you would like clarification on specifics, please contact the course staff.

Disability Accommodations

Students with learning, physical, or other disabilities requiring assistance should contact the instructor as soon as possible. If you're unsure if this applies to you or think it may, please contact the instructor and Disability Resources and Educational Services (DRES) as soon as possible. You can contact DRES at 1207 S. Oak Street, Champaign, via phone at (217) 333-1970, or via email at disability@illinois.edu.