DSC441 – Fundamentals of Data Science, Summer I 2021 (Sections 201, 210)

Instructor Dr. Alvin Chin
Class hours: Tuesday and Thursday 5:45 PM to 9:00 PM on online Zoom (check link in D2L)
Office: Online Zoom (check link in D2L)
Email: alvin.chin@depaul.edu
Will respond within 24 hours or 1 business day, include DSC 441 in the subject!
Office Hours: Wed. 4:30PM-5:30PM or online Zoom by appointment
Course website: https://d2l.depaul.edu/

Summary of Course

Data science is a vast and growing field that focuses on the technology, tools and techniques required to discover patterns and relationships hidden in large databases. With statistics, machine learning and many other computing and applied math disciplines coming together, there is a large toolkit available to practitioners. In this course, we introduce the concepts of data science, and the tools to work with data from the early stages of gaining understanding, through data preparation, applying algorithms, and evaluating and communicating results. Lecture modules cover techniques and algorithms, while tutorials give direct instruction with a powerful toolkit, and homework problems provide the practice needed to hone your skills. Data science takes a lifetime to master, but the core concepts can be put to use in a single quarter.

Topics include data and its storage and exploration, data cleaning and preprocessing, making predictions (SVM, decision trees) and automatically discovering structure (clustering, association rule mining).

Prerequisites: IT 403 (intro statistics) or DSC 423 (regression) or ECO 520 (business analytics) or consent by instructor.

This course assumes that you have had a basic course in statistics along with an introductory programing course (eg. Python).

Learning Goals

Specific learning goals are provided for each module of the course on the D2L website. Overall, by the end of the quarter, students will be able to:

1. Clean, smooth and normalize data, including by accounting for outliers and missing data
2. Choose among clustering algorithms, explain differences between them, and interpret results
3. Choose among classification algorithms, explain differences between them, and evaluate results
4. Identify specific ethical concerns in data mining

Required Textbook and Printed Resources

There is a required textbook for this course. This book is a great resource on data science, with detailed information on the algorithms we are covering and much more, with clear explanations and additional context:
Jiawei Han, Micheline Kamber, and Jian Pei. Data Mining: Concepts and Techniques, 3rd edition. Morgan Kaufmann, 2011.

Textbook webpage: http://www.cs.uiuc.edu/~hanj/bk3/

Digital book copy: http://myweb.sabanciuniv.edu/rdehkharghani/files/2016/02/The-Morgan-Kaufmann-Series-in-Data-Management-Systems-Jiawei-Han-Micheline-Kamber-Jian-Pei-Data-Mining.-Concepts-and-Techniques-3rd-Edition-Morgan-Kaufmann-2011.pdf

The following books are recommended for the course.

- Covers the world of R libraries for data manipulation and more on ggplot:
  R for Data Science: Import, Tidy, Transform, Visualize, and Model Data. Hadley Wickham and Garrett Grolemund. O'Reilly Media, January 2017. https://r4ds.had.co.nz/
- Covers data mining techniques in less detail but includes a business perspective:
  Data Science for Business. Provost and Fawcett. O'Reilly Media, 2013.
- Similar coverage to our book but a bit newer, however, it is connected to Weka, a data mining package we are not using. Still a great resource.
  Data Mining: Practical Machine Learning Tools and Techniques, 4th Edition. Witten, Frank, Hall and Pal. Morgan Kaufmann Publishers, 2017.
- Covers ggplot, the visualization library we are using, with piles of helpful examples.
  R Graphics Cookbook: Practical Recipes for Visualizing Data, 2nd Edition. Winston Chang. O'Reilly Media, November 2018. https://r-graphics.org/

Statistical package

Students will use the R programming language in this course. Two or three lab sessions (recorded for online students) will be scheduled during the quarter. R is available in the CDM labs. If you prefer to install the software on your own machine, go and download it from http://cran.rstudio.com/ . You can then install RStudio (IDE for R) to run R scripts at https://www.rstudio.com/products/rstudio/download/. Choose the Open Source Edition and your platform (MacOS, Windows 10 or Linux). Introductory notes on R will be posted on the course website.

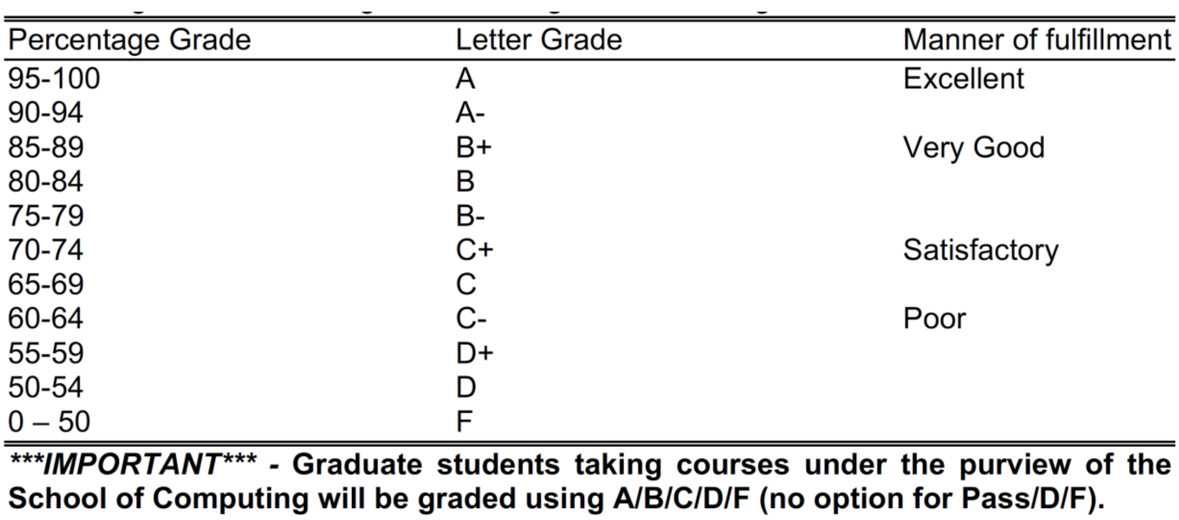
Grading Policy

The final grade has the following components:

- Quizzes (16%) – There will be 4 weekly quizzes online. The quiz will be assigned at the end of Thursday, and you need to complete it by Monday. You have three attempts and your final grade will be the highest of your three scores.
- Final Exam (20%) to be held in Week 5 – Students will use Lockdown Browser to write the exam online and are only allowed a calculator and 1 page (double-sided) cheat sheet (no phones are allowed)
- Homework assignments (60%): There will be 4 weekly assignments to be done individually. Late submissions will be accepted no later than three days after the due date. Late submissions are allowed with a 5%, 10% and 15% penalty for one day, two days and three days, respectively. No submissions will be accepted after 3 days. Extensions may be granted only for exceptional reasons. Requests for extensions must be received BY EMAIL before the due date.

- Participation (4%): Full participation marks will be awarded for asking or answering questions during the online class and/or posting questions and answers in the Discussion section of D2L.

The final grade will be assigned according to the following scale:



Percentage Grade	Letter Grade	Manner of fulfillment
95-100	A	Excellent
90-94	A-	
85-89	B+	Very Good
80-84	B	
75-79	B-	
70-74	C+	Satisfactory
65-69	C	
60-64	C-	Poor
55-59	D+	
50-54	D	
0 – 50	F	

******IMPORTANT****** **- Graduate students taking courses under the purview of the School of Computing will be graded using A/B/C/D/F (no option for Pass/D/F).**

Remarks about homework

Working through homework exercises and applying the statistical concepts to real problems is a critical step for building your understanding of data analysis. Only by trying to apply the statistical techniques you can test if you really understand them. Homework assignments should be regarded as a genuine "learning experience." Study groups are encouraged, but you should, however, be sure that the effort is truly collaborative. The best strategy for completing the assignment is to begin tackling the questions alone, then discussing with others, and finally writing up your answers by yourself. Feel free to consult the instructor if you have any questions.

Students are expected:

- To read this document in full!
- To check email messages regularly and to keep the current email account information on http://campusconnect.depaul.edu.
- To visit the course website and read course announcements on a regular basis.
- To participate actively to class discussions and activities and to work on the practice problems and exercises that are designed to improve students' understanding of the class topics.
- To be familiar with all the course documents and notes posted at the course website.
- To read all the sections in the textbook relevant to the modules. The reading assignments are listed in each module. Notes are meant to complement the course textbook NOT TO REPLACE IT.
- **To contact me regularly and ask me questions related to the course**. You can reach me during my office hours or by appointment at other times. The best way to contact me is through **email at alvin.chin@depaul.edu**. Most emails will be answered within 24 hours.
- To post on the discussion forum messages that are of interest to the entire class.
- To work independently on course assignments and quizzes and not to copy or submit someone else's work as your own. See also University academic integrity policy below.

Homework Assignments and Final Exam Policies

There will be 4 homework assignments during the quarter. Work to be submitted for the course is due one week after it was assigned; late submissions are allowed with a 5%, 10%, and 15% penalty for a one day, two days, and three days, respectively. No late work will be accepted after three days since the assignment was due. The assignments must be submitted online on the D2L site at https://D2L.depaul.edu. Include your name, section number, date, and homework number on the first page of your assignment. It is your responsibility to check that your files are uploaded correctly on D2L; you should always keep a copy of your submission.

There will be a final exam (worth 20%) that will be run using Lockdown Browser on a computer during the last day and last week (week 5) of the course. More information on the content and procedure will be made available in class and on D2L.

Tentative Schedule

The following schedule is tentative. The references are from the course textbook or from other references in D2L.

Lecture (Class date)	Topic	References
Week 1, Lecture 1 (June 15)	Syllabus, Goals & Overview of the course topics. Introduction. Introductory tutorial on R, Data Storage & Variables. Quiz 1 handed out, Assignment 1 handed out.	Chapter 1 (book), Introduction to R tutorial, R references (see above and on D2L)
Week 1, Lecture 2 (June 18)	Data Exploration	Chapter 2 (book)
Week 2, Lecture 3 (June 22)	Data Preprocessing. Quiz 2 handed out, Assignment 2 handed out.	Chapter 3, Chapter 4 (book)
Week 2, Lecture 4 (June 24)	Supervised Learning - Classification Using SVM, Basic Evaluations	Chapter 8, Chapter 9 (book)
Week 3, Lecture 5 (June 29)	Supervised Learning - Classification Using Decision Trees (Part I). Quiz 3 handed out, Assignment 3 handed out.	Chapter 8 (book)
Week 3, Lecture 6 (July 2)	Supervised Learning - Classification Using Decision Trees (Part II).	Chapter 8 (book)
Week 4, Lecture 7 (July 6)	Supervised Learning - Classification Using KNN, Supervised Learning Methods Comparison. Quiz 4 handed out. Assignment 4 handed out.	Chapter 8, Chapter 9 (book)
Week 4, Lecture 8 (July 8)	Unsupervised Learning - Hierarchical Clustering, Unsupervised Learning - K-Means Clustering.	Chapter 10 (book)
Week 5, Lecture 9 (July 13)	Advanced evaluation techniques, association rule mining.	Chapter 11 (book)

Week 5, Lecture 10 (July 15)	Final exam	In class (sync) or on your own (until July 18)

Important Dates

Tuesday, June 15, 2021 – First day of class, 11:59 PM deadline to add classes to SUI 2021 schedule

Friday, June 18, 2021 – Last day to drop SUI 2021 classes with no penalty, Last day to select auditor status for SUI 2021 classes, Last day to select pass/fail option for SUI 2021 classes

Saturday, June 19, 2021 – Grades of "W" assigned for SUI 2021 classes dropped on or after this day

Monday, June 21, 2021 – Deadline to complete Quiz #1 and Assignment #1 by 11:59 PM on D2L

Monday, June 28, 2021 – Deadline to complete Quiz #2 and Assignment #2 by 11:59 PM on D2L

Monday, July 5, 2021 – Independence Day Observed – University officially closed

Tuesday, July 6, 2021 – Deadline to complete Quiz #3 and Assignment #3 by 11:59 PM on D2L, Last day to withdraw from SUI 2021 classes

Monday, July 12, 2021 – Deadline to complete Quiz #4 and Assignment #4 by 11:59 PM on D2L

Thursday, July 15, 2021 – Final exam in class using Lockdown Browser

Sunday, July 18, 2021 – Last day to complete Final exam using Lockdown Browser, End Summer Session I 2021

Tutors

CDM offers free tutoring for many of its courses. The tutors' schedule is at: https://www.cdm.depaul.edu/Student-Resources/Pages/Student-Tutoring.aspx. If you have any difficulty with the course topic, you should contact me or come and talk to me during office hours.

College Policies

Changes to Syllabus

This syllabus is subject to change as necessary during the quarter. If a change occurs, it will be thoroughly addressed during class, posted under Announcements in D2L and sent via email.

Online Course Evaluations

Evaluations are a way for students to provide valuable feedback regarding their instructor and the course. Detailed feedback will enable the instructor to continuously tailor teaching methods and course content to meet the learning goals of the course and the academic needs of the students. They are a requirement of the course and are key to continue to provide you with the highest quality of teaching. The evaluations are anonymous; the instructor and administration do not track who entered what responses. A program is used to check if the student completed the evaluations, but the evaluation is completely separate from the student's identity. Since 100% participation is our goal, students are sent periodic reminders over three weeks. Students do not receive reminders once they complete the evaluation. Please see https://resources.depaul.edu/teaching-commons/teaching/Pages/online-teaching-evaluations.aspx for additional information.

Academic Integrity and Plagiarism

This course will be subject to the university's academic integrity policy. More information can be found at https://offices.depaul.edu/oaa/faculty-resources/teaching/academic-integrity/Pages/default.aspx.

Academic Policies

All students are required to manage their class schedules each term in accordance with the deadlines for enrolling and withdrawing as indicated in the University Academic Calendar. Information on enrollment, withdrawal, grading and incompletes can be found at:
http://www.cdm.depaul.edu/Current%20Students/Pages/PoliciesandProcedures.aspx

Incomplete Grades

An incomplete grade is a special, temporary grade assigned by an instructor when unforeseeable circumstances prevent a student from completing course requirements by the end of term and when otherwise the student had a record of satisfactory progress in the course. All incomplete requests must be approved by the instructor and a CDM Associate Dean. Only exceptions will receive such approval. Information about the Incomplete Grades policy can be found at http://www.cdm.depaul.edu/Current%20Students/Pages/Grading-Policies.aspx.

Students with Disabilities

DePaul University is committed to ensuring equal access to its educational and extracurricular opportunities for students with disabilities. The Center for Students with Disabilities (CSD) offers reasonable academic accommodations and services to support our students. We also serve as a resource to the many university departments that have a responsibility to accommodate students. Please see https://offices.depaul.edu/student-affairs/about/departments/Pages/csd.aspx for Services and Contact Information.

Attendance

It is not mandatory to attend the lecture online, but students are expected to watch the lecture recordings. As the class is online and students can watch the classes on their own time after recording, attendance will be taken based on students having watched the recorded class. This will be done by asking a question from the class and the student having to answer that question correctly. The overall grade for participation drops one-third after any absence. Three absences for any reason, whether excused or not, may constitute failure for the course.

Class Discussion

Even though there are no participation marks, students are highly encouraged to ask questions and offer comments relevant to the class topic. Participation allows the instructor to “hear” the student’s voice when lecturing. Students must keep up with the reading to participate in class discussion.

Attitude

A professional and academic attitude is expected throughout this course. Measurable examples of non-academic or unprofessional attitude include but are not limited to: talking or disrupting class when the instructor is speaking, mocking another’s opinion, and cell phones ringing while watching the class in real-time. If any issues arise a student may be asked to leave the online classroom. The professor will work with the Dean of Students Office to navigate such student issues.

Civil Discourse

DePaul University is a community that thrives on open discourse that challenges students, both intellectually and personally, to be Socially Responsible Leaders. It is the expectation that all dialogue in this course is civil and respectful of the dignity of each student. Any instances of disrespect or hostility can jeopardize a student's ability to be successful in the course. The professor will partner with the Dean of Students Office to assist in managing such issues.

Cell Phones/On Call

If you bring a cell phone to the online class, it must be off or set to a silent mode. Should you need to answer a call during class, students must mute and leave the online classroom in an undisruptive manner.