DSC 424: Advanced Data Analysis
Wednesdays, 6-9pm CST
Lewis 1509

Instructor Information

Instructor:	Ronan Johnson
Office:	CDM, Room 822 (subject to change)
Office Hours:	Tuesday, Wednesday 1:00-3:00pm (in office)
Phone:	(302) 521-5064
Email:	sjohn165@depaul.edu
Course page:	http://d2l.depaul.edu

Course Description

The course will teach advanced statistical techniques to discover information from large sets of data. The course topics include visualization techniques to summarize and display high dimensional data, dimensional reduction techniques such as principal component analysis and factor analysis, clustering techniques for discovering patterns from large datasets, and classification techniques for decision making. The methods will be implemented using standard computer packages.

Course Goals

At the end of this course, the student should be able to conduct and interpret the following analysis techniques and be able to identify which approach is appropriate for a given data set and data analysis task to be performed:

- Multivariate linear regression (least-square estimation & normal equations, model building & variable selection)
- Principal component analysis & Factor analysis (Eigen-values and eigenvectors, scree plots, dimension reduction, factor rotation)
- Discriminant analysis (Fisher's discriminant function)
- Canonical Correlation (to assess the relationship between two sets of variables)
- Cluster analysis (similarity measures, hierarchical clustering, density based and spectral clustering).
- Multidimensional Scaling

Highly Recommended Books

- Brian S. Everitt, Graham Dunn, Applied Multivariate Data Analysis 2nd Edition, Wiley; 2 edition (June 28, 2010), ISBN 978-0470711170. There will be suggested readings from this book which will supplement the course material.

Also Recommended

- Hair, Black, Babin, & Anderson, "Multivariate Data Analysis", Published by Prentice Hall, ISBN-13: 9780138132637, 2010 (8th edition). This is a
- Johnson & Wichern, "Applied Multivariate Statistical Analysis", Published by Prentice Hall, ISBN-13: 9780131877153, 2008 (6^{th} edition). This is an older book that covers the mathematical foundations for many of the topics and is a good reference if you want to dive deeper into the math.

Prerequisites
CSC423: Data Analysis and Regression. We also assume that you have had some experience in linear algebra and in particular, matrix arithmetic.

Grading
Grading in this course will be based on a combination of homework, programming and participation assignments, periodic quizzes, the midterm exam, which will be held during the 6th week of class, and the final project. The final grade will be computed based on the following weights:

- Homework/programming assignments and quizzes: 40%,
- Midterm exam: 30%
- Final project: 30%

The midterm exam is mandatory, and you must take it to pass the course, and you must pass the final project to pass the course. The midterm will be held on **October 13th** during class. Makeup exams will only be given in extreme circumstances which must be documented.

Lectures
Lectures will be traditional in-class presentations, discussions and exercises. They will be held every **Wednesday from 6-9pm in Lewis 1509**. You will be expected to attend and/or watch each lecture in its entirety and to actively participate in each lecture or to go through in-class exercises yourself. In addition to the in-class lectures, there will be some supplemental suggested readings from the recommended text that are not required but which will supplement the lectures.

Examples will be usually posted in a separate activity on D2L and will often be in the form of R code. They are provided for you to study on your own to supplement the specific examples I use in the lectures and tutorials.

Homework/Programming Assignments, Papers' Reviews, and Exam Policies

Homework/quizzes/analysis assignments
There will be homework/analysis assignments and quizzes which are due during the week following the lecture in which they are assigned. Specific due dates will be clearly marked on d2l, but will usually be on **Monday nights at 11:59PM CST.**

Late assignments will be accepted up to one week after the assigned due date with a 25% penalty. No assignments will be accepted beyond a week after the due date. The assignments must be submitted online at **https://d2l.depaul.edu**. No assignments will be accepted via e-mail.

Note that each homework assignment will take a **significant amount of time** and it may not be possible to complete a homework in a single day/weekend, so the best way to work on the homework is to start early and work on the materials over time. Remember, it **may take up to 24 hours** to receive a response from an e-mail or posted question, so do not wait until the day the homework is due to ask questions particularly about the meaning or intent of the homework questions. Homework in this class will take a significant amount of time and you are expected to work on it throughout the time that it is assigned.

Quizzes will accompany each lecture and will count for a portion of the homework grade and will be due the same day as the homework. They will cover the material from each week's lectures.

Midterm:
There will be a midterm exam scheduled **Wednesday, October 13th** on the sixth week of class. The midterm is a closed book exam, the only notes you may bring are a two-sided 3x5 index card. Students are allowed to bring a calculator (no phones or internet connected devices are allowed). You must take the midterm to pass the course and makeups will only be given in extreme circumstances such as illness which must be documented with a doctor's note.

The Final Project:
The final project will be a group project with three to four students to a project. The project will be to thoroughly analyze and apply course techniques to a large dataset. You will be expected to apply a range of techniques from the course and from your own readings to the data, and to draw conclusions from your analysis. Your grade in the project will be based on both individual (50%) and group (50%) performance, including the following components:

- Periodic milestones throughout the quarter which will entail both group and individual work
- Minutes of all team meetings documenting your discussions
- A group final summary report for which all members are expected to contribute
- An individual final report detailing your contributions and individual investigations in the project and reflections on what you learned from the project
- Peer evaluations

The groups will be formed in the first three weeks of the class. Group dynamics play an important role in any project, and you are expected to make every effort to both contribute to the group effort and make the environment safe, comfortable and respectful for your team members.

Final projects will be presented by in person during the 10th class lecture, **November 10th**. Each group must present their projects in person and all students are expected to participate. You will receive feedback from me and also from your peers. I will provide each group with a punchlist of items to address in the final week before the submission of the final project, and addressing this punchlist will form part of the group grade.

Non-performance as a team-member on a project
Usually, the peer evaluation and documentation, including the meeting minutes, in addition to an overall desire for excellence, is sufficient motivation for individuals to contribute a fair share to the team project. However, in extreme cases, individuals have been known to completely cease contributing to a team project. If this is the case, a team has the right to notify the instructor that the individual is no longer contributing and the team no longer wants the individual on the team.

This is not a decision to be made lightly, as expulsion from a team will result in the **loss of 40% of the final project grade**. Because this is such a serious decision, any team that makes this decision will also experience a point deduction for the remaining members. In this situation, **each remaining team member will lose 10% of the final project grade.**

Approximate Weekly Schedule

1. **Review of multiple regression and matrix operations on datasets**
2. **The bias/variance tradeoff and regularized regression. Combatting overfitting**
3. **Multicollinearity, eigenvalues and eigenvector, and Principal Component Analysis.**
4. **Factor analysis and factor rotation**
5. **PCA for ordinal variables and correspondence analysis**
6. **Midterm**
7. **Linear Discriminant Analysis**

8. **Cluster analysis and MDS**
9. **Canonical correlation analysis**
10. **Project presentations**

What to Expect

As with any course in mathematics, data analysis and computer science, you are expected to spend a significant amount of time outside of class reviewing lectures and working on homework/projects. The best way to learn mathematical or statistical techniques is to experiment with them on a variety of problems. You will, of course, have a range of problems posed on the homework, but the more you can experiment with these techniques on both real and synthetic datasets, the better you will learn their nuances, and the better prepared you will be to apply them in novel situations.

The topics in this course build on each other, much in the same way as in any programming or math course. Be sure to monitor your progress carefully in this course and come see me immediately if you miss a class or start to fall behind so that we can discuss getting you caught up. It is important, however, that you do so as soon as possible after getting behind so that we can work on a plan to get you caught up.

This is a graduate level class and you are expected to be an independent learner. This course is not a tutorial in a specific software package, rather it is teaching the fundamentals of data analysis that may be applied to any software or computational platform. Introduction to software techniques will be taught as in-class demonstrations, but techniques required for the homework may build on those that are explicitly taught in class. Further, the class will not be taught in a lab and there may not be time to answer all individual questions or look through individual code during class. You are expected to work through class examples at home and make sure that you can both understand and use the techniques covered. There are a wide range of resources available to you for aid including my office hours, the D2L discussion forums, and a wealth of knowledge in online blogs and examples that can help you work through error messages or issues that you may be having.

Cross-listed DSC 324

This course is cross-listed with the undergraduate section DSC 324. As such, each assignment will include questions/problems that will be required of the graduate students only. These will be clearly marked on the assignments. Undergraduates are encouraged to also complete these parts of the assignments for extra credit.

Software

The use of the RStudio statistical system will be taught in class and will be required for homework problems. **RStudio will be the only officially supported platform for the course.** However, the final project will allow you to use any software you wish to complete the analysis (e.g. SPSS, Python, R, etc.) as long as the software has sufficient features to complete the assignment/project and produces output equivalent to software shown in class. You also may use more than one software package to work on your final project.

Please note that this course is **not** a tutorial in R but rather a course in the foundations of multivariate analysis. You will be provided with some introductory tutorials in R to get you going and provided with many examples that will help you with the homework. You should spend time each week going through the posted examples and discuss them with your classmates.

E-Mail questions

If you have a question for me, you may e-mail me directly at sjohn165@depaul.edu. I will respond to all emails within 24 hours. My phone number is listed at the top of this syllabus. I expect it to be used for extremely urgent matters only. I prefer text messages as I don't typically pick up unknown numbers.

Attendance
It is expected that you will watch the videos, complete the online activities and participate in course discussions weekly. This is the single most important action you can take in mastering the course objectives. You are responsible for all material covered, assignments delivered or received, and announcements made in class sessions that you miss. For distance learning students, this means viewing the classes in a timely manner, participate in the discussion forums, and being sure to email or call in any questions that you have.

Changes to Syllabus
This syllabus is subject to change as necessary to better meet the needs of the students. Significant changes are unlikely but will be thoroughly addressed in class, over email, and through announcements on d2l. Minor changes, especially to the weekly agenda, are possible at any time. You will be informed of all such changes.

School policies:

Attitude and Civil Discourse:
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 to others when the instructor is speaking, mocking another's opinion, cell phones ringing, emailing, texting or using the internet whether on a phone or computer. If any issues arise a student may be asked to leave the classroom. The professor will work with the Dean of Students Office to navigate such student issues.
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 class, it must be off or set to a silent mode. Should you need to answer a call during class, students must leave the room in an undisruptive manner. Out of respect to fellow students and the professor, texting is never allowable in class. If you are required to be on call as part of your job, please advise me at the start of the course.

COVID-19 Health and Safety Precautions
Keeping our DePaul community safe is of utmost importance in the pandemic. Students, faculty and staff are expected to (1) wear a mask as required at all times while indoors on campus; (2) refrain from eating and drinking in classrooms; (3) keep current with their COVID-19 vaccinations or exemptions; (4) stay home if sick; (5) participate in any required COVID-19 testing; (6) complete the online Health and Safety Guidelines for Returning to Campus training; and (7) abide by the City of Chicago Emergency Travel Advisory. By doing these things, we are Taking Care of DePaul, Together. The recommendations may change as local, state, and federal guidelines evolve. Students who do not abide by the mask requirement may be subject to the student conduct process and will be referred to the Dean of Students Office. Students who have a medical reason for not complying with any requirements should register with DePaul's Center for Student with Disabilities (CSD).

Respect for Diversity and Inclusion at DePaul University as aligned with our Vincentian Values
At DePaul, our mission calls us to explore "what must be done" in order to respect the inherent dignity and identity of each human person. We value diversity because it is part of our history, our traditions and our future. We see diversity as an asset and a strength that adds to the richness of classroom learning. In my course, I strive to include diverse authors, perspectives and teaching pedagogies. I also encourage open dialogue and spaces for students to express their unique identities and perspectives. I am open to having difficult conversations and I will strive to create an inclusive classroom that values all perspectives. If at any time, the classroom experience does not live up to this expectation, please feel

free to contact me via email or during office hours.

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/onlineteaching-evaluations.aspx for additional information.

Email
Email is the primary means of communication between faculty and students enrolled in this course outside of class time. Students should be sure their email listed under "demographic information" at http://campusconnect.depaul.edu is correct.

Academic Integrity Policy
I expect that you have read and understood DePaul's policy on Academic Integrity: http://academicintegrity.depaul.edu/ It is part of this syllabus; follow it.

Plagiarism
The university and school policy on plagiarism can be summarized as follows: Students in this course, as well as all other courses in which independent research or writing play a vital part in the course requirements, should be aware of the strong sanctions that can be imposed against someone guilty of plagiarism. If proven, a charge of plagiarism could result in an automatic F in the course and possible expulsion. The strongest of sanctions will be imposed on anyone who submits as his/her own work a report, examination paper, computer file, lab report, or other assignment which has been prepared by someone else. If you have any questions or doubts about what plagiarism entails or how to properly acknowledge source materials be sure to consult the instructor.

Incomplete
An incomplete grade is given only for an exceptional reason such as a death in the family, a serious illness, etc. Any such reason must be documented. Any incomplete request must be made at least two weeks before the final, and approved by the Dean of the School of Computer Science, Telecommunications and Information Systems. Any consequences resulting from a poor grade for the course will not be considered as valid reasons for such a request. Students must formally request an incomplete by filling out a Request for Incomplete Grade form, available at the CDM main office, and submitting it to me.

Preferred Name & Gender Pronouns
Professional courtesy and sensitivity are especially important with respect to individuals and topics dealing with differences of race, culture, religion, politics, sexual orientation, gender, gender variance, and nationalities. I will gladly honor your request to address you by an alternate name or gender pronoun. Please advise me of this preference early in the quarter so that I may make appropriate changes to my records. Please also note that students may choose to identify within the University community with a preferred first name that differs from their legal name and may also update their gender. The preferred first name will appear in University related systems and documents except where the use of the legal name is necessitated or required by University business or legal need. For more information and instructions on how to do so, please see the Student Preferred Name and Gender Policy at http://policies.depaul.edu/policy/policy.aspx?pid=332

Resources for Students with Disabilities

Students who feel they may need an accommodation based on the impact of a disability should contact the instructor privately to discuss their specific needs. All discussions will remain confidential. To ensure that you receive the most appropriate accommodation based on your needs, contact the instructor as early as possible in the quarter (preferably within the first week of class).

Students seeking disability-related accommodations are required to register with DePaul's Center for Students with Disabilities (CSD) enabling them to access accommodations and support services to assist with their success. There are two office locations:

- Loop Campus (312) 362-8002
- Lincoln Park Campus (773) 325-1677
- Email: csd@depaul.edu

Students who register with the Center for Students with Disabilities are also invited to contact Dr. Gregory Moorhead, Director of the Center, privately to discuss how he may assist in facilitating the accommodations to be used in a course. This is best done early in the term. The conversation will remain confidential to the extent possible.

Please see https://offices.depaul.edu/student-affairs/about/departments/Pages/csd.aspx for Services and Contact information.