CSC 334: Advanced Data Analysis
Tuesday 5:45PM-9:00PM

Instructor Information

Instructor:	Dr. John McDonald
Office:	CST, Room 831
Office Hours:	Tuesday, 3:30pm-5:00pm Wednesday, 3:30pm-5:00pm
Phone:	(312) 362-5142
Email:	jmcdonald@cs.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 have a basic understanding of the following topics 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)
- Canonical Correlation (to assess the relationship between two sets of variables)
- Discriminant analysis (Fisher's discriminant function)
- Cluster analysis (similarity measures, hierarchical clustering & non-hierarchical clustering).
- Multidimensional Scaling (if time permits)

Highly Recommended Books

- Johnson & Wichern, "Applied Multivariate Statistical Analysis", Published by Prentice Hall, ISBN-13: 9780131877153, 2008 (6th edition).
- Hair, Black, Babin, & Anderson, "Multivariate Data Analysis", Published by Prentice Hall, ISBN-13: 9780138132637, 2010 (7th edition).

Prerequisites

CSC324: Data Analysis and Regression

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: 30%,
- Midterm exam on February 7th: 35%
- Final project due on March 14th: 35%

The midterm exam is mandatory and you must take it to pass the course. Makeup exams/quizzes are only given in extreme circumstances (severe illness, etc.) which must be documented. Students in the undergraduate section will have a different midterm from the graduate section and the requirements for the homeworks will be different. Some of the graduate level material will be extra credit for the undergraduate section. There will also be different requirements for the final project.

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

Homework/programming assignments

There will be homework/programming assignments, which are due at the beginning of class one or two weeks after they are assigned. Late assignments will be accepted up to one lecture later than 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.

Midterm:

There will be a midterm exam given on Tuesday, February 7th. The midterm is a closed book and notes exam, but students are allowed to bring a calculator (no phones or internet connected devices are allowed).

The Final Project:

The final project will be a group project with two to three 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 (40%) and group (60%) 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
- 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 on the 10th week of class, Tuesday, March 7th. Each group must present their projects. Online students are encouraged to participate either through helping to build the presentation or by recording a part of the presentation to be played during the presentation.

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 30% of the final project grade**. Because this is such a serious decision, any team that makes this decision will also experience a point deduction. In this situation, **each remaining team member will lose 7% of the final project grade.**

What to Expect

As with any course in mathematics and computer science, you are expected to spend a significant amount of time outside of class reviewing lectures and working on homeworks/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 homeworks, 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.

Software
The use of the RStudio statistical system will be taught in class and will be required for some homework problems. It will be the only officially supported platform for the course. However, many of the homework problems and the final project will allow you to use any software you wish to complete the analysis (e.g. SPSS, SAS, R, etc.) as long as the software has sufficient features to complete the assignment/project. You are may use more than one software package to work on your final project. Different packages have different strengths and you should learn to leverage each package for its strengths.

E-Mail questions
I get a lot of e-mail each week and to help insure that you get a prompt response from me, you are required to preface each e-mail's subject line with **CSC334:<subject>** (note: capital letters on CSC and no space). **If you do not do this, you may not get a response to your e-mail.**

Attendance
It is expected that you will attend every class; it 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.

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

School policies:

Online Instructor Evaluation
Course and instructor evaluations are critical for maintaining and improving course quality. To make evaluations as meaningful as possible, we need 100% student participation. Therefore, participation in the School's web-based academic administration initiative during the eighth and ninth week of this course is a requirement of this course. Failure to participate in this process will result in a grade of incomplete for the course. This incomplete will be automatically removed within seven weeks after the end of the course and replaced by the grade you would have received if you had fulfilled this requirement.

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.

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), and make sure that you have contacted either:

PLuS Program (for LD, AD/HD) at 773-325-4239 in SAC 220

The Office for Students with Disabilities (for all other disabilities) at 773-325-7290 Student Center 307