THE UNIVERSITY OF MEMPHIS
COMPUTER SCIENCE
COMP 7/8150: Fundamentals of Data Science

Course Description
The data life cycle; data collection, cleansing, visualization and storage; methods and models for data analysis and management; analysis tools, including statistical packages, experimental design software tools and computer simulations; evaluation and deployment of results; ethical and societal factors. PREREQUISITES: Knowledge of a programming language and descriptive statistics, or equivalent, or permission of instructor.

Why this course
Computer and information technology daily generate enormous amounts of data. The exercise is only useful if it can be turned into information and knowledge in the form of applications and models. This course will provide a broad introduction to key state-of-the-art concepts and methods for testing, evaluation, validation, and/or societal acceptance of data analyses using current state-of-the-art tools.

Syllabus

Date[Week(s)]	Topic
Module 1 [1]	The Big Picture: What is Data Science? The data life cycle: pre-processing, analysis, post-processing
Module 2 [2]	Pre-processing: Data gathering, cleansing, visualization, and understanding (Mean, Variance, Standard Deviation. Percentiles.) Data Storage (Relational databases, e.g. MySQL).
Module 3 [3]	Data Management. A Review of current major Tools for Data Analysis [Case Study: Language R and/or Python]
Module 4 [4]	Probability Models for Statistical Methods: Discrete and continuous probability distributions, density functions. Random variables, expected values, variance, correlation. Data Normalization (z-values, transforms). Central Limit Theorem.
Module 5 [6]	Statistical Inference-Estimation: Sampling: point and interval estimation, confidence levels. Estimation from large and small samples. Normal and student-t distributions.
Module 6 [7/8]	Regression with a single and multiple predictors: Data fitting, residual plots, ANOVA tables, standard error and confidence intervals. Assessment of predictive accuracy.
Module 7 [9]	Statistical Inference-Hypothesis Testing: Decision making: null and alternative hypotheses, critical regions, p-values. Testing from large and small samples. Chi-squared and F-distributions. Contingency tables.
Module 8 [10/11]	An Overview of Other Major Techniques in Data Science: Machine Learning (ML). Supervised Learning.
Module 9 [12]	Machine Learning (ML). Unsupervised Learning.
Module 10 [13]	Closing the data life cycle: Business Intelligence Deployment of Results. Ethical and societal factors.
Week 14	Term Projects/More Case Studies

Textbook
[**MJ**] G.J. Myatt and W.P. Johnson (2014). Making Sense of Data II: A Practical Guide to Data Visualization, Advanced Data Mining Methods, and Applications, 1st ed, John Wiley & Sons.
[**ZM**] Nina Zumel, John Mount (2014). Practical Data Science with R, Manning Publications Co, ISBN-13: 978-1617291562.

Supplementary References

John Maindonald and W. John Braun, Third edition (2010).
(Cambridge Series in Statistical and Probabilistic Mathematics).
ISBN-10:0521762936 ISBN-13: 978-0521762939

In addition, a number of up-to-date readings will be provide in each version of the class with new developments and case studies in data science, such as

[C] L. Cao. Data Science: Challenges and Directions. Comm. of the ACM 60:8 (2017), 59-68. https://cacm.acm.org/magazines/2017/8/219605-data-science/fulltext

[D] D.A. Reed and J. Dongarra. Exascale Computing and Big Data, Comm. of the ACM 58:7 (2015), 56-68.

Other Resources

The use of R or A computing platform (e.g. Python) is required for this course. R is a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques. R can be downloaded from http://CRAN.R-project.org

Evaluation

Final grades (+/- will be used) will be assigned based primarily on a term project that requires application to a well-defined problem of the full data life cycle on a corpus of data procured by the student(s). Homeworks (including programming assignments) and presentation of results will also be required. Plus/minus grading will be used. A typical grading rubric will be

- 10% Class Participation
- 30% Homeworks
- 40% Term Project (Proposal and Final Report)
- 20% Reports (Progress Reports and Final Presentation)

Students taking the 8150 version of the class will be expected to carry out term projects requiring deeper and/or more advanced analyses with higher quality outcomes and/or a research component.

Plagiarism/Cheating Policy

P*lagiarism or cheating* behavior in any form is unethical and detrimental to proper education and ***will not be tolerated*.** All work submitted by a student (projects, programming assignments, lab assignments, quizzes, tests, etc.) is expected to be a student's own work. Plagiarism is incurred when any part of anybody else's work is passed as your own (no proper credit is listed to the sources in your own work) so the reader is led to believe it is therefore your own effort. Students are allowed and encouraged to discuss with each other and look up resources in the literature (including the internet) on their assignments, but ***appropriate references must be included for the materials consulted****,* and appropriate citations made when the material is taken verbatim.

If plagiarism or cheating occurs, the student will receive a failing grade on the assignment and (at the instructor's discretion) a failing grade in the course. The course instructor may also decide to forward the incident to the University Judicial Affairs Office for further disciplinary action. For further information on U of M code of student conduct and academic discipline procedures, please refer to: **http://www.people.memphis.edu/~jaffairs/**