UNIVERSITY OF MEMPHIS

COMPUTER SCIENCE

COMP 7/8150 – Fundamentals of Data Science

Credit Hours: 3

Course Description

The course provides an overview of the **data life cycle**, including data collection, cleansing (outlier detection), visualization, and storage. Students will study methods and models for data analysis and management, use tools such as Python or R, and explore statistical techniques, machine learning, evaluation, and deployment of results. Ethical and societal implications of data science will also be emphasized.

This is class is also taught fully online (M50). M50 classes impose an **additional requirement** for success in it, namely a fair a amount of **self-discipline** and **personal initiative** for **strict scheduling of time** in your busy schedule to study the materials and follow up with assignments in a timely manner. The materials **stack up** very quickly and each unit requires good understanding of the preceding content. That's why the modules are NOT released all at once but in a weekly schedule.

Prerequisites

- Knowledge of a programming language (Python preferred or R)
- Basic understanding of descriptive statistics
- Or permission of the instructor

Learning Outcomes

By the end of this course, students will be able to:

1. Understand and apply the stages of the data life cycle.
2. Perform data collection, cleansing, visualization, and storage.
3. Utilize modern tools (Python/R) for statistical analysis and data management.
4. Apply probability, statistical inference, regression, and hypothesis testing.
5. Develop supervised and unsupervised machine learning models.
6. Evaluate and interpret results with awareness of ethical and societal considerations.
7. Complete a term project involving a real-world dataset covering the entire data life cycle.

Weekly Schedule

Module 1 [Week 1]

- Introduction to Data Science & the Data Life Cycle
- Understanding data science problems

Module 2 [Week 2]

- Mathematical and statistical Preliminaries
- Probability Models: distributions, random variables, expectation, correlation, Central Limit Theorem (CLT)

Module 3 [Week 3]

- Data Preprocessing: Collection, cleansing, visualization, descriptive statistics (mean, variance, standard deviation etc.)
- Data storage (Relational databases, e.g., MySQL)

Module 4 [Week 4]

- Scores and ranking
- Data Management and Tools (Case Study: R and/or Python)

Module 5 [Week 6]

- Statistical Inference – Estimation
- Sampling, confidence intervals, estimation with large/small samples

Module 6 [Weeks 7–8]

- Regression (single and multiple predictors)
- Model fitting, residual analysis, ANOVA, predictive accuracy

Module 7 [Week 9]

- Hypothesis Testing: null/alternative hypotheses, p-values, chi-squared, F-tests

Module 8 [Weeks 10–11]

- Machine Learning (Supervised Learning)

Module 9 [Week 12]

- Machine Learning (Unsupervised Learning)

Module 10 [Week 13]

- Achieving Scale, deployment of results

- Ethical and societal implications

Week 14

- Term Project Presentations & Case Studies

Textbook

- M. Garzon, C.Y. Yang, D. Venugopal, N. Kumar, K. Jana, L.Y. Deng (2022). *Dimensionality Reduction in Data Science*. Springer-Verlag. ISBN 978-3-031-05370-2

Supplementary References:

- Chirag Shah (2020). *A Hands-On Introduction to Data Science*. Cambridge University Press.
- G.J. Myatt and W.P. Johnson (2014). *Making Sense of Data II*. John Wiley & Sons.

Software Requirements

- Python (Anaconda recommended) or R (download: http://CRAN.R-project.org) or

---

Grading

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). Quizzes/Homeworks (including programming assignments) and presentation of results will also be required. Plus/minus grading will be used. A typical grading rubric will be

- **Homeworks:** 50%
- **Final examination:** 15%
- **Project proposal and progress report.**: 5%
- **Final Report & Presentation:** 30%

Graduate students (COMP 8150) will be expected to complete more advanced analyses and include 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/**