COMP 4151: Introduction to Data Science. Fall 2019

Time: TR 1pm – 2:25pm. **Location**: DH 119
Instructor: Vinhthuy Phan (vphan@memphis.edu)
Office: Dunn Hall 309. **Office Hours**: 10AM Monday or by appointment

Course Description
COMP 4151. A hands-on and programming-intensive introduction to data science and applications of data mining and machine learning techniques to analyze real data sets. Specific topics include data collection, cleaning, manipulation, and visualization, clustering and developing models to make predictions, and ethical aspects of data science.

Prerequisites: COMP 2150, MATH 4614 or MATH 4635, or permission of instructor.

In this course, students will learn how to visualize and analyze real-word data. Examples include building models and tools to predict home prices based on available information (location, size, neighborhood, etc.), recommending movies based on a person's preference, predicting admissions into graduate schools, and many more.

Course Objectives:

- Students will understand fundamental data structures such as series and dataframes.
- Students will be able to manipulate dataframes with groupby.
- Students will be able to manipulate dataframes with pivoting.
- Students will be able to select data from dataframes.
- Students will be able to create charts to visualize numerical and categorical data.
- Students will be able to build models to cluster data using existing machine learning libraries.
- Students will be able to build regression models and make predictions using existing machine learning libraries.
- Students will be able to build classification models and make predictions using existing machine learning libraries.

Recommended Textbook:
Python for Data Analysis, 2nd Edition, 2017. by Wes McKinney, O'Reilly.

Grading:

Homework assignments	30%
Exams	40%
Project	30%

Grading scale:
A ≥ 94 A- ≥ 90 B+ ≥ 86 B ≥ 83 B- ≥ 80 C+ ≥ 76 C ≥ 73 C- ≥ 70 D+ ≥ 60 D ≥ 50 F < 50

Tentative agenda

1. Python, iterative patterns
2. Dictionaries, Series and Data frames (Pandas)
3. Groupby: movielens dataset
4. Pivoting: movielens dataset
5. Visualization of numerical data: iris dataset

6. Visualization of categorical data
7. Linear regression: California housing dataset
8. Project
9. K-means clustering, clustering evaluation (Rand, silhouette)
10. Hierarchical clustering, connectivity, DBScan, density
11. Classification: K-nearest neighbor, K-nearest neighbor classifier
12. Decision trees.
13. Cross validation, precision, recall.
14. Support vector machine
15. Ensemble methods: random forest, AdaBoost.
16. Feature selection: decision tree, random forest, chi-squared test
17. Dimensionality reduction
18. Parameter optimizations

Academic misconduct:

- Plagiarism or cheating behavior in any form is unethical and detrimental to proper education and will not be tolerated. If plagiarism is found on a problem of an assignment/exam/project, a grade of 0 is given for that assignment/exam/project. A serious offense might also be reported to the appropriate authority in the department and university.
- Verbatim usage of code or writing is considered plagiarism.
- Superficial changes to a solution, piece of code or writing, is also considered plagiarism.
- Students must carry out the main tasks of an exercise, assignment, and project.
- If students need help on smaller, supportive parts of the main tasks, they must give credit and/or citation to collaborators and/or sources. If a piece of code or writing is used and not properly cited (to a classmate, collaborator or source), especially when it's implied that the entire or even partial work is your own, it is considered plagiarism.

Special accommodation:

If you need special accommodation, please let the instructor know immediately.