Syllabus

B365: Data Analysis and Mining
Spring 2020

Instructor C. Raphael

Section 1 (11324) MW 9:30 – 10:45 GR 103 (Gresham Hall in Foster Quad)

Section 2 (11325) TR 4:00 – 5:15 TE F160 (Teter Quad)

Office 315 Miles Brand Hall

Office Hours MW: 4-5

AI for Section 1 Yucong Jiang (yujiang@indiana.edu) Office Hours: Wednesday 2-4PM Myles Brand 313

AI for Section 1 Vijay Krishna (vgopala@iu.edu) Office Hours: Thursday 11:30-1:30 Luddy 2069

AI for Section 2 Eman Hassan (emhassan@indiana.edu) Office Hours: Monday 6-8PM Robotics House (611 Park 2nd floor left)

AI for Section 2 Changchang Ding (dingchan@indiana.edu) Office Hours: Tuesday 1-3PM Luddy 3051N

Course Overview
This course serves as an introduction to Data Analysis and Data Mining, in which we understand real-world data in algorithmic and visual ways. We will use *probability* as the language that supports and unifies our treatment. This subject will be introduced in class without assuming prior familiarity. We will study probabilistically-formulated data models for classification, regression, and clustering, turning these models into algorithms for data analysis. The aim is for students to master some basic probabilistic grounding, learn several useful algorithms, and develop experience thinking critically about the overall process of understanding and interpreting data.

There will be a significant computing component to this class, allowing us to look at many different real and simulated data sets, as well as implementing algorithms to carry out our ideas. The computing will occur both in class and in the homework. This will be done with short programs written in the R language. No prior exposure to R is assumed. The R language has become very popular over years, making it a valuable language to know.

Prerequisites Students should have basic programming skills such as would be acquired through CSCI-C 200, C-211 or INFO-I 210.

Course Notes
I will be placing typed lecture notes for the class on Canvas as we go along. These will contain the entire content for the class will serve as an essential study guide. These notes will be used in place of a text. You are expected to study these notes carefully. The notes present a view of the class similar to the one I will use in lectures, as well as containing the material the homework will draw from.

Grading

Homework	40%	
Midterm	25%	Section 1: March 11
		Section 2: March 12
Final	35%	Section 1: May 8, 8:00AM
		Section 2: May 5, 5:00PM

Homework

Homework for this class will be a mix of written and computing work, consisting of 8 regularly spaced assignments. For the computing component will use the R programming language for problems and experiments. Homework must be submitted to Canvas by the deadlines given online. Homework must be submitted by the due date to receive full credit. Homework will be accepted for 3 days after the due date, but such late submissions will be marked down. Homework will not be accepted more than 3 days after it is due. Do not email your homework to me or the AI. If you do so it will not be considered as submitted. **All homework must be solely the work of the student submitting the assignment.**

Course Outline

1. Data Representations and Simple Visualization
   (a) Types of Variables (Continuous, Categorical, Ordinal, etc.)
   (b) Data Matrix
   (c) Boxplots
   (d) Scatterplots and Pairs plots
2. Probability
   (a) Basic Discrete Probability
   (b) Probability through Counting
   (c) Probability estimates through Simulation
   (d) Joint Probability, including Marginal and Conditional Probability
   (e) Confidence Intervals for Probability Estimates
   (f) Bayes' Rule
   (g) Independence and Conditional Independence
   (h) Probability Tables and Mosaic Plots,
   (i) Intuition Building: Simpson's Paradox, Law of Rare Disease
3. Classification
   (a) Bayes Classifier and Naive Bayes
   (b) Nearest Neighbor Classifiers
   (c) Decision Tree Classifiers
   (d) Overfitting, Regularization, and Cross Validation
4. Regression
   (a) Linear Algebra and the Data Matrix
   (b) Least Squares and the Normal Equations
   (c) Polynomial Regression, Derived Predictor Variables
   (d) Model Selection
5. Clustering
   (a) K-Means Clustering Algorithm
   (b) Gaussian Mixture Models