BUAN-4310 Data Mining and Big Data Analytics

Syllabus and schedule/Fall 2018

Course Detail

Professor: Diala Ezzeddine**,** PhD
Room: PIGT, 328
Time: TTH 10:15 am -12:20 pm
Email: ezzeddinedia@seattleu.edu
Office Hours: By appointment
Website: All assignments and class materials will be posted on Canvas

Course Description and Objective

This course is an introduction to machine learning with R. It introduces the modern concepts and practices of machine learning and data mining. In this course, we will cover well-known supervised machine learning and unsupervised machine learning methods. The focus of this course is to be familiar with the most known data mining and machine learning methods and demonstrate its application in R codes. In addition, we will introduce the Hadoop framework used to deal with big data.

In this course, we are going to write R codes and present results using the R markdown. Classes will include small lectures based on the textbook (Machine Learning with R) followed by in-class R Labs. Those R labs will be group exercises and practices of applying the machine learning technique of each week. By the end of this course, everyone will have a detailed paper to publish (final project) on analyzing real data using all the machine learning methods explained during this course.

Required readings and materials

-Textbook:

- **Machine Learning with R,** by Brett Lantz, second edition, 2015, PACKT publishing, ISBN-13: 978-1784393908

- Here is a list of recommended references:

- **An introduction to R,** by W.N. Venables, D.M. Smith and the R Core Team, 2018, version 3.5.1, CRAN.R-project.org
- **An Introduction to Statistical Learning,** by Gareth James, Daniela Witten, Trevor

Hastie and Robert Tibshirani, 2013, 1st edition, Springer, ISBN-13: 978-1461471370

- **Hadoop in Actions,** by Chuck Lam, 2011, Manning publications, 1st edition, ISBN-13: 978- 1935182191

- Course materials including assignments and links will be made available at the course web page on Canvas, which can be found at: http://seattleu.instructure.com.

Grading

Points will be assigned to individual assignments, not letter grades. A final grade will be assigned at the end of the term based on the student's total points and relative standing in class. A tentative grading schedule is:

A range: 90-100% of total points
B range: 80-89% of total points
C range: 70-79% of total points
D range: 60-69% of total points
F range: less than 59% of total points

This grading schedule is subject to change during the quarter based on the overall performance of the class, but it will NOT be made more difficult.

Grades will be based on the following assignment points:

Component	Points Each	Total	Percent
Assignments (8)	**20**	**160**	**16%**
Midterm Exam	**250**	**250**	**25%**
Final Exam	**250**	**250**	**25%**
Final project	**100**	**100**	**10%**
In-class labs (8) (Attendance/Participation)	**30**	**240**	**24%**

In-class labs

Almost every lecture is followed by in-class R lab and exercises. Class activities will be posted on Canvas before class. You are expected to do the R labs in class and post the answer on Canvas. You can discuss class exercise problems with your fellow students.

Final project

Through the project, you will be applying machine learning techniques to analyze real data. This project will be a summary of all the methods used in this course. Each one of

your assignments will be a piece of your final project. You will select a real dataset and perform different machine learning method using R and present an analysis report on (1) main goal of your project, (2) description of the features of data, (3) all the data mining techniques used (4) Explain the method you think best fit your goal and why, and (5) final findings or conclusions.

You are expected to submit your project on Canvas on **December 10th** .

More details about the final project will be announced during the first lecture.

Course Policies and Expectations

- No late submissions, no make-ups, and no late work will be accepted.
- All assignments are to be turned in **on time** through Canvas.
- Participation is an indication of your future business success and you are encouraged to participate actively in discussion in class.
- You are expected to attend all the sessions and come to class before it starts. Please do not come late to class.
- The use of computers during the class hours is only limited to course materials. If a student misuses a computer (such as web browsing, chatting, doing homework, etc.), it will lower his/her final grade (10 percent off for each incident from your attendance/participation).
- Cell Phones and PDAs: Cellular phones and PDAs must be turned off before entering the classroom.

If you have, or think you may have, a disability (including an invisible disability such as a learning disability, a chronic health problem, or a mental health condition) that interferes with your performance as a student in this class, you are encouraged to arrange support services and/or accommodations through Disabilities Services staff in the Learning Center, Loyola 100, (206) 296-5740. Disability-based adjustments to course expectations can be arranged only through this process.

Computer

We will use computers during each class meeting, make sure you have programs like Excel, R/RStudio installed, or can connect to the SU visual display infrastructure (VDI).

University Resources and Policies

Academic Resources

- Library and Learning Commons (http://www.seattleu.edu/learningcommons/)
- Academic Integrity Tutorial (https://www.seattleu.edu/academicintegrity/)

Academic Policies on Registrar website

- https://www.seattleu.edu/registrar/academics/performance/

Tentative Course Outline

A detailed schedule will be provided on a separate handout

- Introduction
- Chapter 1 Introducing Machine Learning
- Chapter 2 Managing and understanding Data
- Chapter 3 Lazy learning
- Chapter 4 Probabilistic learning
- Chapter 5 Divide and Conquer
- Chapter 6 Regression Methods
- Chapter 11 Ensemble Methods
- Intro to Hadoop
- Chapter 8 Finding Patterns
- Chapter 9 Finding Groups of Data