Data Sci 8010: Data Analytics from Applied Machine Learning

Mar 14 – May 13, 2022| University of Missouri System | MUIDSI

Instructor Information

- **Instructor:** Tozammel Hossain, PhD
- **Department:** Institute for Data Science and Informatics
- **Email:** hossaink@missouri.edu, khx3p@umsystem.edu;
- **Zoom sessions:** During office hours and by request
    - **Online:** Wed/Thu/Fri, 7:00 PM – 8:00 PM
    - **On-campus:** Wed/Fri, 7:00 PM – 8:00 PM

Course Information

- **Duration:** 8 weeks
- **Modality:** Fully online (asynchronous)
- **Credit:** 3 credit hours
- **Location and Schedule:** All coursework in Jupyterhub, presented in 8 modules that cover 1 week each
- **Prerequisites/Co-Requisites:** DATA_SCI 7020
- **Restrictions/Exclusions:** None

Course Overview

Synopsis

This course leverages the foundations in statistics and modeling to teach applied concepts in machine learning (ML). Participants will learn various classes of machine learning and modeling techniques, and gain an in-depth understanding how to select appropriate techniques for various data science tasks. Topics cover a spectrum from simple Bayesian modeling to more advanced algorithms such as support vector machines, decision trees/ forests, and neural networks. Students learn to incorporate machine learning workflows into data-intensive analytical processes.

Course Objective

1. To illustrate the basic methodology of machine learning, beginning from a raw dataset to prediction or latent structure learning
2. Understand the basic concepts and methods of ML
3. Evaluate and interpret various ML models
4. Hands-on practices to help you identify recurring themes of data analytics/machine learning in different problem domains
5. Operationalize ML Models

Course Structure

The course will be conducted in the Canvas Learning Management System (LMS) and Jupyterhub. First, get access to your course repo on Jupyterhub system. Then, review this course syllabus. The information provided in the Syllabus and Start_Here in the course repo should give you all the information you need to get started in the course and know what is expected. If you still have questions after reviewing these materials, however, please contact your instructor by email or post questions in the discussion forum (Slack).

Time Commitment

Over the course of each one-week module, you should anticipate spending an average of 6 to 10 hours on readings, coding, discussions, and activities.

Course Requirements

Required Texts

- No textbook is required. All the required materials are freely available online and introduced in the relevant module.

Required Technology

You are expected to have access to the Data Science & Analytics Computing Environment (DSA Environment). If you do not have access to this system, please reach out to me. In addition, you will need a computer browser to access the course materials on the DSA environment. You spend most of your time with this system for your learning activities.

Netiquette

- Respect the privacy of the participants in the course. Use FERPA guidelines as a benchmark for what you are (and are not) able to share with others inside and outside of this seminar.
- Don’t be afraid to ask questions. If you have a question, chances are one of your course mates has the same, or similar, question.
- Respect diversity and opinions that differ from your own. Communicate tactfully, and base disagreements on scholarly ideas or researched evidence.
- Be polite and professional in *all* communications, written and/or verbal.

Module delivery, Email, and Discussion Forum Expectations

Modules will be released incrementally. Each module will be deployed at the Gitlab system on Sat at 6:00 am. Using Git, you are required to fetch each module on the DSA environment. The assignments are due by the following Sat at 11:59 PM.

If you have technical questions or logistics questions for this course head over to **#8010-aml-oncampus-sp22** or **#8010-aml-online-sp22** channel on Slack (Preferred Method). This is fastest way to resolve your issues as TAs and the instructor check this channel frequently. Also, your fellow classmates could answer questions.

If required, you can email the TAs and the instructor (Please check course homepage for emails). While emailing us please write a subject line beginning with "AML-8410".

Course Modules

Module 1: Introduction to Machine Learning

- Machine Learning Overview
    - ML versus PR, DM, and Stat
    - Supvervised vs Unsupervised vs Reinforcement
    - Machine Learning Workflows
        - Training
        - Testing
    - Supervised ML
        - Baseline classifier
        - Classification

Module 2: Supervised Learning: Classification & Regression

- Cross-Validation
- Supervised Learning: Classification
    - Naive Bayes
    - Logistic Regression
- Supervised Learning: Regression
    - Linear Regression

Module 3: Feature Selection and Dimensionality Reduction

- Feature Selection
    - Forward Selection
    - Backward Elimination
    - Mutual Information
- Feature Extraction/Dimensionality Reduction
    - PCA
    - Kernel Methods
    - Factor Analysis

Module 4: Unsupervised Learning: Clustering & Anomaly Detection

- Clustering
    - K-Means
    - Hierarchical
    - DB-Scan
- Anomaly Detection
    - Isolation Forest
    - Local Outlier Factor

    - Elliptic Envelope

Module 5: Grid Search and Pipeline

- Class-Imbalance Problem
- Grid Search
- Random Search
- Pipeline

Module 6: Final Project - Part I

- Operationalizing an ML pipeline

Module 7: Final Project - Part 2

- Operationalizing an ML pipeline

Module 8: A brief Intro to Neural Networks

- Perceptron / Neuron
- Neural Networks
- Feed Forward
- Back Propagation
    - Gradient descent

Grading & Feedback

Turnaround Expectations

Feedback will be provided via Canvas system. We aim to have feedback on Canvas by the end of the following module. For instance, feedback on the Module 1 activities will be provided by the end of Module 2. If I am delayed on providing feedback, I will post an announcement to let you know.

Grading Policy

Each module has readings, labs, practices, and exercises. The tasks in labs and practices are solved when a module is releases. For some modules, students are required to participate in a Canvas Discussion on a topic from the readings. Students are expected to go through labs and practices before solving the exercises. Here is a tentative score distribution.

- Practices (10%)
- Exercises (60%)
- Final Project (30%)