IS 733 Project Requirements

The aim of the IS 733 project is to give you hands-on experience in data mining, and to help synthesize and apply what you are learning in the course. This course is about data mining techniques for finding hidden patterns in complicated data sets in order to solve real-world data analytic tasks, and your projects should reflect this. Projects will address either the application of data mining techniques to solve a particular task, or the development of new techniques, validated by experimental results. A typical project might involve:

- Selecting an interesting data mining task to work on, e.g. with industrial significance, or motivated from computational biology, sociology, etc. Acquiring a dataset (often by downloading an existing dataset, but could instead involve web scraping, etc).
- Exploratory analysis and/or visualization to get to know the data (see Lesson 3).
- Data preprocessing (see Lesson 3).
- Implementing and applying one or more data mining techniques to solve the task.
- Rigorous experimental evaluation. For most projects, this will involve cross-validation or train-tests splits, validation sets, and comparisons to baseline methods (see Lesson 8).

Note that you may have to read ahead to start your project. All readings are listed on the syllabus at the course webpage. Some important information:

- The project will be done in groups of 2.
- There are four deliverables for the project, worth 35% of the total overall course grade:

• Project proposal	Due 2/23/2021	(5%)
• Midterm progress report	Due 4/6/2021	(5%)
• Poster presentation	Due 5/11/2021	(10%)
• Final report	Due 5/14/2021 (Friday!)	(15%)

- The project may overlap with other research you are doing, but not any other class project. Projects that are related to any research you or your teammates may be working on, or your job, are, in fact, encouraged – I hope this course helps you solve the data science tasks you are already interested in. However, your project work needs to be your own, and not that of your labmates or collaborators.

Project proposal

Project proposals are to be sent to me by email (nroy@umbc.edu), and approved by the deadline. Please send me a short summary (< 1 page) of your proposed study, including the goals of the project, why this is important to tackle, the data you plan to use, the type of models/algorithms you plan to use, and the names of the group members. *If you need help selecting a project, please arrange to discuss it with me before the deadline, or come to office hours.* We'll share our group projects together in class on 2/23/2021.

Midterm progress report

This a 1-2 page report. Please summarize the progress made, results obtained, and methods tried. Include a discussion of any challenges faced, and plans to resolve them.

Poster

The poster will be presented in class, with a digital copy due at the same time. Posters will be evaluated on informativeness, attractiveness, understandability, delivery of presentation, and technical merit.

Final report

Think of the final report as a near-final draft of an academic research paper. The report must be no more than 15 pages, including references. If you plan to publish your work, you should follow the formatting guidelines of the publication venue. It must include a discussion of the relevant work in the literature, and how the work goes beyond this. The report will be evaluated on clarity in addition to technical merit. Late final reports will not be accepted. The report should begin with a title and author names, and include at least the following sections: abstract (500 words maximum), introduction, background and related work, methodology, experimental results, references.

Project ideas / guidelines

Many types of projects are possible, subject to my approval. It is important that your methods are thoroughly evaluated, including comparisons with simpler approaches, and with competing methods from the literature (if possible). There should generally be an implementation component to the project. Note that research is challenging and uncertain, and it's ok if the project does not yield positive results.

Example topic areas include outlier detection, privacy preserving data mining, time series/sequence analysis, mining biological data, mining financial data, healthcare/clinical applications, business applications, social network analysis, recommender systems, etc. In previous versions of this course, the projects have been a highlight and in some cases have lead to publications in prestigious data mining conferences and journals. Some students have used real-life examples from their jobs as project ideas. Here are a few possible sources of data sets which might help get you started:

- UCI Machine Learning Repository: `http://archive.ics.uci.edu/ml/`
- IEEE Dataport: `https://ieee-dataport.org/`
- PhysioNet: `https://physionet.org/`
- Awesome Public Datasets, a collection of data from a variety of domains with over 100 contributors: `https://github.com/caesar0301/awesome-public-datasets`
- Johns Hopkins University & Medicine COVID-19 Datasets: `https://browse.welch.jhmi.edu/datasets/Covid19`
- Yahoo Webscope datasets: `http://webscope.sandbox.yahoo.com/`
- Yelp Dataset Challenge: `https://www.yelp.com/dataset_challenge`
- Stanford Network Analysis Project (SNAP): `http://snap.stanford.edu/index.html`
- GroupLens recommendation system data sets: `http://grouplens.org/datasets/movielens/`
- ImageNet: `http://www.image-net.org/`
- The Stanford MOOCPosts Data Set: `http://datastage.stanford.edu/StanfordMoocPosts/`
- Microsoft Learning to Rank: `http://research.microsoft.com/en-us/um/beijing/projects/letor/`
- Julian McAuley's webpage: `http://cseweb.ucsd.edu/~jmcauley/`
- Tagged.com social spam data set: `https://obj.umiacs.umd.edu/tagged_social_spam/index.html`
- NLP datasets: `https://github.com/niderhoff/nlp-datasets` and `https://nlp.stanford.edu/links/statnlp.html`
- Mulan's multi-label learning datasets: `http://mulan.sourceforge.net/datasets-mlc.html`
- Kaggle Datasets: `https://www.kaggle.com/datasets`

Titles of previous projects done in this course include:
Predicting Relations Between the Factors Involved In Gerrymandering, *[A Data Mining Study of] Crime in Los Angeles*, *Housing Prices and Crime*, *Voting Predictions Implemented in a Recommender System*, *Food Access and Crime in Baltimore MD*, *Fatality Prediction on Car Crash Data*, *A Descriptive and Predictive Analysis of Chicago Crime Occurrences*, *Diabetic Retinopathy Detection Using Convolutional Neural Networks*, *Assessing the Impact on House Valuation based on Multi Model Data Fusion*, *H1B Visa Petition Prediction*, *Storm Data Analysis for United States.*

Titles of previous projects done in IS 698: Smart Home Health Analytics course include (published in IEEE/ACM):
Smartphone-based Mobile Gunshot Detection, *SoccerMate: A Personal Soccer Attribute Profiler using Wearable*, *HappyFeet: Recognizing and Assessing Dance on the Floor*, *Developing Machine Learning based Predictive Models for Smart Policing.*