CSE514 Fall 2022 Data Mining
Introduction to Data Mining

In a nutshell, Data Mining (DM) aims to find patterns, or a model, embedded in a set of given data, so as to explain previous events and/or predict future events.

- From a statistician's point of view, we are looking for a model *M* so that *prob*(*M*|*D*) is maximized for a given dataset *D*
- From an engineer's point of view, we want to reverse engineer the underlying process that generated the data

Why do we Data Mine?

- Targeted acquisition of knowledge is often difficult, expensive, and slow
- Processes of interest can be too complex for human understanding
- Domain experts take time and money to train
- Advancements in technology means it's increasingly faster and cheaper to generate and store data
- Advancements in Machine Learning (ML) means it's increasingly faster and cheaper to analyze data
- Successful data mining results in new insights, new stories, and profit

What differentiates Data Mining from other fields?

- Probability and statistics form the math foundation of most Data Mining models, but it's generally assumed that:
    1. A data miner works with much larger datasets than a statistician
    2. A data miner is searching for more complex models and patterns than a statistician
    3. The larger datasets targeted by data miners are more likely to have practical issues like dirty data, aka data that has missing values, duplicate values, incorrectly parsed values, or just typos.
    4. Data mining is done with more diverse data, such as text, images, and audio
- Machine Learning approaches overlap with Data Mining approaches, but it's generally assumed that:
    1. Data Mining is focused on applying Machine Learning to a practical application, rather than improving the Machine Learning algorithms
    2. A data miner does not have control over data acquisition and instead must choose the best model to extract knowledge from a given dataset. In machine learning, the model is usually chosen first, and data is compiled for the purpose of training/testing that model
    3. A data miner specializes in data analysis. A Machine Learning researcher specializes in algorithm optimization

Classes of most Data Mining tasks:

- Outlier detection
  Identification of unusual data that may indicate something is worth further investigation
  Ex. An email that is unusual compared to all other emails from the same contact could be an indication of a hacked email account
- Association rules
  Searching for relationships between variables without assuming causality
  Ex. Time phrases ("act now" and "limited time") in emails are associated with retail phrases ("buy one get one free" and "20% off")
- Clustering
  Discovering groups of "similar" objects within data
  Ex. Emails received can be clustered into groups used features like length, time received, and whether the sender is in the receiver's list of contacts
- Classification
  Predicting a qualitative label for an unlabeled sample
  Ex. New emails can be classified as personal or work, based on how many features it shares with known personal/work emails
- Regression
  Predicting the quantitative value of dependent variables from independent variables
  Ex. Predict the number of emails that will be received on a particular day
- Summarization
  Producing a more compact description of data
  Ex. The percentage of emails that are spam for the average user, for comparison with a specific user's experience