CP400R: Data Mining and Enterprise Computing

by Scott King

Winter Term 2017

- Data mining is important, **data** is a buzz word

1 Introduction to Data Mining

1.1 Why mine data? Commercial viewpoint

- It is mined to be able to better tailor experiences to the specific user (generally); based on what you Tweet and who you follow, suggest users; what you buy online, give suggestions that you have a chance of buying, etc
- It can also be useful for developing trends, which can be helpful for determining target markets, customer retention, etc
- From a scientific viewpoint, we're looking more at qualitative information (think LHC data, or measurements, etc)

1.2 What is data mining?

- A non-trivial extraction of implicit, previously unknown data and *potentially* useful information
- With mining large data sets, we can gain some insights that were not obvious from face-value
- Leverages a lot of statistical methods, machine learning and organizing data via database technologies
- General challenges of data mining include: scalability, dimensionability, complex data, data purity, privacy preservation, streaming or distributed data

1.2.1 General flow of data mining applications

1. Selection →
2. Processing →
3. Transformation →
4. Data mining →
5. Evaluation

1.2.2 Data mining tasks

- Predictive tasks:
    - Classification

    - Regression
    - Deviation detection
- Descriptive tasks:
    - Clustering
    - Association rule discovery
    - Sequential pattern discovery

1.3 What is classification?

- Given a training set of data, find the **class** (essentially defining attribute)
- Find a *model* for the class attribute as a function of values of other attributes
- The end goal is to have previously unseen records have an associated class as accurately as possible
    - A test set of data is used to determine the accuracy of the model
- Classification techniques:
    - Decision tree methods
    - Neural networks
    - Naive Bayesian algorithms
    - Support vector machines
- Example of when classification is used: direct marketing/customer churn

1.4 What is regression?

- Predict the *next* value given previous [continuous variables] values resembling a linear or non-linear model of dependency
- Example of when regression is used: time series prediction of stock market indices, predict sales of item based on advertising efforts

1.5 What is clustering?

- Given a set of data (all of which have attributes), find similar ones, *clusters*, such that intracluster distsances are minimized and intercluster distances are maximized
- You can sometimes measure the similarity of points using the *Euclidean Distance* between two points
- Some clustering algorithms include:
    - K-means
    - Hierarchical clustering

    - Spectral-clustering

- Examples of when clustering is used: market segmentation, stock data (cluster based on whether the price has increased or decreased and use the similarity measure to correlate behaviour)

1.6 What is association rule discovery?

- Given a set of records, each of which contain some set of items of a given collection; produce dependency rules which will predict the occurance of an item based on occurences of other items

- A common algorithm for this is the *Apriori* algorithm

- Examples of when association rule discovery is used: supermarket shelf management

1.7 What is sequential pattern discovery?

- Given a set of objects, where each object has a timeline of events, find rules that predict strong *sequential dependencies* among different events

    - These rules are formed by discovering patterns in which the event occurences in the patterns are governed by timing constraints

- Examples of when sequential pattern discovery is used: point-of-sale transaction sequences (if Guy purchases A and B, he's likely to purchase C)