Overview of Data Mining

Charles Sutton
School of Informatics

Semester 2

Definition of data mining

[Hand, Manilla, and Smyth]: the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner.

Things to notice about this definition:

- Data usually from observations rather than experiments
- Mining not always primary reason data was collected
- Human interpretability is central
- Need to find nontrivial patterns

Examples of data mining

- Fraud detection [e.g., Fawcett and Provost, 1997]. A rule learning method to identify cell phone spoofing
- Cortes and Pregibon, 1998. Classifying residence versus business of phone company customers based on who they call and when
- Scientific data mining, e.g., astronomy, biology, drug discovery
- Recommender systems. e.g., Amazon, Netflix

Data Mining Tasks

- Prediction, e.g., classification, regression
- Description, e.g., clustering
- Pattern mining, e.g., rule mining, market basket analysis

Process of Data Mining

- Data collection
- Data cleaning (e.g., Looking for outliers, missing data)
- Data management (e.g., Details of data store / database, algorithms for access)
- Mining
- Evaluation

Relationship to other fields

- Statistics
- Machine learning
- Databases

The boundaries between these fields are fluid, but in general, data mining tends to be a little bit less theoretical, a bit more focused on large scale data, and have a bit more prominence in industry

Reading:

- Chapters 1-2, HMS