Data Mining
(Knowledge Discovery)

CISC437/637, Lecture #22

Ben Carterette

Introduction to Data Mining

- **Data mining** is the *automatic* exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and understandable patterns in data
  - **Valid**: the patterns hold in general
  - **Novel**: something we didn't know before
  - **Useful**: we can use the patterns to accomplish something
  - **Understandable**: we can comprehend why the patterns hold
- Example: if degree=MS and income>75000, then credit=Excellent (with 82% probability)

Two Classes of Patterns

- **Predictions** about new data items based on knowledge about existing data items
  - **Classification** algorithms predict categorical variables
  - **Regression** models predict numeric variables
- **Associations** between two or more field values across records
  - **Clustering** discovers groupings of records according to field values

Variable Types

- Variable type depends on domain:
  - **Numerical**: domain is numeric (integer, real, etc)
  - **Nominal** or **categorical**: domain is a finite set with no natural ordering (e.g. occupation, gender, etc)
  - **Ordinal**: domain is ordered (e.g. preference scales, severity of injury, etc)
- The types of variables determine what kinds of data mining techniques might be used

Classification

- The goal of classification is to predict the value of a *categorical* variable given one or more numeric, ordinal, or categorical variables
- Definitions:
  - C = categorical variable/class labels
  - X_1, …, X_k = variables/features/attributes
  - F: $X_1 \times X_2 \times \ldots \times X_k \rightarrow$ C is the classification function
- Given **training data** with features $X_1 \ldots X_k$ and known class C, find the best F

Classification

- We do not expect that the classification will be exact
  - Errors will occur; our goal is to minimize them
- The function F should:
  - Be able to classify items with high accuracy
  - Produce information (the class) that we don't have and that is useful for something
  - Be interpretable by those using it (i.e. it should make sense)

