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 & Enterprise Computing

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 *consequent* is an item that can be used to determine what can be done; an *antecedent* is an item that can be used to see what would be effected if something about the item changes

- 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)

2 *Data*

2.1 What is data?

- A collection of data objects, each with attributes

2.2 Attribute values

- *Attribute values* are numbers or symbols assigned to an attribute

- The same attribute can be mapped to different attribute values (ex. height can be measure multiple ways)

- Different attributes can be mapped to the same value (ex. attribute values for an ID and age are integers, but properties of attribute values can be different)

- There are a few types of attribute values:

    - **Nominal**: IDs, eye colours, zip codes (distinctness)

    - **Ordinal**: rankings, grades, height (distinctness and order)

    - **Interval**: calendar dates, temperature in C or F (distinctness, order and addition)

- ◦ **Ratio**: length, times (distinctness, order, addition and multiplication)

- A *discrete* attribute is one where there are a finite set of values
- A *continuous* attribute has real numbers as attribute values

2.3 Characteristics of structured data

- Dimensionality: *Curse of Dimensionality*
  - When dimensionality increases, data becomes increasingly sparse
  - Because of this, important characteristics (ex. density, distance between points) starts to become less meaningful (pertains less to sample of data)
- Sparsity: only presence counts
- Resolution: patterns depend on scale

2.3.1 Dimensionality reduction

- Reduce the time and effort spent by mining algorithms
- Methods for this include: principle component analysis and singular value decomposition

2.3.2 Types of structured data

- Data matrix:
  - When data objects all have the same attribute values, the data can be represented by an $m \times n$ matrix
- Document data:
  - Each term in the document becomes a vector where the value of each vector is the number of times it appears in the document
- Recommendations data:
  - Sparse matrix where each column is trivial (book, movies, etc) and the value is the rating

2.4 Data quality

2.4.1 Dirty data

- Incomplete data: hardware/software problems, N/A when gathering data
- Noisy data: incorrect values that may appear from faulty collection tools, human/computer error
- Inconsistent data may come from different data sources or a functional dependency violation

2.4.2 Noise

- Refers to the modification of original values

2.4.3 Outliers

- Outliers are data objects with characteristics that are considerably different from other objects

2.4.4 Data preprocessing is important!

- *No quality data, not quality mining results!*
- Often times, the quality of data correlates to the quality of results gained from mining it
- Data extraction and transformation comprises most of the work

2.4.5 Handling redundancy in data integration (combining data)

- This can often occur with merge of multiple databases
- *Correction analysis* may be able to filter out most of the duplicate data

2.5 Aggregation

- Basically a functional reduce on multiple objects (combining)
- A few purposes we'd use this for are: data reduction, change of scale, increased reliability of data

2.6 Sampling

- Sampling is the main technique to obtain data; most of the time it comes down to the expense of processing the entire set of data
- The key principle for effective sampling is:
  - Make sure the sample is more or less representative of the entire set
  - It it consider *representative*, it roughly describes the entire set
- Types of sampling:
  - Simple random: equal probability of selecting an item
  - Sampling without replacement: as an item is selected, it is removed from population
  - Sampling with replacement: same as above, but objects can be selected multiple times
  - Stratified sampling: partition data and sample from each partition

2.7 Dimensionality reduction

2.7.1 PCA

- The goal is find the projection that captures the largest amount of variation among the data
- Find eigenvectors of the covariance matrix where these eigenvectors will define a new space

2.7.2 ISOMAP

- Construct a neighbourhood graph
- Find each pair of points in the graph and compute the shortest path distances (geodisic distances)

2.7.3 Feature subset selection

- Redundant features: duplicate most, if not all, information contained within the attributes
- Irrelevant features: remove features that do not contain any information pertaining to the task at hand
- Some techniques for doing this include:
  - **Brute force**: try all feature subsets as input to data mining algorithm
  - **Embedded approaches**: feature selection is a complementary feature of algorithm
  - **Filtering**: features are selected before the algorithm is run
  - **Wrapper approaches**: use algo as a "black box" to figure out the best attributes

2.8 Attribute transformation

- A function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with new values
- Standardization and normalization with some simple functions: x^k, $\log(x)$, e^x, $|x|$

2.9 Data similarity and dissimilarity

- *Similarity* is to measure how alike two objects are
  - Can be measured numerical where, the higher the number, the more alike two objects are, often the range $[0, 1]$ is used
- *Dissimilarity* is to measure how different two objects are
  - Similar to *similarity*, but the number is lower when objects are more like, lower bound is 0, upper limit can vary
- Proximity refers to similarity/dissimilarity

2.9.1 Measurements of similarity

- *Manhattan distance* is defined as:

$$\text{dist} = \sum_{k=1}^{n} |p_k - q_k|$$

  where n is the number of attributes (dimensions) and k reprensents the k^{th} objects of sets p and q

- *Euclidean distance* is defined as:

$$\text{dist} = \sqrt{\sum_{k=1}^{n}(p_k - q_k)^2}$$

where the variables are similar to the above. Although, standardization may be required.

- There are some common properties of Euclidean distances:
  1. $d(p,q) \geq 0 \ \forall\ p$ and q and $d(p,q) = 0$ only if $p = q$ (positive definiteness)
  2. $d(p,q) = d(q,p) \ \forall\ p$ and q (symmetry)
  3. $d(p,r) \leq d(p,q) + d(q,r) \ \forall$ points p, q and r (triangle inequality)

  where $d(p,q)$ is the dissimilarity between two objects p and q. A distance that satisfies hese properties is called a **metric**.

- *Minkowski distance* is defined as:

$$\text{dist} = \left(\sum_{k=1}^{n}|p_k - q_k|^r\right)^{\frac{1}{r}}$$

where r is a parameter, n is the number of attributes (dimensions)

- *Mahalanobis distance* is defined as:

$$\text{mahalanbois}(p,q) = (p-q)\Sigma^{-1}(p-q)^T$$

where Σ is the covariance matrix of the input data, X. It can be defined as such:

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{k=1}^{n}\left(X_{ij} - \bar{X}_j\right)\left(X_{ik} - \bar{X}_k\right)$$

2.9.2 Correlation

- *Correlation* measures the linear relationship between two objects
- To compute correlation, we standardize data objects, p and q, and then take their dot product

$$\begin{aligned} p'_k &= \frac{(p_k - \text{mean}(p))}{\text{std}(p)} \\ q'_k &= \frac{(q_k - \text{mean}(q))}{\text{std}(q)} \\ \text{correlation}(p,q) &= p' \bullet q' \end{aligned}$$

- When two data sets of data are strongly linked together, we they *high correlation* or:
  - Values in $[-1, 1]$
  - Correlation is *positive* when the values increase together, and *negative* when one values decreases as the other increases

2.9.3 Density

- Density-based clustering often uses one of: euclidean density, probability density, graph-based density

3 Exploring Data