Data Mining

INFS602 Physical Database Design

Learning Outcomes

- Distinguish between OLAP and Data Mining
- Discuss some typical Data Mining Applications
- Discuss some commonly used techniques in Mining

Data vs. information

- Society produces huge amounts of data
- Sources: business, science, medicine, economics,geography, environment, sports, …
- Potentially valuable resource
- Raw data is useless: need techniques to automatically extract information from it
- Data: recorded facts
- Information: patterns underlying the data

What is Data Mining?

- Data mining is the process of non-trivial extraction of implicit, previously unknown and potentially useful information from data stored in repositories using
- pattern recognition technologies, as well as
- statistical and mathematical methods

Data Mining

- Needed: programs that detect patterns and regularities in the data
- Strong patterns can be used to make predictions
- Problem 1: most patterns are not interesting
- Problem 2: patterns may be inexact (or even completely spurious) if data is garbled or missing

Machine Learning Techniques

- Supervised learning
  - Basically a synonym for classification
  - Supervision comes from labelled examples in the training data set.
  - Eg. Programming a computer to automatically recognise postal codes:
    - Labelled examples = a set of handwritten postal code images + their corresponding machine-readable translations
    - Used as training examples which supervise the learning of the classification model

Machine Learning Techniques…

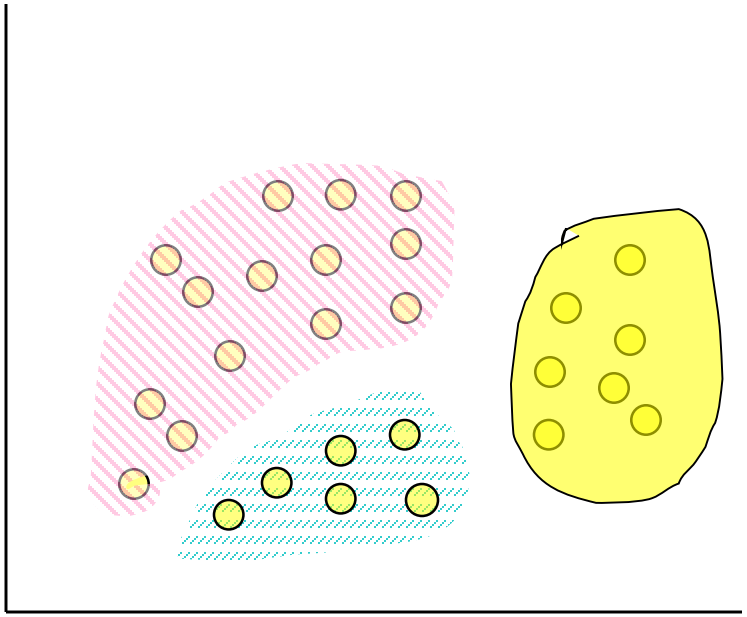
- Unsupervised learning
  - Basically a synonym for clustering
  - Unsupervised as the learning examples are not class labelled.
  - Typically use clustering to discover classes within the data
    - Eg. Input = set of images with handwritten digits. Suppose the unsupervised method finds 10 clusters of data. These may correspond to the 10 digits 0-9, respectively.

    However, since the training data are not labelled, the learned model cannot tell us the semantic meaning of the clusters found.

Clustering

- Elements grouped together according to different characteristics
  - Every cluster shares the same values (homogenous)
- Used most frequently for:
  - Consolidating data into a high-level view
  - Group records into likely behaviors

Clustering



Find “natural” groupings of instances given un-labeled data

Classification - Processing Loan Applications

- Given: questionnaire with financial and personal information
- Problem: should money be lent?
- Simple statistical method covers 90% of cases; borderline cases referred to loan officers
- But: 50% of accepted borderline cases defaulted!
- Solution: reject all borderline cases?

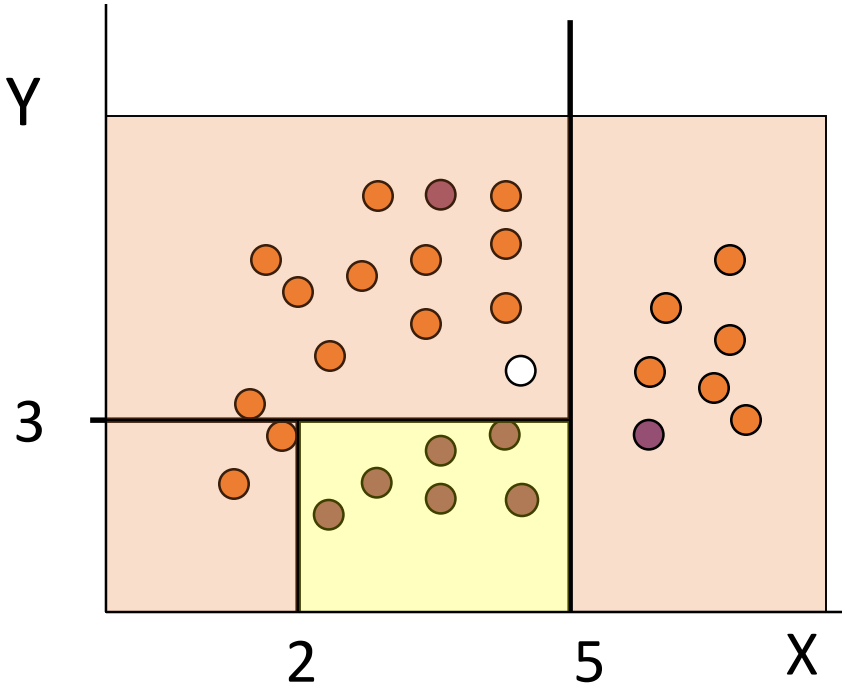
Classification - Processing Loan Applications

- Input consisted of 1000 training examples of borderline cases
- Each example tracked 20 attributes that included: *Age, Years with current Employer, Years with the bank, other credit cards, etc.*
- A machine learning procedure produced a small set of classification rules that made predictions on an independently chosen test set
- Correct predictions were made on around 67% of this test set
- These rules not only improved the success rate, but also helped the company to explain to customers why their applications were declined

Decision Trees

- A way of representing a series of rules that lead to a class or value
- Structure:
  - Decision node, branches, leaves
- Example: A loan officer wants to determine the credit of applicants

Decision Trees…



If X > 5 then accept
Else If Y > 3 then accept
Else If X > 2 then reject
Else accept

Decision Trees…

- Help to induce the tree and its rules to make predictions

Classification vs Association

- Classifiers categorise data into classes by predicting the value of a data attribute

```
If income > 40,000 and debt = high
then accept = no
```

- Association problems on the other hand attempt to find relationships among different data attributes (or items)

```
If windy = false and play = no
then outlook = sunny and humidity = high
```

The Market Basket Problem

- A *market basket* is a collection of items purchased by a customer in a transaction
- The goal here is to track items that are frequently purchased together – hence this falls into the *Association Rules* category
- This will enable Managers to better target customers and thereby increase sales

Market Basket Example

Transid	Custid	Date	Item	Qty
111	201	5/1/03	Pen	2
111	201	5/1/03	Ink	1
111	201	5/1/03	Milk	3
111	201	5/1/03	Juice	6
112	105	6/1/03	Pen	1
112	105	6/1/03	Ink	1
112	105	6/1/03	Milk	1
113	106	6/1/03	Pen	1
113	106	6/1/03	Milk	1
114	201	7/1/03	Pen	2
114	201	7/1/03	Ink	2
114	201	7/1/03	Juice	4

Finding Frequent Itemsets

- Need to first define “frequent” – the *support* of an itemset is the fraction of transactions that contain all the items in the itemset
  - For example, we may be interested in finding all itemsets which have 70% or more *support*
- Thus for our basket, we need to identify all frequently occurring
  - Single items,
  - Pairs of items
  - Triples and so on

Finding Frequent Itemsets

- A simple method is to scan the database as many times as the maximum desired itemset size, N
- For example with N=3, scan the database 3 times, and each scan i, we generate all possible itemsets of size i and then determine whether each of these itemsets are frequent

Frequent Itemsets – Worked Example

- Suppose that we set the *support* level to 70%
- In the first scan we identify {pen}, {ink} and {milk} as frequent itemsets of size 1
- In the next iteration, we extend each frequent itemset with an additional item and generate the following candidate itemsets:

  {pen, ink}, {pen, milk}, {pen, juice}, {ink, milk},
  {ink, juice}, {milk, juice}
- We now scan the database once again and determine that {pen, ink} and {pen, milk} are frequent pairs

Frequent Itemsets – Worked Example

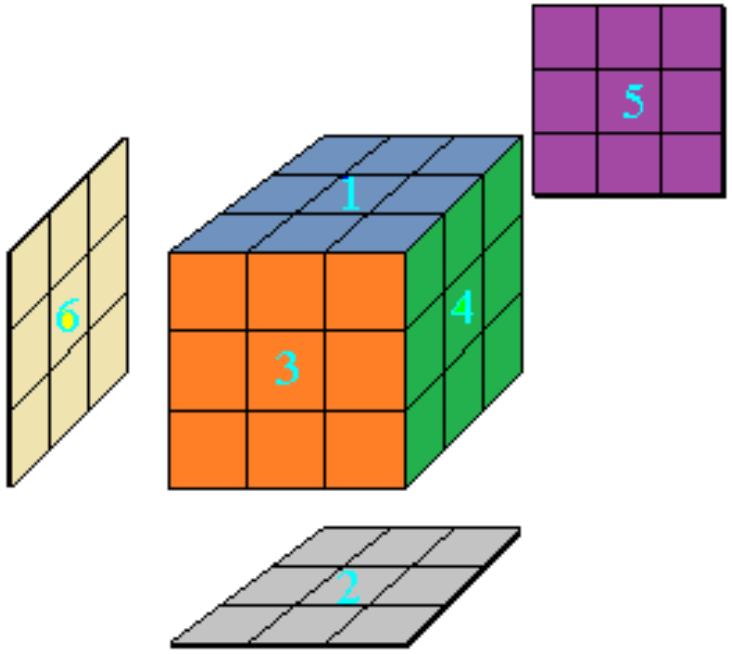
- In the third iteration, we generate the following itemsets

  {pen, ink, milk}, {pen, ink, juice}, {pen, milk, juice}
- A third scan of the database will now establish that none of the above triples are frequent
- Optimizations?
  - A more efficient method would be use the *a priori* property that states that every subset of a frequent subset must also be a frequent subset

Data Mining versus OLAP

OLAP - Online Analytical Processing

- Provides a very good view of what is happening, but cannot predict what will happen in the future or tell us why it is happening



Data Mining versus OLAP

Data Mining

- Originally developed to act as expert systems to solve problems
- Less interested in the mechanics of the technique
  - If it makes sense then let's use it
- Does not require assumptions to be made about data
- Can find patterns in very large amounts of data
- Requires understanding of data and business problem

References

1. Oracle 11g Data Mining
2. Chapter 27, Fundamentals of Database Systems (4th Edition) Elmasri and Navathe
3. Chapter 26, Database Management Systems (3rd Edition) Ramakrishnan and Gehrke
4. Data Mining Concepts and Techniques (3rd edition) Han, Kamber & Pei