Data Mining An Introduction

Read:

Chapter 33 Database Systems A Practical Approach to Design Implementation and Management by Connolly Begg

Chapter 1 Discovering Knowledge in Data by Larose

What is Data Mining?

- "…the process of discovering meaningful new correlations, patterns, and trends by sifting through large amounts of data…" (Gartner Group)
- "…the analysis of observational data sets to find unsuspected relationships and to summarize data in novel ways…" (Hand et al.)
- "…is an interdisciplinary field bringing together techniques from machine learning, pattern recognition, statistics, databases, and visualization…" (Cabana et al.)

Why Data Mining?

- "…we are drowning in information but starved for knowledge." (Naisbitt, author Megatrends)
- Not enough trained analysts available to translate data into knowledge
- Data mining fuelled by several factors
  - Explosive growth in data collection
  - The storage of enterprise-wide data in data warehouses
  - Increased availability of Web clickstream data
  - The tremendous growth in computing power and storage capacity
  - Development of off-the-shelf commercial data mining software products

Elicitation of Knowledge in Data – An Example

“A particular customer transaction at the Tallaght branch on Friday the 14 July 2007 at 9:15am, listed a punnet of strawberries and a carton of cream with a total purchase price of €3.10 which was paid in cash”

- Such a specific collection of ***facts*** is known as a ***case***
- Facts are essential raw material for the elicitation of knowledge from data
- Knowledge can be identified as patterns or regularities in the data
- Patterns (regularities) in data that can be expressed by a statement of the form (IF x THEN y) called a **Rule** e.g.

```
IF a customer purchases a punnet of strawberries
THEN they will also purchase a carton of cream
```

Knowledge and Data ‘contd

- **SUPPORT**
  - The relative frequency where both left and right hand side of the association are found in the database
  - A low level of support may indicate that the pattern is not significant
  - Normally expressed as % or decimal e.g. A=>B

$$\frac{\textbf{No. of transactions containing both A and B}}{\textbf{Total no. of transactions}}$$

- **CONFIDENCE**
  - Given the occurrence condition A (antecedent), how often does condition B occur (consequent)

$$\frac{\textbf{No. of transactions containing both A and B}}{\textbf{No. of transaction containing A}}$$

- **Exercise:** Overall there was 25,000 transactions. If out of 7,000 that included the purchase of strawberries, 5,000 also had a carton of cream Calculate the confidence and support for the rule (already described).

Knowledge and Data ‘contd

A RULE in Association Rules Discovery -An Example

- **The rule describes a simple association within data**
  - **IF** a customer purchases a punnet of strawberries **THEN** they will also purchase a carton of cream (confidence:71%)

- In summary:
  - Association rule does not determine the behaviour of each case BUT an observation that is generally true across all cases
  - Data in a database can be viewed as set of atomic facts about the subjects of interest
  - Knowledge can be viewed as **patterns or regularities** in the data as a whole, which can be expressed as a series of rules that hold in **most** cases

Conventional Data Analysis and Data Mining

- Conventional Data Analysis in eliciting knowledge
  - Hypothesis Verification (human input required)
    - Formulate a hypothesis
    - Test the hypothesis
    - Refine Hypothesis
- Data Mining in eliciting knowledge
  - Hypothesis Generation
    - Hypothesis formulated and refined as part of the process without requiring human input.
    - Data Mining tool formulates and refines the hypothesis (aka Knowledge Discovery)
    - BUT Human Analyst must be able to decide if patterns discovered are significant (statistically)

Conventional Data Analysis and Data Mining

- Conventional Data Analysis Analyst might ask
  - "Did the sales of cream increase in June?"
  - "Did the sales of cream increase when strawberries were available?"
  - "Did the sales of cream increase when there was promotion on strawberries?"
- Using a Data Mining Tool the analyst might just ask the following single question:
  - "What are the factors that determine the sales of cream?"

Hypothesis Verification

- Traditionally, patterns identified using database queries i.e.
  - Formulate an Hypothesis about the data
    - A Candidate Rule (pattern)
    - Convert into one or more SQL queries that are run against the database
    - What are the limitations of this process ?
  - Alternatively, patterns can be identified by statistical techniques

Hypothesis Verification Exercise

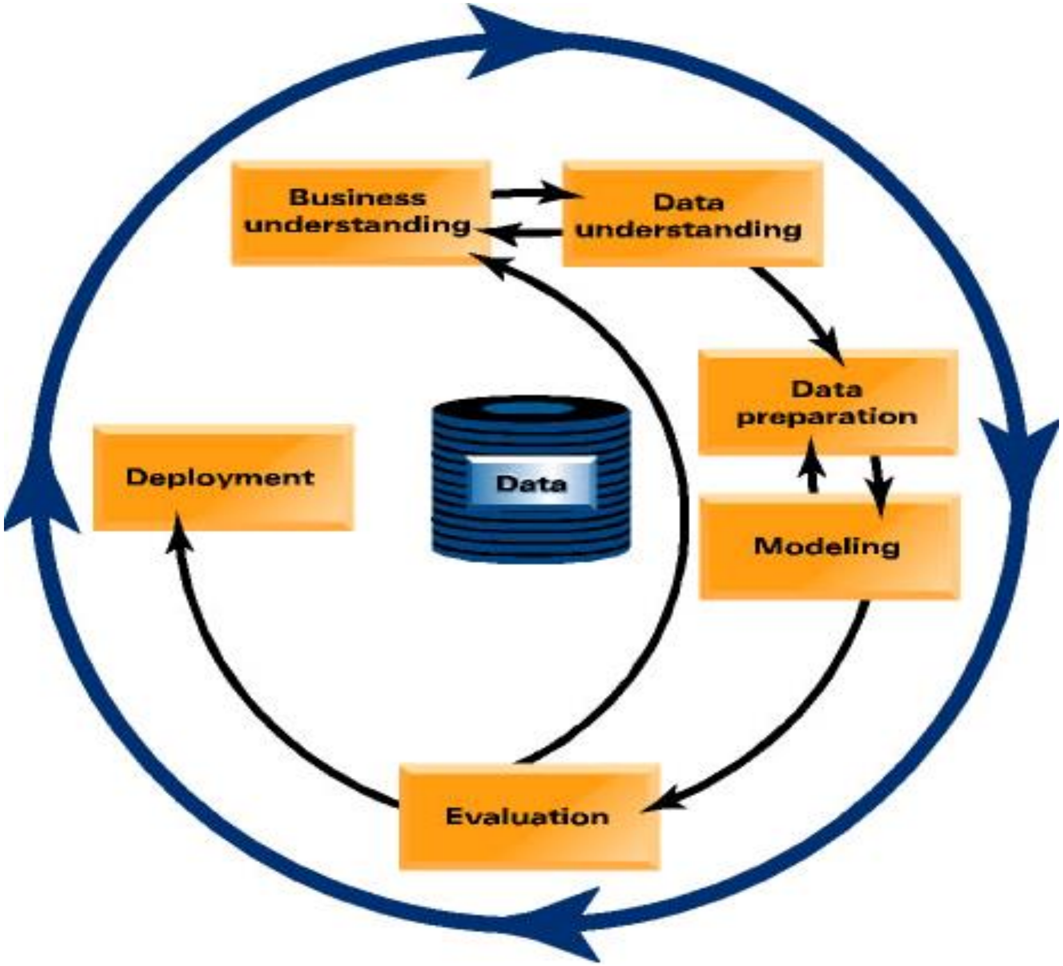
- Suppose a data warehouse for Tallaght Superstores is built according to the star schema provided (from last week). Describe how you would calculate the proportion of cases supporting the hypothesis for the following rule, giving the SQL queries that you used
  - *IF a customer purchases a punnet of strawberries THEN they will also purchase a carton of cream on the same shopping trip*

- Each shopping trip (transaction) is identified by the store_code and time_code columns of the sales table, the purchase of a punnet of strawberries by the description column of the product table having a value of 'strawberries' and the purchase of a carton of cream on the same shopping trip

Methodologies in Data Mining

- There are a number of methodologies for data mining projects:
  - CRISP-DM (**CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining)
  - The SAS SEMMA (**S**ample, **E**xplore, **M**odify, **M**odel, **A**ssess) process
  - KDD (**K**nowledge, **D**iscovery in **D**atabases) process
- Why use a methodology?
  - Framework for recording experience
    - Allows projects to be replicated
  - Aid to project planning and management
- “Comfort factor” for new adopters
  - Demonstrates maturity of Data Mining

CRISP/DM


Business understanding
Data understanding
Data preparation
Modeling
Evaluation
Deployment
Data

The Data Mining Task

- Identify patterns (regularities) in the target data
  - *Nontrivial*, *interesting* and whose *confidence* is above a predefined threshold
- Main goals of the data mining task
  - **Prediction**
    - Involves building a set of rules or a model to predict future values of variables of interest
    - 2 principle techniques
      - Classification
      - Regression
  - **Description**
    - Identifying rules and models that describe the data
    - 2 principle techniques
      - Association Rules Discovery
      - Clustering

The Data Mining Task – The 3 Phases

1. **Training Phase (All Techniques)**
   - Searches a data sample for a set of rules or a model (depending on the technique)
   - Descriptive Analysis Techniques consider <u>all</u> of the cases
   - In contrast, Predictive Analysis Techniques consider a <u>sample</u> of data from some population i.e. a subset of the cases called the *training set that are already pre-classified*

2. **Testing Phase**
   - The models are applied to new known cases
   - By comparing the models' predictions with those of the known cases, the analyst is able to verify that models are still valid in the light of new cases.
   - **Note: Predictive data mining techniques (i.e. Classification and Regression) follow this phase**

3. **Application Phase**
   - The models identified by the training phase are applied to all or individual unseen cases to predict unknown or future values of the data.

Classification Technique

- Classification requires <u>categorical target variable</u> such as *Income Bracket*
  - e.g. Partitioned into three values include “High”, “Middle”, “Low”
- Data model examines records containing input fields and target field
- Table shows several records from data set

Subject	Age	Gender	Occupation	Income Bracket
001	47	F	Software Engineer	High
002	28	M	Marketing Consultant	Middle
003	35	M	Unemployed	Low
…	…	…	…	…

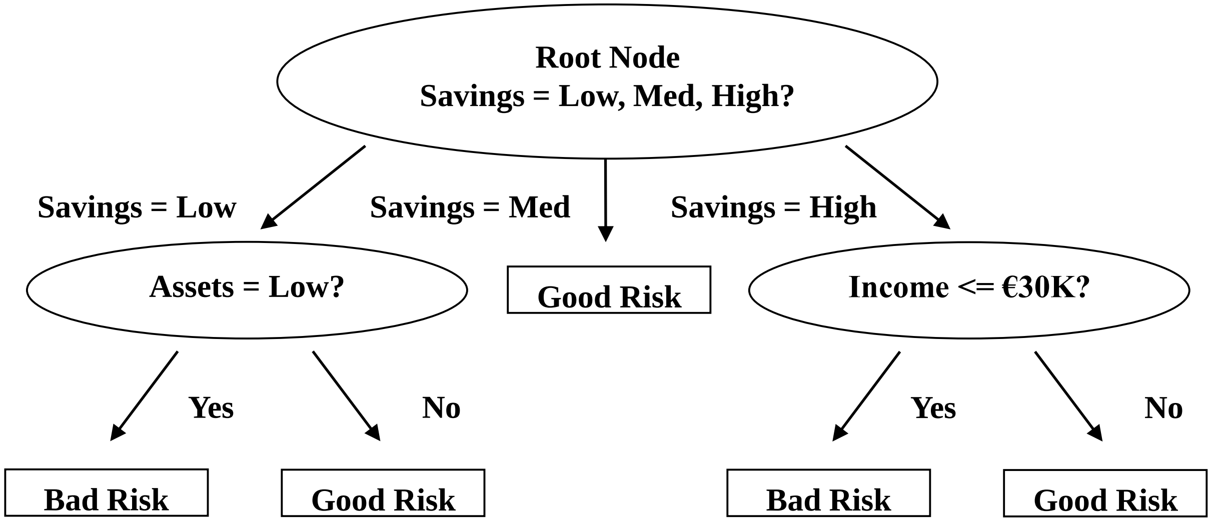
- Examples of Classification Tasks
  - Assessing whether a mortgage application is good or bad credit risk
  - Diagnosing whether a particular disease is present
  - Identifying whether or not certain financial or personal behaviour indicates a possible terrorist threat
- Common Data mining methods for Classification
  - Decision Tree, Neural Network, k-nearest Neighbour

Classification: Decision Trees

- Pre-classified target variable must be included in training set
- The target variable must be categorical
- Decision trees learn by example, so training set should contain records with varied attribute values
- If training set systematically lacks definable subsets, classification becomes problematic
- Classification and Regression Trees (CART) and C4.5 are two leading algorithms used in data mining

Classification:Decision Trees

- Assessing Credit Risk in a Loan Approval Application Example
  - *Credit Risk* is the target variable
  - Customers are classified as either "Good Risk" or "Bad Risk"
  - Predictor variables are *Savings* (Low, Med, High), *Assets* (Low, High) and *Income*



Simple Decision Tree

Decision Trees *(cont'd)*

- Records with *Savings* = "Low" tested at second-level decision node to determine whether *Assets* = "Low"
- Those with low assets classified "Bad Risk, while others classified "Good Risk"
- Second-level decision node in right branch tests whether customers with *Savings* = "High" have *Income* <= $30,000
- Those with *Income* less than or equal to $30,000 classified "Bad Risk". Others classified "Good Risk"
- If no further splits possible, algorithm terminates

CLASSIFICATION– Data Mining Task
The 3 Phases - An Example

- **A bank wants a model that will determine whether a new applicant for a loan is of good credit risk or bad credit risk**
  - Must have a dataset the is already pre-classified i.e. good credit risk or bad credit risk

  Training Phase
  - Take a representative subset (called the training set) of the pre-classified dataset and run a Classification Algorithm e.g. CART over it.
  - Training set is analysed to identify the model that classifies each customer of expressed as rules or a decision tree.

  Test Phase
  - Use the remaining cases (i.e. the test set) in the pre-classified dataset i.e. these test cases are different from the training set.
  - Apply the Model from Training Phase on this Test Dataset.
  - For each test case the known (or expected) class is compared with the predicted class
  - If there is a significant discrepancy between the known and predicted class of the test cases the training phase is repeated with a different training set or a different Classification Algorithm is used.

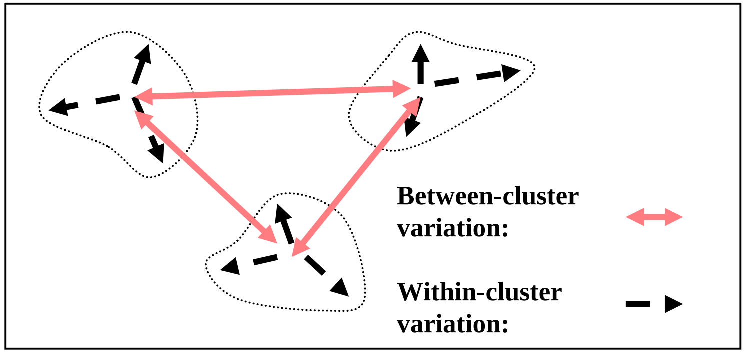
  Application Phase
  - If satisfied with the model identified, it is then used to classify unseen cases
  - Credit Risk predicted using the model that classifies each customer using information on the loan application.

Clustering Technique

- Clustering refers to grouping records, observations, or tasks into classes of similar objects
- Cluster is collection records similar to one another
- Records in one cluster dissimilar to records in other clusters
- Clustering is unsupervised data mining task i.e.no target variable specified
- Clustering algorithms seek to segment records and maximize homogeneity in subgroups
- Similarity to records outside cluster minimized
- Examples of Clustering methods

  K-means clustering and Kohonen

  networks



Clustering Technique (*cont'd*)

- Differs from Classification techniques
  - The Classification scheme is already known with Classification Techniques
- Examples of Clustering Tasks
  - Identification of target markets of a niche product for a company with small marketing budget.
  - From accounting auditing purposes, to segment financial behaviour into benign and suspicious categories

- Clustering is often performed as a preliminary step in a data mining process
  - For example, the resulting clusters being used as further inputs into a Classification technique

Clustering– Data Mining Task The 3 Phases - An Example

- **Training Phase**
  - Identifies data groups with similar characteristics
  - Applied to **whole** dataset
  - Clusters discovered by the technique are used to create rules that characterise the data associated with each group
    - e.g. For the Loyalty Card example using age and income the following customers could characterise the data associated with a particular group of cases
    - IF age> 16 AND age <=25 AND income < €20k THEN cluster=10
- **Application Phase**
  - Records the identity of the cluster that each case belongs to.
    - As cases are usually rows in the database clustering typically means adding a new column to record the cluster
- **Testing Phase**
  - Note there is no Testing Phase required for this technique

Association Analysis

- Find out which attributes "go together"
- Market Basket Analysis commonly used in business applications
- Quantify relationships in the form of Rules

IF *antecedent* THEN *consequent*

- Rules measured using support and confidence
- For example, discover which items in supermarket are purchased together
- Thursday night 200 of 1,000 customers bought diapers, and of those buying diapers, 50 purchased beer
- Association Rule: "IF buy diapers, THEN buy beer"
- Support = 200/1,000 = 5%, and confidence = 50/200 = 25%

Association Analysis *(cont'd)*

- **Generally used in an unsupervised manner**
- **Association Tasks in Business and Research:**
  - Investigating proportion of subscribers to cell phone plan responding positively to service upgrade offer
  - Predicting degradation in telecommunication networks
  - Discovering which items in supermarket purchased together
  - Determining proportion of cases where administering new drug exhibits serious side effects

- Two commonly-used algorithms for generating association rules
  - A Priori and Generalized Rule Induction (GRI)