Master of Technology in Knowledge Engineering

KE5107: Data Mining Methodology and Methods

Data Mining Methodologies

Fan Zhenzhen
Institute of Systems Science
National University of Singapore
E-mail: zhenzhen@nus.edu.sg

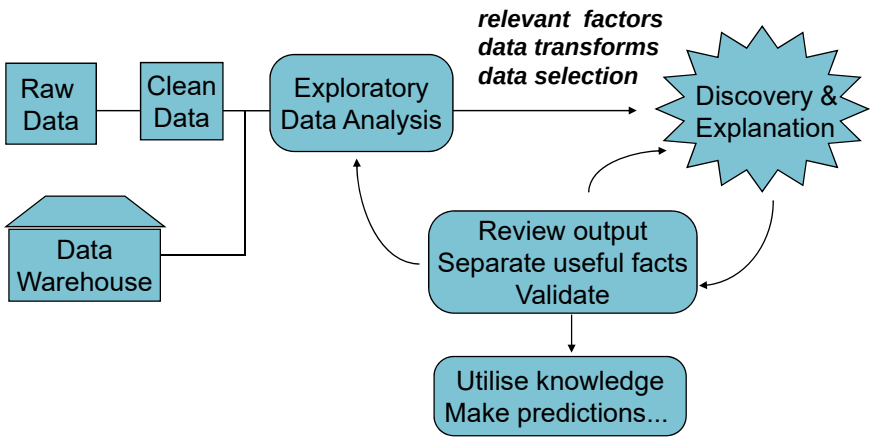
Module Objectives

- To introduce the methodologies for performing data mining based on the business requirements
- To be aware of the common steps in the data mining process

Module Topics

- Data mining methodologies
- CRISP-DM in details
- Agile data mining

Data Mining Process

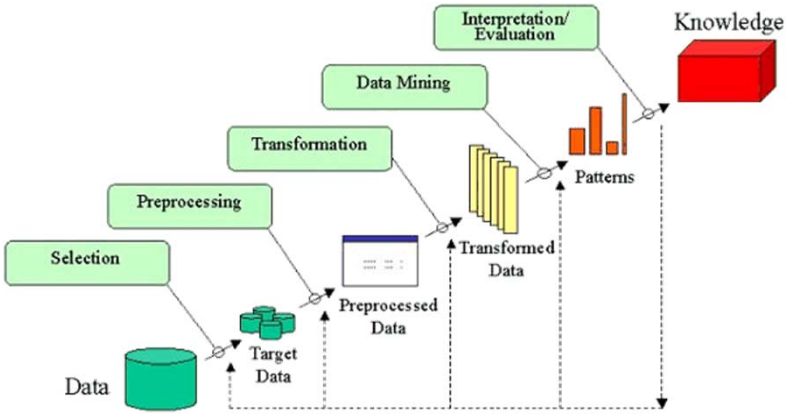


Existing Methodologies In Use

What main methodology are you using for your analytics, data mining, or data science projects ? [200 votes total]
2014 poll 2007 poll

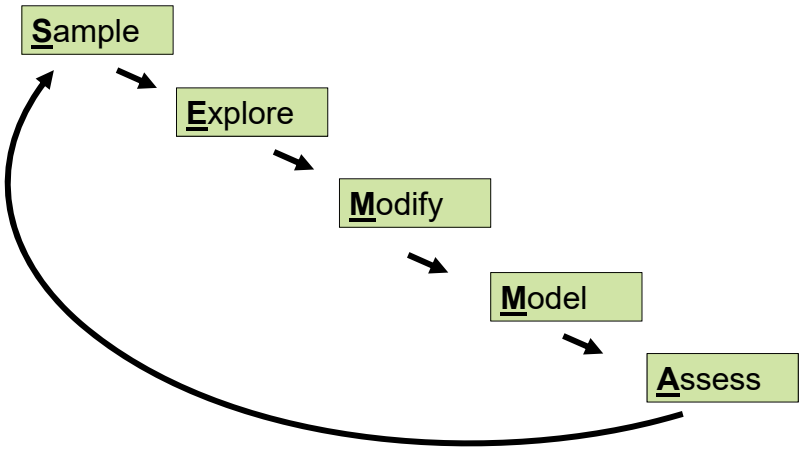
CRISP-DM (86)	43% / 42%
My own (55)	27.5% / 19%
SEMMA (17)	8.5% / 13%
Other, not domain-specific (16)	8% / 4%
KDD Process (15)	7.5% / 7.3%
My organizations' (7)	3.5% / 5.3%
A domain-specific methodology (4)	2% / 4.7%
None (0)	0% / 4.7%

KDD Process



Fayyad et al, 1996

SAS SEMMA

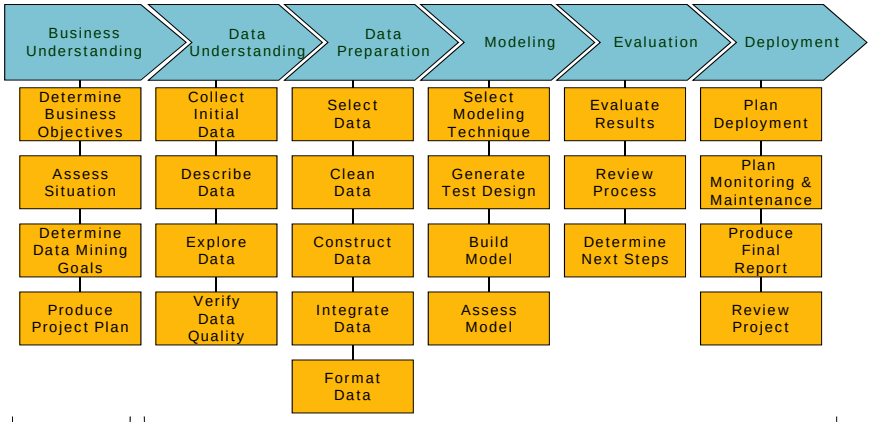


CRISP-DM

- **CR**oss-**I**ndustry **S**tandard **P**rocess for **D**ata **M**ining
- Began as an EC funded project to define a standard process model for carrying out data mining projects (industry neutral and tool neutral)
- Members: SPSS (Integral Solutions), NCR, DaimlerChyrsler, OHRA
- Began July, 1997. The first version of the methodology was completed in June 1998.

CRISP Process Overview

CRISP-DM: Recap



CRISP Process Overview

Data Understanding

- Get familiar with the data:
  - identify data quality problems
  - discover first insights into the data
  - detect interesting subsets to form hypotheses for hidden information
- Steps
  1. Collect initial data
  2. Describe the data
  3. Explore the data
  4. Verify data quality

Collect Initial Data

- List the data that you need
  - What is available?
  - What is the right level of granularity?
  - how much data is needed?
  - over what time period should it apply?
  - what attributes are needed for the data mining goals? What must the data contain?
- Acquire the data.
  - Some data may not yet exist! - Data gathering may be needed
- Resolve inconsistencies between data from different sources, *e.g. different formats for same items (addresses, dates etc)*

Describe Data

- Examine the gross properties of the data
  - How many records, fields, free text fields ?
- Examine the attributes (fields)
  - What do they measure? Are they relevant to the mining?
  - Are they numeric or symbolic?
  - What are their legal values or numeric ranges?
  - Compute basic statistics: averages, variances, skewness (*will help detect anomalies, outliers, suggest hypotheses)*
- Examine relationships/correlations between attributes
  - Are any attributes redundant, duplicated? *(i.e. can omit)*

Exploring the Data

Some data mining questions can be addressed by querying, visualisation and reporting alone:

- Formulate initial hypotheses
- Perform basic analysis to verify hypotheses
- Analyse interesting attributes in detail
- Identify interesting data subsets for further examination

Verifying the Data Quality

- Is the data complete?
- Does it contain errors?
- Are there missing values?
- Do all attributes agree with common sense?
  (e.g. teenagers with high incomes)
- Are there inconsistencies?
- Is there noise? Which attributes are affected?

The results of the Data Understanding phase will feed into the Data Preparation phase.....

CRISP Process Overview

Data Preparation

- Construct the final dataset, for input into the modelling tools, from the raw data.
- Steps
  1. Select data
  2. Clean data
  3. Construct data

Data Selection

Decide on the data to be used for modelling:

- decide which fields to include
- collect additional data if needed
- select data subsets to use
- consider use of sampling techniques to reduce size of data set
- decide if data balancing is required

Data Cleaning

- Correct missing data
  - use defaults, interpolation or other data models
  - can omit faulty records if data is abundant
- Correct, remove or ignore noisy records
- Deal with special values and their meanings
  - e.g. 99 may represent unknown data
  - e.g. 1 = Male, 2 = Female, 3 = ???
- Document all decisions

Construct the data set for Modelling

- Combine data from various sources
- Reformat data to meet the requirements of the modelling tool *(do not change the data meaning)*
- Reorder records or attributes if necessary for a specific modelling tool *(e.g. sort records)*
- Improve the data set by **transforming** existing attributes and/or creating new ones

Data Transformations

- Transforming existing attributes, e.g:
  - Convert number ranges to categories
    Age: 0→18 ~ child, 18 →65 ~ adult, > 65 ~ pensioner
  - Convert text fields into categories
  - Normalise or scale attributes if necessary
  - Turn non-linear variable into a linear one (e.g. log transform)

- Create new attributes by combining existing ones:
  - E.g. If a known relevant fact is absent in the data

    *area = length * width*

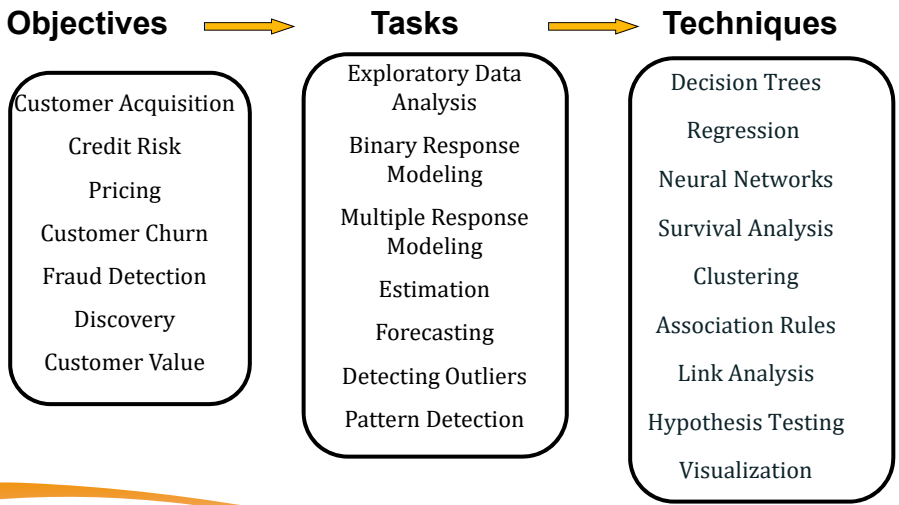
    total cost = costA + costB + costC

CRISP Process Overview

Modelling

- Select and apply various data mining algorithms to build models for prediction, classification, association finding, segmentation etc…
- Steps
  1. Select modelling technique
  2. Generate a test plan
  3. Build the model
  4. Test and Assess the model

Recap: DM Tasks Lead to Specific Techniques

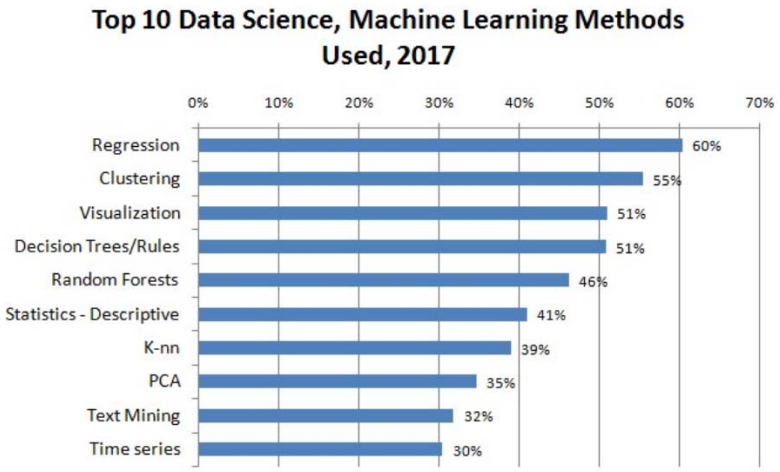


Selecting a Modeling Technique - (not tool selection)

Main techniques	Suitable for
Decision Tree	classification, prediction
Rule Set	classification, prediction
Neural Network	classification, numerical prediction, function estimation
Linear Regression	numerical prediction
Association Rules	dependency analysis
K-NN clustering	segmentation

Will be covered in more details in a later lecture

KDNuggets Poll: Top ML Methods



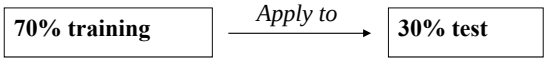
Designing a Test Plan (1)

- How to test and validate the model?
  - Visual inspection?
  - Empirical testing (accuracy)? How reliable is the result?
- Test Plan:
  - divide data into **training** and **test** sets, build a model using the training set and test on the test set
  - measure the performance of the model using a quality measure, e.g. error rate
  - decide how many iterations, individual tests?

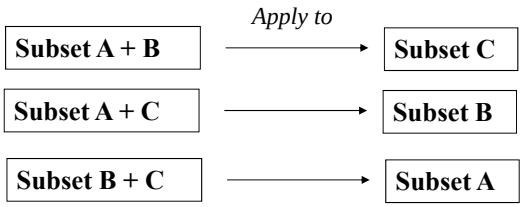
Designing a Test Plan (2)

Divide data into training and test sets. Common methods are:

Simple split:



Cross Validation: split into N equal sized subsets and the testing results are averaged over the rounds

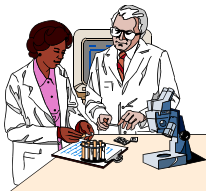


Model building

- Run the modeling tool on the prepared dataset to create one or more models
- Document reasons for your selection of parameter settings
- Experiment with different parameter settings, e.g. pruning levels

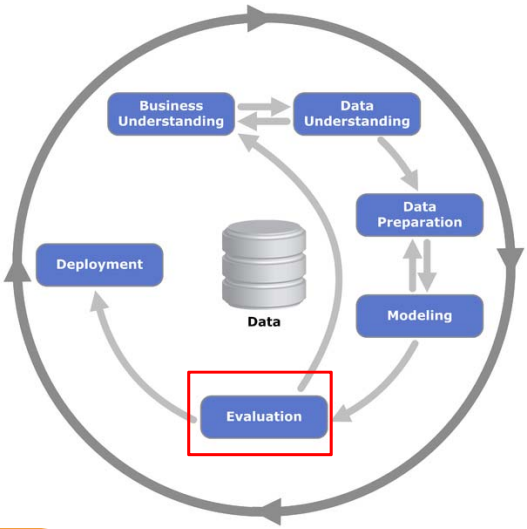
Model Assessment

- Does the model meet the data mining success criteria?
- If the models can be examined (e.g. rules) then do the models make sense?
- Revise parameter settings and iterate model building until you find an acceptable model
- Empirically test each model using the test plan



More details in module *Predictive Analytics and Data Mining*

CRISP Process Overview



Evaluation

- Thoroughly evaluate the model and review the steps taken to build it, ensure it achieves the business objectives.
- Determine if there is an important business issue that has not been sufficiently considered.
- Decide on the use of the data mining results
- Steps
  1. Evaluate results
  2. Review the Process
  3. Decide next steps

Evaluating the results

- Assess the degree to which the model meets the business objectives
- Is there any business reason why the model is deficient?
- Results include more than the model results:-

RESULTS = MODELS + FINDINGS

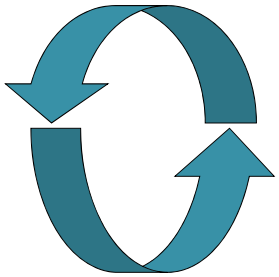
Findings may include things not related to the business objectives but which may unveil information or hints for future directions etc.

Review the process

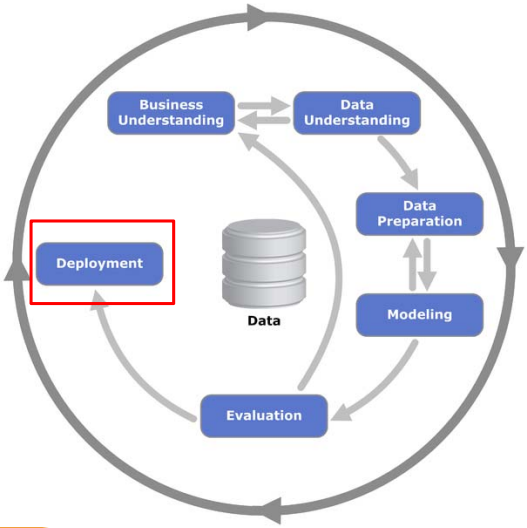
- Has any important factor been overlooked?
- Was the model built correctly
- Have we only used valid attributes and those available for future analyses?
- What are the failures?
- What alternative actions can we take?

Determine Next Step

- Are resources available for more iterations?



CRISP Process Overview



Deployment

- Organize and present knowledge gained in a way that the customer can use.
- Can be as simple as generating a report or as complex as implementing a repeatable data mining process.
- If the customer undertakes the deployment, ensure they understand up front how to use the created models

Plan the deployment

- Produce a deployment plan
  - Making direct use of the generated models, e.g integrate into DSS
- Produce monitoring & maintenance plan
  - what can change in the environment?
  - when should a model be discarded?
    (e.g. drop in accuracy, change in data)

Produce final report

- Write up a final report
  - Identify results
  - State if goals have been met
  - Describe the process followed & deviations from original plan
  - Show costs incurred
  - Make recommendations for future work
- Identify who should receive the report
- Make final presentation (if needed)

Review Project

- What went right and what went wrong?
- What needs to be improved
- Document the whole experience including hints, pitfalls
- Gather feedback

Report Guidelines(1)

- List of contents
- Executive Summary (1 page)
  - Business goals
  - Findings
- Introduction
  - Purpose of Analysis, business & data mining goals
- Main Body
  - Description of each mining step and the results
    - Use Suitable diagrams & summary tables
    - State interpretation & interim conclusions clearly

Report Guidelines (2)

- Conclusions
  - What was found
  - What actions should be taken out
  - What was not found out
  - How the results can be used (describe an implementation plan)
  - Further recommended research
- List of references
- Appendices
  - Details, listings, figures
  - Proper indexing & referencing

Agile Data Mining

- Data mining is an **agile** activity
- Agile concept from the agile software engineering principles
  - Lightweight software development methodology, such as SCRUM, to help cater to today's dynamic and demanding business needs/requirements
  - Iterative and incremental development in short development cycles delivering working softwares
  - Rapid and flexible response to change, evolving requirements and solutions
  - Close collaboration between developers and business people with emphasis on face-to-face conversation

http://www.agilealliance.org/the-alliance/the-agile-manifesto/

Agile Data Mining

- Agility in data mining
  - Evolution or incremental development of the problem requirements
  - Requirement for regular client input or feedback
  - Testing of models as they are being developed
  - Frequent rebuilding of the models to improve their performance
  - Allied pair "programming" with two miners on the same data in a friendly, competitive, and collaborative approach to build models
  - Emphasis of face-to-face communication

Summary

- Data mining is in nature an iterative process.
- It's never conducted in isolation to the business context where the data you mine on is generated.
- Data understanding and data preparation are as important as building the models.
- The results obtained from the data mining process need to be evaluated before being deployed.

Reference

- CRISP-DM 1.0 Step-by-step data mining guide (ftp://ftp.software.ibm.com/software/analytics/spss/documentation/modeler/14.2/en/CRISP_DM.pdf)

CA Information

- Assignments for this course worth 40% in total
  1. Zhenzhen & Fangming (20%)
  2. Barry (10%): Bayesian Networks
  3. Rita (10%): PCA & clustering
- For assignment 1 (20%), you need to work in a team (5±1 students):
  - Find a suitable dataset (online or from work) with at least 5000 records and 50 variables
  - State your mining objectives on the dataset
  - Perform exploratory data analysis to understand your data
  - Prepare your data and build predictive models with appropriate validation
  - Discuss any knowledge discovered from your data through mining

CA Assignment 1

- You should use R as the main tool for mining and follow CRISP-DM in the process.
- You need to submit into IVLE:
  - A project report
  - The dataset (both raw and clean)
  - R codes
  - Other supporting documents (if any)
- Deadline for submission: Mar 12, 2018 (Monday)