Data Mining

= finding useful patterns and relations in large amounts of data (large datasets) in an automated way.

Applied in:

- Medical world: determine the probability of diseases
- Economical world:
  Who will accept a certain offer?
  Average expenditure for these persons?
  Fraud, e.g. in insurance
- Information technology:
  Classification of spam – non spam e-mails
  What data packs constitute an attack?
- …

Bordering:

- statistics,
- linear and logical regression,…
- machine learning, artificial intelligence
- neural networks, decision trees,…
- database management

As a result: multiple terminology!

- Variable, characteristic, attribute, “field”

→ column in a dataset

- Observation, “record”

→ row in a dataset

- Output variables, target variables, dependent variables
- Input variables, predictor variables

Differences statistics ↔ data mining

Statistics:

♉ “small” quantities of data (= sample)
♉ same sample for prediction and for reliability of the results
♉ hypothesis tests, confidence interval
→ difficult
→ many restrictions

Data mining

♉ “large” quantities of data
♉ a large sample to “fit” model
♉ second sample for performance efficiency of model
→ much more comprehensible
→ fewer restrictions
♉ **Danger! Overfitting !!** (see later)

Data mining is statistics at scale, speed
and simplicity!

Most important ‘domains’ for data mining:

Classification

♉ Predicting class
♉ Based on data with known class
→ ‘fitting’ model
♉ Apply to data with unknown class
♉ Examples:

- Spam – non spam
- Attack on network – no attack

Prediction

- Analogous to classification
- No class, but **continuous** value

Predicting variable

- Examples:
- Average expenditure client
- Prediction of viewer ratings

Affinity Analysis – Association Rules

- What variables are associated with one another?
- Examples:
- Arrangement products supermarket
- Amazon (Market Basket)

Clustering Analysis

- Grouping ‘similar’ data
- Examples:
- Market segmentation-research
- http://www.music-map.com

(Data Exploration – Data Visualization)
(Data Dimension Reduction)

Supervised ↔ Unsupervised learning

Supervised

- Classification and Prediction
- “Training set” where the value of target variable is known

→ is used to “train” model
→ model learns from data in “training set”

- Then ‘tune’ model
- Use model for prediction target variable in new data
- example: linear regression

Unsupervised

- No target variable to make predictions
- Model can’t learn from a “training set” with known target variable
- Examples: clustering analysis

Steps in the data mining process

1. Clearly define the purpose of the analysis

2. Building dataset for analysis
- Random sampling from large database
- Combine data from databases
- Internal and external data

3. Exploring, ‘cleaning’, preparing data
- Missing data (missing values)?
- Outliers?
- Relations between variables?
- …

4. Removing or creating variables
Partition data in training, validation, test set
‘Transform’ variables

5. What data mining task?
- Association, Prediction, Clustering?

6. What data mining technique?
- Regression, Naïve Bayes, Neural net?

7. Implementing and executing data mining algorithms

8. Choosing best algorithm/technique

9. Bringing model in production / translate into decision rules

In Sas: **SEMMA**-methodology:
Sample
Explore
Modify
Model
Assess

Random Sampling

♉ data mining: many variables and records
♉ restrictions on processing capacity, software

→ sampling, smaller data set

Oversampling of rare events

♉ frequent in classification: often ‘0’, few ‘1’

→ possibly random sample with few ‘1’

→ little information on records class ‘1’

→ difficult to train a good model for classification ‘0’ and ‘1’

♉ solution: “oversample” class ‘1’
♉ can be important:

→ missing a ‘1’ can be costly! (see attack computer network)

→ erroneously classifying ‘0’ as ‘1’ is less harmful!

Preparing data
Continuous variable

→ possibly discrete variable

Discrete variable with n classes

→ n-1 binary dummy variables

Outliers

♉ can indicate an error in measurement/input
♉ can have large effects on the model
♉ possibly find explanation for outlier
→ involvement of a domain expert important!

Missing data (missing values)

♉ if few records with missing data

→ remove records

♉ **but** suppose 30 variables, for every record,

For every variable 5% probability missing value

→ probability missing value in given record?

♉ possibly replace by e.g. average

(= ‘imputed value’)

→ downside: no new information for that variable

→ upside: information of other variables will not be lost!

Normalizing data

♉ comparing records can only be done on the same scale

♉ otherwise 1 variable can dominate

→ normalize

→ (value – average)/standard deviation

→ scale: “number of standard deviations of the average”

♉ whether necessary or not depends on the technique

How many records needed for training model?

♉ e.g. Delmater and Hancock for classification

Number of records = at least 6*M*N

M = number of classes

N = number of variables

→ preferably minimise number of variables

Overfitting

♉ use training data to fit model
♉ problem: training data = signal + noise
♉ danger: modelling noise instead of signal

= **overfitting**

♉ there is always a model that does fit training data
♉ but: model needn’t only model training data!
♉ as a result: model not applicable to future data
♉ important: At what point do you stop fitting?
♉ example: linear regression and regression of higher order
♉ the problem can also surface when too many ‘predictor’ variables

→ model better fits with more var.

→ but: possible intake of variables in the model
that aren’t important when applying the model on future data

Training Set – Validation Set – Test Set

"Partitioning"

Specifically for data mining (large datasets!)

- Training set
- use to train model
- adapt model / fit to the data
- danger: overfitting
- Validation set
- study performance of found model
- refining / tuning found model
- danger: again overfitting
- Test Set
- not always present
- compare performance best models for every technique

Partitioning at random or according to a variable

Typical ratios:

- training/validation (60/40)
- training/validation/test (50/30/20)