Chapter 2 Business Problems and DS Solutions

10/25/19 12:31 AM

Transform business problems into data mining tasks

Classification problems (supervised)

- Classify individuals into several exclusive classes
- Very close to scoring models, which estimate the probability

Regression problems (supervised)

- Estimate a value (continuous)

Similarly matching problem (both)

- Identify similar individuals, most common for making product recommendations (based on similarity of individuals to recommend product purchased by another person)

Clustering (unsupervised)

- Group individuals based on similarity but not for any purpose
- Useful in preliminary domain exploration, to observe whether there are natural segments or groups formed, in order to find potential data mining problems

Co-occurrence grouping (unsupervised)

- Identify entities that appear together (usually in transaction), e.g. ground meet and hot sauce
- Can also be used in recommendation system

Profiling (unsupervised)

- Characterize behavior of certain individuals or groups. Behavior can be more than one description
- Often used in anomaly detection such as fraud detection (based on consumption behavior, detect fraud activity)

Link prediction (both)

- Decide whether a link between data should exist or not. E.g. have lots of common friends, suggest they should be friends as well
- Can also estimate the stength of a link

Data reduction (both)

- Reduce size of data, to have smaller set but contain most important information from original data set. Usually will involve trade-off

Causal modelling (supervised)

- Determine the causal relationship, commonly by randomly control

Supervised v.s. Unsupervised Methods

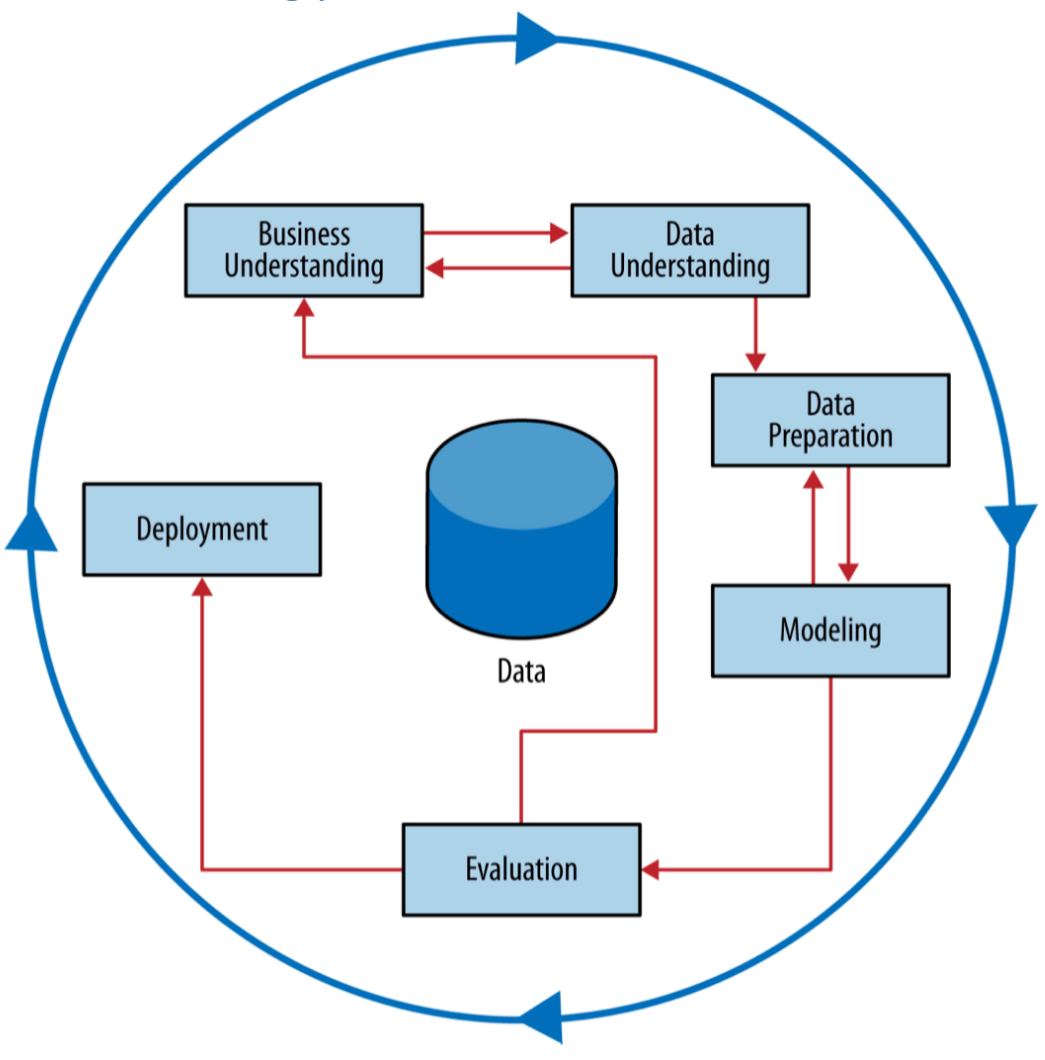
Supervised learning

- requires a specific target to be provided
- Target must exist in data

Unsupervised learning

- grouping but no guarantee to be meaningful for specific purpose

General data mining process



Other Techniques and Technologies

1. *Statistics*

- Summary statistics: sums, averages, rates, etc.
  - Should be chosen with close attention to business problem as well as distribution of data to be summarized
- Statistical hypothesis testing
  - What is the chance that the difference of (e.g. churn rate) is due to random variation?

2. *Database Querying*

- Query: special request for a subset of data or for statistics about the data
- Tools are usually frontends to database systems, based on Structured Query Language (SQL) or graphical user interface (GUI)
- Differs from data mining as there is no discovery of patterns or models

On-line Analytical Processing (OLAP)

- Done in real time, so answers can be found efficiently
- The dimensions of analysis must be pre-programmed into OLAP system (In contrast, SQL is "ad-hoc" querying)

3. *Data Warehousing*

- Collect and coalesce data from across an enterprise
- Can be seen as a facilitating technology of data mining, although not always necessary

4. *Regression Analysis*

- Explanatory modeling v.s. predictive modeling

5. *Machine Learning and Data Mining*

- KDD: Knowledge Discovery and Data Mining
- Machine Learning and KDD shares lots of overlaps
  - Machine Learning also covers subareas not included in KDD, such as robotics and computer vision
  - KDD concerns more about the application of machine learning and the entire process of data analytics

Some sample questions and solution strategies

- Who are the most profitable customers?
  - Database Querying, retrieve a set of customer records from database
- Is there really a difference between the profitable customers and the average customer?
  - Statistical hypothesis testing
- Who really are these customers? Can I characterize them?
  - Database Querying can extract characteristics of profitable customers
  - Also can involve data mining techniques for automated pattern finding
- Will some particular new customer be profitable? How much revenue should I expect this customer to generate?
  - Data Mining techniques to produce predictive models of profitability