ISAD353 - Introduction to Data Mining

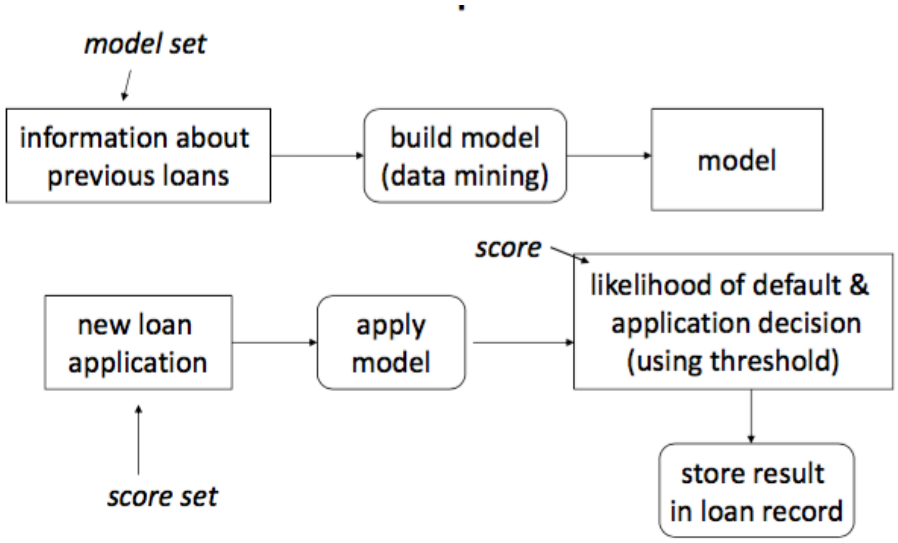
Date: 05-10-16

Data Mining

Examples of data mining

A bank wishes to assess loans application for risk

- The bank has information about previous loans (loans details, customer details and whether or not the loan default)
- Using this information **Data Mining** tries to build a *predictive* model relating the loan & customer details to the target attribute default
- Input information
  - AKA model set
- Model
  - Business logic



Supermarkets with automated checkout

- Vast amount of data collected from purchases
- Want to find **Useful information** from this large amount of data
  - Correlations between products for marketing etc.

What is data mining?

Data mining is the process of explorationand analysis, by automatic and/or semi-automatic means, of large quantities of data to discover meaningful, actionable patterns and rules (that were previously unknown or unexpected)

- Applied disciplain

The prerequisites

- Relevant data sets are being produced
  - Normally in large volumes
  - captured autmatically, and stored in databases and warehouses
- Computing power is available
- The value of *hidden* information is increasingly recognised

Mining vs reporting

Reporting - SQL is getting the infromation that you specifiy in the query

Directed forms of data mining

- In directed mining we have a *prior* view of what we are trying to do/find
  - Defined by a target variable

Classification

- Building a model that enables us to take a new record, and assign it to a class
- The set of classes is pre-defined
- Model set consists of pre-classified records (target value is already known)

Directed forms of mining

- Estimation
  - As in classification, but the target variable takes continuous values
- Prediction
  - Looking to the future

Undirected forms

- No prior view of what we are looking for; no target variable
- Trying to make sense of data collected in various forms

  - Can analysis the data in different ways (*as described below*)
- Trying to understand something from the data
  - Because we think the understanding of the data will be of benefit to us
- Association analysis/rules; link analysis
  - Used to understand which things go together
- Clustering (or segmentation)
  - Clustering a group of individuals into subgroups that are more homogeneous: no predfined classes
- Description
  - To imprive our understanding of the data
  - AKA visualisation

Black vs clear box techniques

- Sometimes we just want the answer
  - Put your input data into the model, get out put and processes is not of interest, only the answer
  - We don't care about the intneranl workings of the model
  - A black box technique is fine
- Sometimes the inner details of the models give useful insights or explanations
  - Clear box
  - Understanding how the answer was arrived at
  - *e.g* Doctors understanding a prescription estimated by a machine

Applications of data mining

- Limited type of data mining approaches
- Many applications boil down to the same thing but under a different guise
  - *e.g.* Directed vs undirected / classification

The data mining cycle

- Identify suitable business problems(s)
  - Where data analysis might provide business value
- Transform data into actionable results
  - Using data mining
- Act upon these results
- Measure the impact of the actions

  `**Repeat the steps**`

- Starts with the real world
- Ends with the real world
  - **Applied**

Some manual or semi manual activities

- Helping form the data using human input to assist decisions made by a machine
- Business insight is needed

Identifying the right business problem

- What is important to the business
- Which segments are of interest
- The relevant business rules
  - Data mining has a habit of finding known patterns
  - Business rules can direct the mining effort
- Is the required data available?
- Is a mining effort necessary?

Transforming data into actionable results

- Obtain, validate & clean data
- Premliminary analysis

Repeat - Choose Modelling technique - Prepare the model set - Build model & evaluate performance

- Pick "best" model(s) and apply to score set

Acting upon the results

- insight
  - Tells you something about the data
  - Very valuable but does not provide you an action to use
- One-time
- Remembered results
  - Something that in the past has proved beneficial
  - Act again on it in the future
- Periodic prediction
  - What will they want to buy next
  - Seasonal
- Real-time

  - Fraud on credit card purchases
  - Scores it as it happens
- Understanding the data can mean that some of the data used is *garbage*

Measuring the impact of the action

- Often overlooked because of its long term value
  - But feeds into the next cycle by highlighting what works and what doesn't