Outline

1. Data Analysis of Industrial Data
   - Analysing Industrial and Commercial Data
   - The Data Analysis Process
   - The Data Preparation and Merging Problem

2. Difficulties in Data Preparation and Analysis
   - Analysing Industrial Data
   - Some Common Difficulties in Data Preparation
   - Examples

3. A Methodology For Data Preparation
   - An Overview of the Data Preparation Methodology
   - Description of Data Preparation Operations

4. Modelling, Validation and Deployment
   - Modelling and Validation
   - Deployment and General Recommendations
The purpose of data analysis and data mining is to extract answers and useful patterns, such as regularities and rules, in data.

These patterns can then be exploited in making predictions, diagnoses, classifications, anomaly detection, hypothesis testing, process structure detection etc.

Several terms describe essentially the same process: Data analysis, data mining, knowledge discovery in databases (KDD) etc.

To some degree, data mining and knowledge discovery can be described as practical application of Machine Learning methods on (often large) data sets.
Working Industrial and Commercial Applications (1)

- Virtual sensors, i.e. an indirect measurement computed based on values that are easier to access.
- Predictive maintenance and weak point analysis through e.g. maintenance and warranty databases.
- Incremental step-wise diagnosis of equipment such as car engines or process plants.
- Intelligent alarm filtering and prioritization of information to operators of complex systems.
Working Industrial and Commercial Applications (2)

- Fraud and fault detection in e.g. data communication systems and eBusiness.
- Sales and demand prediction, e.g. in power grids or retail.
- Speed-up through model approximation in control systems.
- Clustering and classification of customers, e.g. for targeted pricing and advertising, and identification of churners, i.e. customers likely to change provider.
Several attempts have been made to describe the complete analysis process, e.g. process models such as CRISP-DM and SEMMA, as well as several textbooks.

The complete process usually includes:

1. Problem understanding.
2. Data preparation and understanding.
4. Evaluation.
5. Model deployment.

We will first focus on the second step, data preparation and understanding.
Data Analysis and Modeling

- Quite often, the exact choice of model/learning system is not critical.
- Once preprocessing has turned data into something reasonable, simple models may be sufficient.
- With limited amounts of independent data, the number of free parameters must be kept low.
Preparation and Transformation of Data

- The process of collecting and preprocessing data as well as understanding the domain and the problem is a huge and the by far largest part of a data analysis project.
- Preparation of data for analysis is often very time consuming since
  - Available data has been collected for various other purposes and is unlikely to fit the new needs perfectly.
  - Data bases need to be merged, using different encodings and ways of identifying items.
- Efficient methodologies and tools are critical.
**Data preparation**

- *Data preparation* is the process of constructing well structured data sets for modelling.

- The data preparation process includes
  - *Acquisition* of relevant data.
  - *Cleaning* data from *noise, errors* and *outliers*.
  - *Resampling, transposition* and other transformations.
  - *Dimensionality and volume reduction*.
  - *Merging* several data sets into one.
Descriptions of Data Preparation

- The guidelines given in current data mining process descriptions and textbooks are described on a too general level to be useful in practice.
- More detailed descriptions exist, but
  - Do not provide a practical guide on what steps to perform and when.
  - Do not provide more than hints on how to create an application that supports these steps.
  - Focus on corporate and business related data rather than industrial data.
Analysis of Industrial Data

- Generally a mixture of logs from sensors, operations, events and transactions.
- Several completely different data sources, modes and encodings are not uncommon.
- Sensor data will reflect that sensors drift, deteriorate, and are replaced and serviced.
- Anomalies often depend on earlier parts of the process.
- Still, data analysis is critical when
  - Trying to increase production process efficiency.
  - Explaining phenomena that are not completely understood.
Correct typing of parameters is essential for effective modelling.

Plain-text encoding can cause severe interpretation problems.

Often several encodings for one type of parameter.

Encodings may be surprisingly complex and unusable for direct modelling.

Many choices on representation can only be made with a good understanding of the domain or with the help of a domain expert.
Noise, Errors and Anomalies

- Industrial and commercial data usually contain a large amount of noise and errors.
- Some of these errors may be corrected using domain knowledge or other information sources. This can be very important in data bases with few effective samples.
- Remaining errors should usually be removed to ensure correct modelling.
- Noise can sometimes be reduced using redundancy in parameters.
- Noise can appear in the most surprising forms and for very unusual reasons.
Missing values

- Real databases more commonly than not contain missing or invalid values.
- These may be explicitly encoded, but often use a rather ad-hoc representation.
- Missing values are handled in the same way as invalid values:
  - The record can be removed.
  - The value can be derived from other data.
  - The value can be set to a default value.
- Missing or partial data can sometimes be found via external sources.
Data Relevance

- Sensors drift and are replaced. Sales patterns show periodic behaviour.
- There may be sampling bias in the data set.
- A large amount of parameters is no guarantee for successful modelling. What we actually might need to measure may still not be included.
- Validation is often very complicated due to a large and complex state space. We may have very few effective samples.
Examples of difficult encodings / errors

<table>
<thead>
<tr>
<th>Attribute 1</th>
<th>Attribute 2</th>
<th>Attribute 3</th>
<th>Attribute 4</th>
<th>Attribute 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.28272</td>
<td>2002080612220500</td>
<td>10.47</td>
<td>Nej</td>
<td>red</td>
</tr>
<tr>
<td>2.28269</td>
<td>2002080612220622</td>
<td>15.39</td>
<td>Valsbyte</td>
<td>red</td>
</tr>
<tr>
<td>2.28253</td>
<td>2002080612220743</td>
<td>12.66</td>
<td>hasp temp 680</td>
<td>blue</td>
</tr>
<tr>
<td>2.28242</td>
<td>2002080612220886</td>
<td>-120.0</td>
<td>Hasp-temp</td>
<td>green</td>
</tr>
<tr>
<td>22.8232</td>
<td>2002080612221012</td>
<td>-120.0</td>
<td>kyln</td>
<td>blue</td>
</tr>
<tr>
<td>2.28197</td>
<td>2002080612221136</td>
<td>-120.0</td>
<td>Kylning</td>
<td>rde</td>
</tr>
<tr>
<td>2.28157</td>
<td>1858111700000000</td>
<td>13.49</td>
<td>karkylning på alltid</td>
<td>Blue</td>
</tr>
<tr>
<td>2.28212</td>
<td>1858111700000000</td>
<td>25.85</td>
<td>Ja</td>
<td>red</td>
</tr>
<tr>
<td>2.28233</td>
<td>2002080612221631</td>
<td>22.98</td>
<td>ej i dragläge</td>
<td>red</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
The Data Preparation Process

- Generally include a number of steps, some of which have to be iterated several times.

- Data preparation is often revisited quite far into the analysis and modelling process.

- To handle this, a methodology will include a number of
  - Distinct operations.
  - Explicit iterations.

- Intended to provide a structured context to support the data preparation process.
Overview of the Data Preparation Methodology

Figure: Methodology Overview
Preparatory Steps

1. Possibly identification of target parameters
2. Selection of data sources
3. Problem Classification
4. Unit standardisation
5. Determination of case identifier

Figure: Preparatory Steps
Analysis of type, range, and format of parameters is often surprisingly difficult.

The analysis may be complicated by e.g.

- Damaged or missing data.
- Ad-hoc encodings.

Should serve as input to the repair and recoding step.
Data generally contain deficiencies and anomalies that can be rectified.

An entry for a particular parameter being out of range or missing may mean that a subset of all other parameters are useless.

Automatic detection of possible errors and data cleaning are highly useful.
Selection

- A refinement of choices made in the selection of data sources.
- After the type analysis and repair steps we are in a better position to assess quality and usefulness of data.
- Reducing the sizes of data sets might be considered.
An Overview of the Data Preparation Methodology

Description of Data Preparation Operations

Merging

- Often the most complex step of data preparation.
- Common merge problems include
  - Different sample rates.
  - Transposition.
  - Event / series representations.
- Differences and errors in textual representations can cause severe problems.
Final Steps

- Often a good idea to revisit data selection at this point.
- Suitable point for data partitioning.
- Introduction of domain knowledge using calculated parameters.
The methodology outlined earlier specifies *what* needs to be done to the data. What remains to describe is *how* to perform the necessary preparations.

- We can identify a small set of generic operations, with a focus on data transformation, of which all data preparation tasks are special cases.
- These operations can be composed into a reproducible sequence of transforms, making it possible to move between preparation steps in a controlled manner.
- The set of operations is for all practical purposes complete, and contains overlap in the functionality of the operations.
Visualisation, Inspection and Other Supporting Operations

- Very important for data preparation and analysis, but outside the scope of this presentation.
- Not only for discovering structure and dependencies, but to get a more basic understanding of what the parameters represent and on what form.
- Examples of useful plots are scatter plots, parallel coordinates plots, histograms, and simple series plots.
- Anomaly detection and structure analysis are critical operations.
The Modelling Phase

- Highly dependant on the choice of model family / parameterisation.
- Practical models often
  - Have a relatively low number of free parameters.
  - Take considerable amounts of domain knowledge into account.
  - Are highly specialised for the specific domain and application.
- Validation is critical.
The Limits of Data-Driven Methods

- Data-driven models by nature need access to relevant historical data.
- Relevant historical data is often not available in significant amounts due to:
  - Significant amounts of noise and errors in data.
  - Specifics of the domain.
  - The fact that the system is new and has not generated any data.
- Knowledge about the domain (e.g. physical structure) must be incorporated into the models.
Validation

- Critical to guarantee that performance will be satisfactory.
- Can be very difficult due to dependant data.
- Choice of validation mechanism is often application dependant, but in general leave-one-out cross validation is recommended if possible.
Deployment

- Not necessarily a step that has to be performed, as the goal of the data analysis process was e.g. to test a hypothesis or increase domain knowledge.
- Usually done into existing application infrastructure, and not quite as often into stand-alone applications.
- Requires a new level of detailed testing on real-world data.
General Recommendations

- Thoroughly understand the problem you are working on and try to understand the process that generated the data.
- Take extreme care with validation.
- Test the application on as much real-world data as possible.
Summary

- Data preparation and transformation is critical for the quality of the data analysis.
- Data preparation and understanding is time consuming and is usually the largest part of a data analysis project.
- Models are often relatively simple and specialised for the application.
Sales prediction for supply chains

Daniel Gillblad    Anders Holst

Swedish Institute of Computer Science
Reliable estimates of demand is very important for effective supply chain management.

We will discuss a decision support system for production, sales, and marketing managers.

Increasing the usefulness of the prediction by providing rich information to the user, e.g. uncertainty measures and information on the basis of the prediction.

Integration of related data sources such as tracking data.
Predicting demand

- Predict customer demand expressed as expected future sales.
- Enhance the flow of products in the supply chain
  - Earlier order placements
  - Less production for stock
  - Better production scheduling
  - Adapting marketing strategies
- Useful in several places within a company, but perhaps mainly for sales managers, sales agents, marketing managers and the persons responsible for production planning.
A decision support system

- Sales predictions are in most cases not suitable for automated decision making
  - Predictions are uncertain, only a human can decide whether they are plausible or not.
  - The predictions should be used for decision support.

- When creating a prediction system for decision support, we need to
  - Increase trust in the prediction - not always perfect predictions but useful anyway.
  - Provide tools for increased understanding.
Constructing a usable tool

- Provide graphs of prediction development.
- Provide uncertainty measure.
- Explanations of predictions.
- Explanations of the data the prediction is based on.
- No “Black Box”: The prediction model should be (to some degree) understandable for the user.
The sales prediction problem

- Although difficult, it is often possible to predict future sales.
- Faces some difficult problems:
  - Very noisy data.
  - Lack of relevant data.
  - Non-representative historical data:
    - Changes in product lines.
    - Changes in market behaviour.
Modeling considerations

- Keep the model complexity low.
- In the model, regard recent sales (or other data) as more important than older sales.
- To make use of historical sales data, group the articles (e.g. categories and segments).
- Use all available informative data.
What data can we use?

- Historical sales data.
  - Very relevant (We can’t do much without it!)
  - Might be difficult to find relevant historical sales data.

- Other information, e.g. sales activities, advertisements, campaigns, weather etc.
  - Encoding problems.
  - Availability of data.

- Marketing investments and brand tracking.
Placing the prediction engine

- Predictions should preferably
  - Be available to many or most users.
  - Use all data available, including the latest information.
  - Be integrated in the users normal work tool.

- Indicates placement as a centrally located server.
  - Always access to the latest data.
  - No need for direct interaction with the prediction engine.
Sales scenarios

- The sales campaign / seasonal scenario
  - Sales agents visit a predefined number of potential customers during a sales campaign.
  - Defined start and end date.

- The continuous order scenario
  - Orders arrive continuously at a varying rate.
Continuous order prediction
Season based sales prediction

Daniel Gillblad, Anders Holst
Sales prediction for supply chains
An example interface
Summary

- Reliable, useful sales prediction is possible, but precision may vary.
- Demand prediction is a useful tool for SCM even with limited precision.
Emulating process simulators with learning systems

Daniel Gillblad  Anders Holst  Björn Levin

Swedish Institute of Computer Science
There is quite often a need to find optimal production parameters in process industries. Many of the settings are scheduled some time in advance, e.g. to produce different qualities. Near-optimal parameter schedules can possibly be found using an optimiser that iteratively tests different scenarios in a simulator, gradually converging to an optimal solution. Unfortunately, although the simulator in question is faster than real time, it might still not be fast enough.
Modeling process behaviour

- Two different ways to model process behaviour:
  - First principles simulator of some sophistication.
  - Approximate the input-output mapping in the process with a mathematical function without considering the physical path.
- The latter can be done with a learning system.
There is a number of implementation choices for the learning system model:

- Model the actual outputs of the process or associate simulator states to objective function values directly.
- Use either real process data or generate data with a simulator for training.
The Jämsänkoski paper mill, overview

TMP 1
5 Refiner Lines
@ 115 t/day each
4 Reject Refiners
Bleaching
Pulp Storage

Mixing Tanks

PM4
121,000 t/yr
900 m/min
532 cm wide
57-80 g/m²

Wet Broke
Dry Broke

Mixing Tanks

PM5
245,000 t/yr
1350 m/min
835 cm wide
49-56 g/m²

Wet Broke
Dry Broke

Mixing Tanks

PM6
325,000 t/yr
1600 m/min
930 cm wide
39-60 g/m²

Wet Broke
Dry Broke

TMP 2
5 Refiner Lines
@ 190 t/day each
3 Reject Refiners
Bleaching
Pulp Storage

Mixing Tanks

Daniel Gillblad, Anders Holst, Björn Levin
Emulating process simulators with learning systems
The selected sub-system

TMP 1
5 Refiner Lines
@ 115 t/day each
4 Reject Refiners
Bleaching
Pulp Storage

TMP 2
5 Refiner Lines
@ 190 t/day each
3 Reject Refiners
Bleaching
Pulp Storage

Tank 2
Tank 3
Tank 4
PM6
Tank 5
Tank 6
Tank 7
Tank 8

Daniel Gillblad, Anders Holst, Björn Levin
Emulating process simulators with learnings systems
The optimisation problem and data generation

- The cost function is constructed so that electricity costs for running the refiners are minimized while maintaining consistency of schedules and tank levels.

- Essentially, the problem can be reduced to predict a number of tank levels at the end of the optimisation horizon from initial tank levels and production schedules.
Data was generated by manually modeling the joint distribution of the inputs and the initial state of the simulator, taking random samples from this distribution and generating data in the simulator based on these.

Control sequences were modeled using Markov processes, simple distributions were generally used for other parameters.

In total, around ten million samples were generated, representing about six years of operations.
Models and data representation

- Three different kinds of models tested:
  - $k$-Nearest Neighbour
  - Mixtures of Naive Bayes models
  - Multi-Layer Perceptrons

- Initial tests with step-by-step recursive predictions were not promising.

- Data was re-coded as *events*.
The test results for the event representation were not as good as expected. So what goes wrong?

<table>
<thead>
<tr>
<th>Tank</th>
<th>Events</th>
<th>Monotonous Events</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MLP</td>
<td>k-NN</td>
</tr>
<tr>
<td></td>
<td>$\sigma$</td>
<td>$\rho$</td>
</tr>
<tr>
<td>2</td>
<td>22.5</td>
<td>0.50</td>
</tr>
<tr>
<td>3</td>
<td>16.0</td>
<td>0.70</td>
</tr>
<tr>
<td>4</td>
<td>15.9</td>
<td>0.48</td>
</tr>
<tr>
<td>5</td>
<td>14.9</td>
<td>0.34</td>
</tr>
<tr>
<td>6</td>
<td>15.8</td>
<td>0.61</td>
</tr>
<tr>
<td>7</td>
<td>22.0</td>
<td>0.54</td>
</tr>
<tr>
<td>8</td>
<td>19.4</td>
<td>0.69</td>
</tr>
</tbody>
</table>
Oscillation of tank levels

Figure: An example of the characteristic oscillations of tank levels

- A modified event representation was used (see results).
Substituting simulators using learning systems

- Test results show that this is difficult on real-world problems.
- Are learning systems a suitable solution?
  - Most “real” processes are described by a system of non-linear differential equations.
  - Such systems will display chaotic behaviour.
  - The time horizon for accurate predictions is likely to be short.
- This may not be a problem, e.g. if mean values or aggregates can be predicted.
- Dividing the problem into smaller sub-problems, each solved with learning systems, may be effective, but usually require domain knowledge.
Summary

- The idea of replacing a slow simulator with a faster learning system is certainly appealing, and can potentially result in much faster parameter optimisation systems.

  However, it is by no means an easy process:

  - Generating suitable training and validation data can be laborious and difficult, which also applies to finding a suitable representation.
  - A learning system might not be suitable for approximating a process simulator.