Inductive modelling of complex systems

Knowledge discovery

Data analytics

Data mining
Environmental Science - controlled experiment

System

Experimental - how do we know?
What happens about that? Collect data. System data analytics - "natural experiment". Known pieces to capture "emergent" behavior. Simulate complex systems using software.
- Cancer
- Genome
- next code
- outcome

- medical

- geographical
- geography

- engineering

- astrophysics

- science

Applications of dark matter
- Recommendations
- Different pricing
- Business
- Social science/politics/natural language
The Tasks:
"Empty space"

System addresses

Clusters

- Clustering/explanation
- anomaly detection

- least

- most

- some

- stimuli
Clustering and Prediction
Let us a
2 reasons to
most dynamic scenario
Outcome doesn't necessarily match the expected state
No workford, but a flag

i rigid units

attitudes

an Siena score for each of the

Similarly in many bin, chromium forms


<table>
<thead>
<tr>
<th>$\omega_1$</th>
<th>$a_i$</th>
<th>$\omega_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\omega_1$</td>
<td>$a_1$</td>
<td>$\omega_1$</td>
</tr>
</tbody>
</table>

- Compute apples vs apples
- Stair doors units

Der. $f(a_i) = M(a_i)$ Standard

Value of $a_i$ means
Cleaning

by their mode

Altematives: replace values of an attribute

normal distribution

But implicit assumption above
<table>
<thead>
<tr>
<th></th>
<th></th>
<th>5172</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>5</td>
<td>7834</td>
</tr>
<tr>
<td>1</td>
<td>9821</td>
<td></td>
</tr>
<tr>
<td>2985</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>3846</td>
</tr>
</tbody>
</table>

- delete the record
- records are plentiful
- delete the attribute

---

Missing data
Missing bit at random

Easiest

Missing at random
- don't answer sensitive questions
- swing dives
- delete cols (or rows) with many missing values

- data imputation
  - fill missing values with the mean of
- what I want to predict
- large attributes

Prediction

Data

consistent with anything else in the column
- some techniques fill in missing values
class label

local attribute is categorical

historical data when I knew the answer

objects
Training to be fast

- Regression

Logistic vs. linear
Out bag sampling

Training set / Test set split

$\frac{20}{30} \rightarrow \frac{1}{3}$

train 80 70 66 
as if fast set $n \delta = \delta \sigma$.

Stage 2

Better than a

Process 1

New

20/3 under

Records

Conditions

\textbf{Training}

\textbf{Decoder}

\textbf{Test}