Data mining observational vs. experimental data

Data mining

- The process of **automatically** discovering non-**trivial useful** information in **large data repositories.**

Two aspects of data mining

- **Predictive**
  - Classification
  - Regression
- **Descriptive**
  - Association rules
  - Clustring
  - Anomaly detection
  - Visualisation

Why data mining?

Scientific answer:

- Huge amounts of data are continuously being collected (GB/h)
  - satellite sensors
  - radar telescopes
  - simulation data
  - DNA experiments

- Traditional statistical methods impractical
- Data mining can help scientists to
  - explore, cluster and classify data
  - formulate hypotheses

Why data mining?

Commercial answer:

- Huge amounts of data concerning:
  - purchases
  - surfing och searching the Internet
  - bank and credit card transactions
- Computers have become and more powerful
- Commercial pressure to provide better and customized services

Data mining?

Exercise:

Give an example of something you did today or yesterday that resulted in data that could be mined to discover useful information.

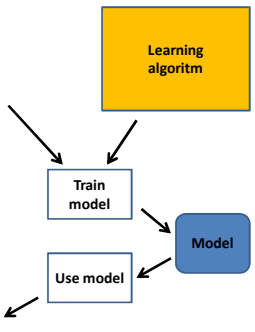
Classification: Finding tax evaders

#	Refund	Marital status	Income	Cheated ?
1	Yes	Single	600K	No
2	No	Married	400K	No
3	No	Single	300K	No
4	Yes	Married	420K	No
5	No	Skild	380K	Yes
6	No	Married	220K	No
7	Yes	Skild	800K	No
8	No	Single	360K	Yes
9	No	Married	240K	No
10	No	Single	340K	Yes

Jim: No refund, divorced, earns 120K?

Classification

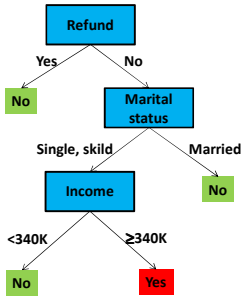
#	Refund	Marital status	Income	Cheated?
1	Yes	Single	600K	No
2	No	Married	400K	No
3	No	Single	300K	No
4	Yes	Married	420K	No
5	No	Skild	380K	Yes
6	No	Married	220K	No
7	Yes	Skild	800K	No
8	No	Single	260K	Yes
9	No	Married	240K	No
10	No	Single	360K	Yes
11	No	Married	250K	?



Decision tree

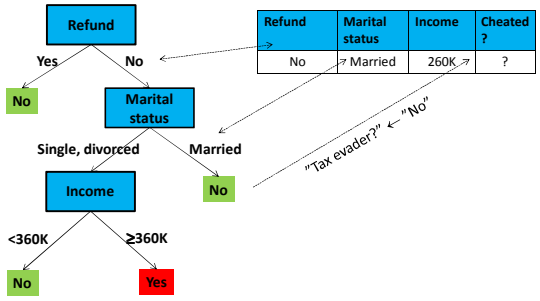
#	Refund	Marital status	Income	Cheated ?
1	Yes	Single	600K	No
2	No	Married	400K	No
3	No	Single	300K	No
4	Yes	Married	420K	No
5	No	Skild	380K	Yes
6	No	Married	220K	No
7	Yes	Skild	800K	No
8	No	Single	360K	Yes
9	No	Married	240K	No
10	No	Single	340K	Yes

Träningsdata



Modell: beslutsträd

Using the tree for classification



Observational vs. experimental data

- Data mining yields **observational** data
- Observational data can be user to infer **correlations** between variables
- Experimental data can be used to infer causal relationships (cause → effect)

Experiments vs. data mining

Experiments:

- We know what we are looking for
  - Formulate null hypothesis
  - Sampling
  - Reject or accept the hypothesis
- Systematically vary the predictor variables and study the effect on the result variable.

Data mining:

- We do **not** know what we are looking for
- Data come from uncontrolled **observations**

Observational data

- Which conclusions can be drawn from the following observations:
  - Autopsies show that deceased patients who have suffered from Alzheimer's disease have high levels of aluminium residues in their brains.
  - Historical data show high levels of CO_2 in the atmosphere during periods of increased average temperature.
  - A questionnaire show that obese persons tend to prefer Coke Light before ordinary Coke.
  - A French consumer organisation reported that owners of red cars were more likely to default on their car loans.

Observational data

- Which conclusions can be drawn from the following observations:
  - Autopsies show that deceased patients who have suffered from Alzheimer's disease have high levels of aluminium residues in their brains.
  - Historical data show high levels of CO_2 in the atmosphere during periods of increased average temperature.
  - A questionnaire show that obese persons tend to prefer Coke Light before ordinary Coke.
  - A French consumer organisation reported that owners of red cars were more likely to default on their car loans.

Observational data

- Which conclusions can be drawn from the following observations:
  - Autopsies show that deceased patients who have suffered from Alzheimer's disease have high levels of aluminium residues in their brains.
  - Historical data show high levels of CO_2 in the atmosphere during periods of increased average temperature.
  - A questionnaire show that obese persons tend to prefer Coke Light before ordinary Coke.
  - A French consumer organisation reported that owners of red cars were more likely to default on their car loans.

Observational data

- Which conclusions can be drawn from the following observations:
  - Autopsies show that deceased patients who have suffered from Alzheimer's disease have high levels of aluminium residues in their brains.
  - Historical data show high levels of CO_2 in the atmosphere during periods of increased average temperature.
  - A questionnaire show that obese persons tend to prefer Coke Light before ordinary Coke.
  - A French consumer organisation reported that owners of red cars were more likely to default on their car loans.

Experiment

- Will a daily dosis of vitamin C lead to fewer infections?
- How can we design an experiment to test this?
- **Suggestion 1:** Do a web questionnaire
  - ”Vitamin C makes me healthier.”
  - ”Vitamin C doesn’t affect my health.”
- **Suggestion 2:** Gather some subjects and have them take a daily dosis of vitamin C during a couple of months. Then evaluate whether the subjects have had fewer days of infection compared to the corresponding period the preceding year.

Experiment

- **Suggestion 3:** Gather some subjects. Let each person decide whether she wants to take a daily dosis of vitamin C (group C) or not (group N). At the end of the trial period, we measure whether group C had fewer days of infection than group N.

Experiment

- **Suggestion 4:** Find some subjects. Let the **experiment leader** decide who is going to have a daily dosis of vitamin C (group C) and who will not (group N). At the end of the trial period, we measure whether group C had fewer days of infection than group N.

Experiment

- **Suggestion 5:** Find some subjects . **Randomly** decide who is going to have a daily dosis of vitamin C (group C) and who will not (group N). At the end of the trial period, we measure whether group C had fewer days of infection than group N.

Experiment

- **Suggestion 6:** Find some subjects . **Randomly** decide who is going to have a daily dosis of vitamin C (group C) and who is going to have a pill that doesn't contain any active ingredient (grupp P). The subjects **do not know** whether they belong to group C or group P. At the end of the trial period, we measure whether group C had fewer days of infection than group P.

Experiment

- **Suggestion 7:** Find some subjects . **Randomly** decide who is going to have a daily dosis of vitamin C (group C) and who is going to have a pill that doesn't contain any active ingredient (grupp P). The subjects **do not know** whether they belong to group C or group P, and **neither does the experiment leader.** At the end of the trial period, we measure whether group C had fewer days of infection than group P.

Design principles for experiments

- Control group
- Randomly select who is part of the experiment group and who is part of the control group
- Placebo
- Double blind tests

Example – Navigation system

- We wanted to design a system guiding a pedestrian P from A to B
- P has a mobile phone with GPS
- It is known from previous studies that people prefer instructions involving **landmarks**
- ”Walk towards the cafe on the corner” rather than ”Turn right” or ”Go north” eller ”Take Sveavägen”.

Example – Navigation system

- On which landmark should we base the instruction?

Example – Navigation system

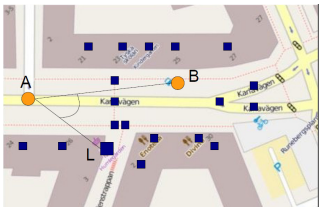
- Data collection:
- A number of subjects walked a predefined route and described the way as they walked (this was recorded).
- We noted which landmarks they used (and did not use) in their descriptions.
- Each landmark could be described by a number of features (size, type, distance, etc.)
- From this information we could construct a mathematical model to predict which landmarks the erson would use in new, unseen situations.

Example – Navigation system

learning instance:

(169822,6.0,4.0,9,1,1,0,0,0,0,0,0,0)
(180196907,0,6.0,109,1,1,0,0,0,0,0,0,0)
(713456474,6.0,0,0,1,0,0,0,0,0,0,0,1)
(928531207,5.0,5.0,40,1,0,0,0,1,0,0,0,0)
(1340895384,6.0,5.0,22,1,0,0,1,0,0,0,0,0)
(1340899520,6.0,4.0,15,1,0,0,0,1,0,0,0,0)
(1340903227,6.0,4.0,13,1,0,0,0,1,0,0,0,0)
(1525463050,6.0,5.0,27,1,0,0,1,0,0,0,0,0)
(1525463077,6.0,5.0,40,0,0,0,0,0,1,0,0,0)
(1755176084,5.0,5.0,25,0,1,0,0,0,0,0,0,0)
(1755176087,5.0,5.0,9,1,1,0,0,0,0,0,0,0)
(1755176089,5.0,5.0,6,0,1,0,0,0,0,0,0,0)
(1755176091,3.0,6.0,82,1,1,0,0,0,0,0,0,0)
(1756229411,5.0,6.0,63,0,0,0,0,0,1,0,0,0)
(1768982670,7.0,5.0,13,1,0,0,0,0,0,0,0,1)
....

L = 928531207



Example – Navigation system

- Was this data collection an example of data mining or was it a controlled experiment?