Statistical Databases and Registers with some datamining

a course in
Survey Methodology and Official Statistics
Pages in the book: 9-18

Department of Statistics
Stockholm University

October 2010

What is data mining

Data mining (short for knowledge discovery from data)

Alternative names may be: **K**nowledge **D**iscovery in **D**atabases, knowledge extraction, data/pattern analysis, data archeology, information harvesting, business intelligence, etc.

Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data

Data mining: a misnomer?

Statistical learning is better, since we learn from data with aid of statistical methods.

Why data mining

Following are reasons for new data methods:

- explosive growth, from terabytes to petabytes
- manual collection systems are replaced by automatic collection systems.
- data is made public and available over the Web

Major sources of abundant data

Business: e-commerce, transactions, stocks, ...

Science: Remote sensing, bioinformatics, environmetal, ...

Society: News, digital cameras, YouTube, ...

We are drowning in data, but starving for knowledge!

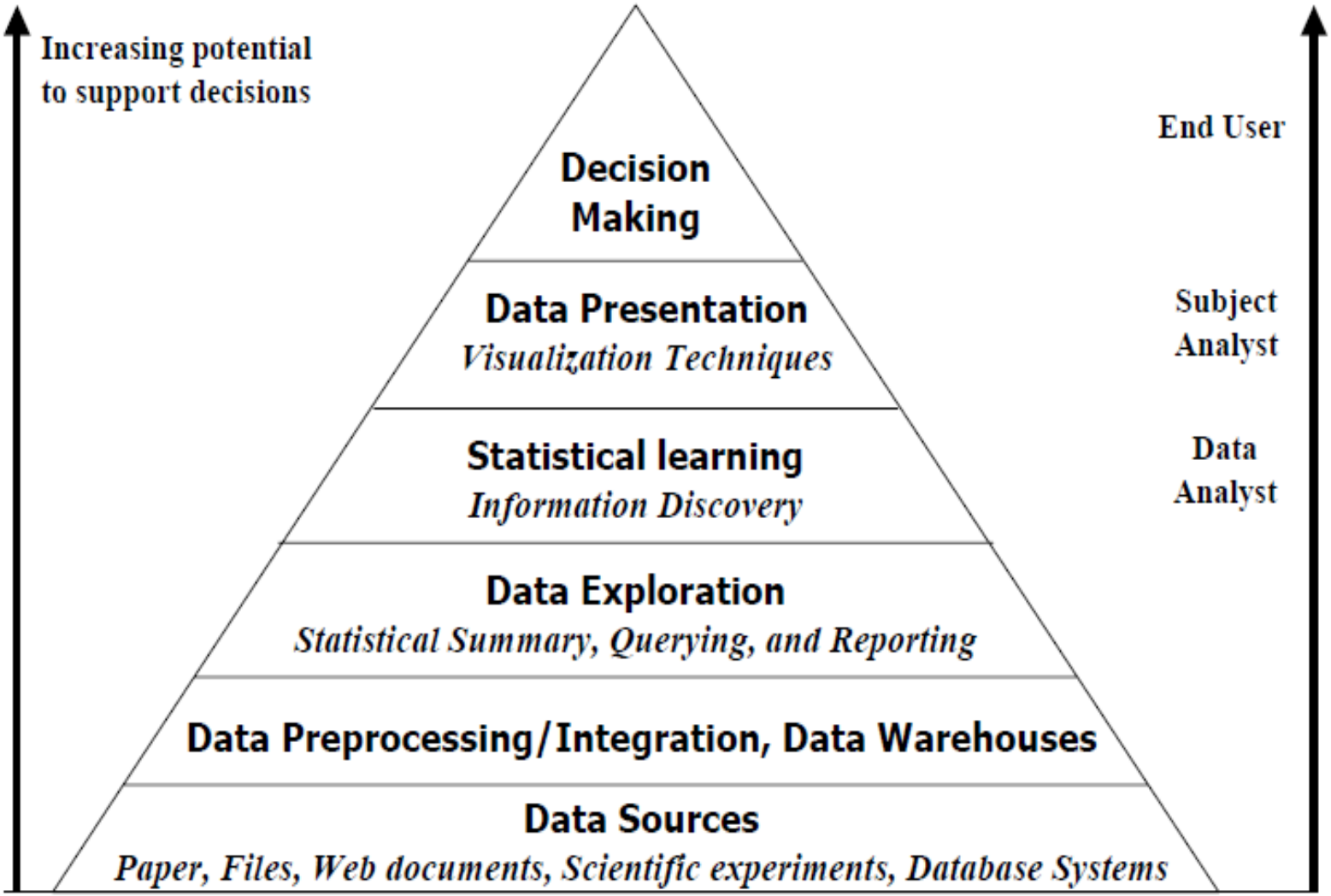
Evolution of database technology

- 1960s: Data collection, database creation, IMS and DBMS
- 1970s: Relational data model, relational DBMS
- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, . . .)
- 1990s: Data mining, data warehousing, multimedia databases, and Web databases
- 2000s:
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems

Why not traditional statistics

- Tremendous amount of data
  - Algorithms must be highly scalable, to cope with tera-bytes of measurements
  - Vectors of high dimensions (up to tens of thousands)
- High complexity of data
  - Data streams and sensor data
  - Time-series data, temporal data, sequence data
  - Structure data, graphs, social networks and multi-linked data
  - Spatial, spatiotemporal, multimedia, text and Web data
  - Software programs, scientific simulations
- New and sophisticated applications

Statistical learning in decision making



The task for statistical learning

The task for statistical learning is: from an overwhelming amount of data find useful patterns.

This is the essence of statistics.

But tradional statistics starts with a model and tries to refute it. If it is not possible to refute the model it is accepted on probable causes.

Statistical learning is the other way around. Learn possible models from data.

Then, due to your objective, different actions will be taken.

Statistical learning examples

A few examples where statistical learning is used

- **Business**: When working with customer data in a local store it was found that many customers on Fridays bought diapers and beers.
- **Physics**: Using readouts of a double-pendulum gives Newton's second law of motion and the law of conservation of momentum

  Eureqa, a program that distills scientific laws from raw data

- **OCR**: To recognize the handwritten post code on letters.
- **PageRank**: How to bring order to the World Wide Web.
- **Medicin**: Identify the risk factors for prostate cancer, based on clinical and demographic variables.

Two types of statistical learning

Supervised learning is when we have an outcome that will guide us in the learning process.
Examples are

- Linear Regression and Nearest Neighbour
- Logistic Regression
- Neural Networks

Unsupervised learning is when we try to describe how the data is organized or clustered.
Examples are

- Cluster Analysis
- Principal Component Analysis
- The Google PageRank Algorithm

Definitions and Notations

Input variables will be denoted by X and output variables by Y. When these variables are vectors we write $\mathbf{X}$ and $\mathbf{Y}$ and when they are matrices we write $\mathbb{X}$ and $\mathbb{Y}$.

We use small letters to signify observations eg $\mathbf{x}$.

When the output variables are quantitative we talk about *regression* and when they are qualitative *classification*.

The goal is to use the inputs to predict the values of the outputs.

For this we need *training* data to construct a prediction rule.

Example: Linear Regression

Our prerequisite are Y =income and (X_1, X_2) geographical coordinates.

Our task is to send an advertisement to the areas (ZIP code) where people have more than average income.

A ZIP code in Sweden consists of 5 numbers where each digit signify a more detailed division.

Eg Stockholm starts with a 1 and the ZIP code for Kungsgatan 32-54 is 11135.

We know who lives at 11135 and we also know their income

How to choose these areas of prospective customers?

Example (cont): Linear Regression

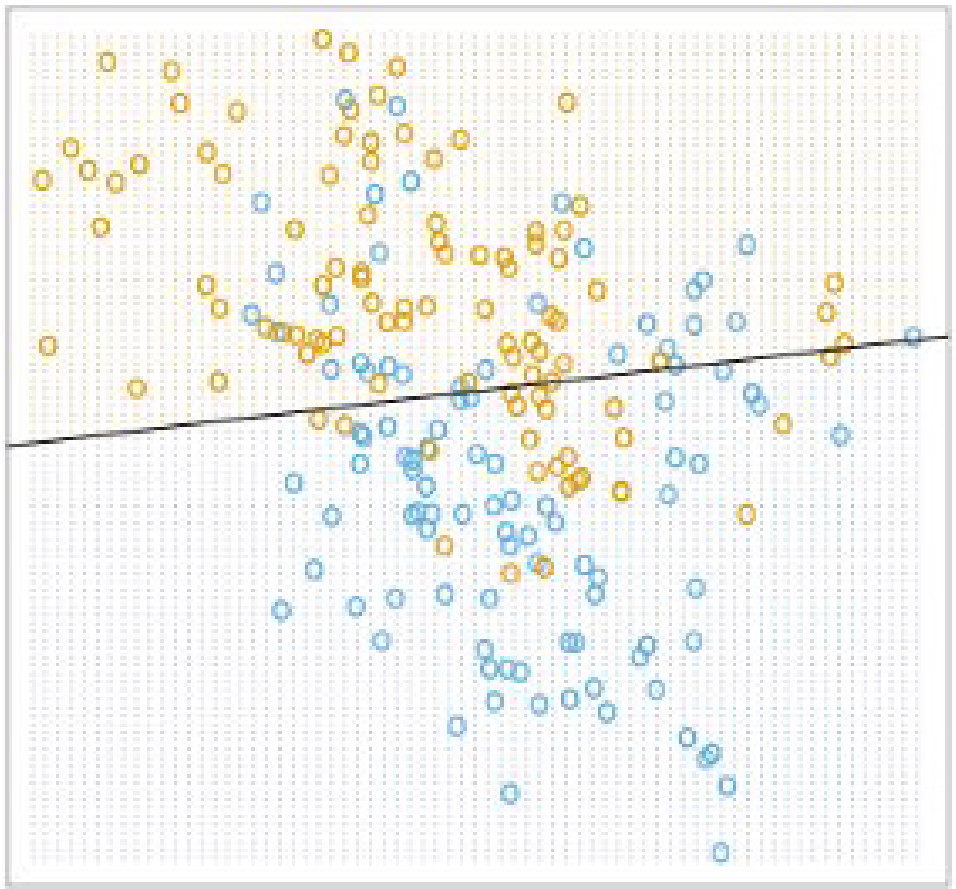
As a model we choose

$$\mathbf{x} = (x_1, x_2)$$

$$y = \begin{cases} \text{orange} \\ \text{blue} \end{cases}$$

Where orange if $x^T \hat{\beta} > 0.5$[1] and blue otherwise.

Here orange mean prospective customer.



[1] There is nothing sacred with the number 0.5. Choose whatever you want.

Example (cont): Linear regression

From the picture it is easily seen that there is a crude division between orange and blue circles.

To find the decision border we assume the model

$$Y = \mathbf{X}^T \boldsymbol{\beta} + \boldsymbol{\epsilon}$$

and the classification

$$G = \begin{cases} 1 & \text{orange} \quad \text{if } \hat{Y} > 0.5 \\ 0 & \text{blue} \quad \text{if } \hat{Y} \leq 0.5 \end{cases}$$

This classification give us the decision border

$$\beta_0 + \beta_1 x_1 + \beta_2 x_2 = 0.5$$

which is seen as the line in the figure.
Our model misclassify a lot of data, on both sides.

Example: Nearest Neighbours

The main idea is, given, $\mathbf{x} = (x_1, x_2)$ find the k closest points $\mathbf{z} \in N_k(\mathbf{x})$ where

$$N_k(\mathbf{x}) = \{\mathbf{z} \mid \|\mathbf{z} - \mathbf{x}\| \text{ are the } k \text{ smallest distances}\}.$$

Then we compute the average for these points

$$\hat{Y}(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i.$$

Lastly the point $\mathbf{x}$ is reclassified, as orange or blue, by the same algorithm as before

$$G = \begin{cases} \text{orange} & \text{if } \hat{Y} > 0.5 \\ \text{blue} & \text{if } \hat{Y} \leq 0.5 \end{cases}$$

Example (cont.): Nearest Neighbours

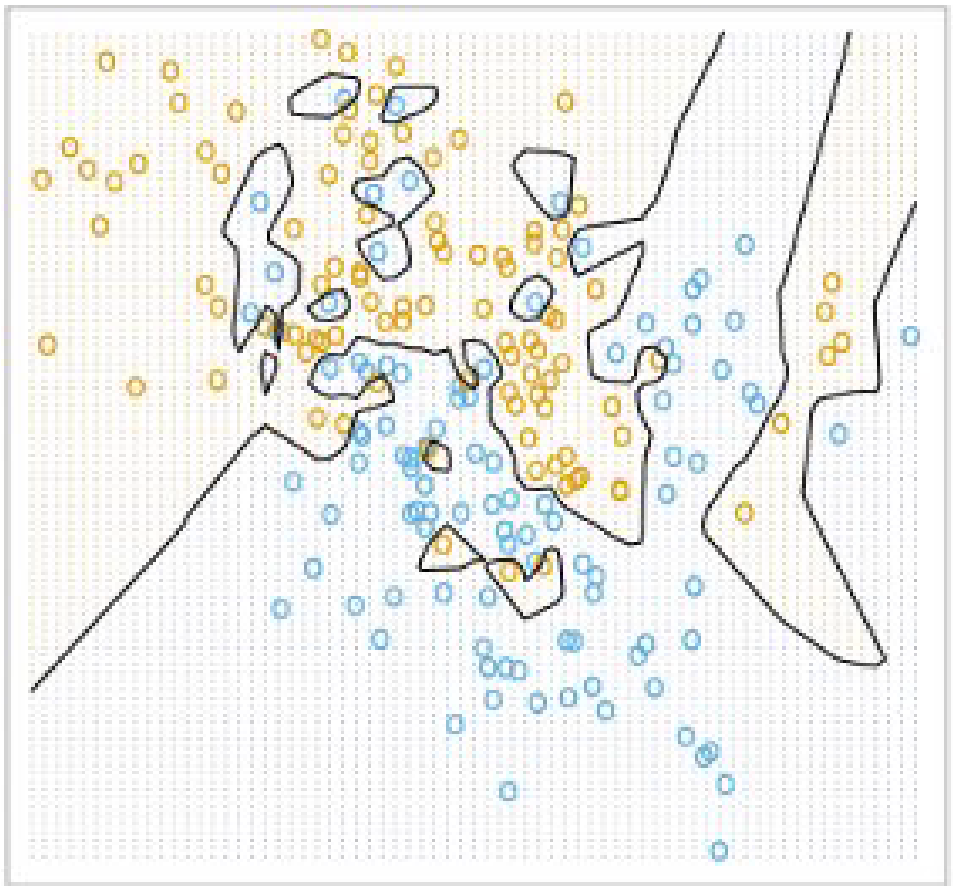
Start with $k = 1$.

Classify each point $\mathbf{x}$ in $\mathbf{X}$ (orange=1, blue=0):

$$\hat{Y}(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i$$

$$G = \begin{cases} 1 & \hat{Y} > 0.5 \\ 0 & \hat{Y} \leq 0.5 \end{cases}$$

Draw a border between points in the dataset



Now every orange circle is separated from the blue circles.

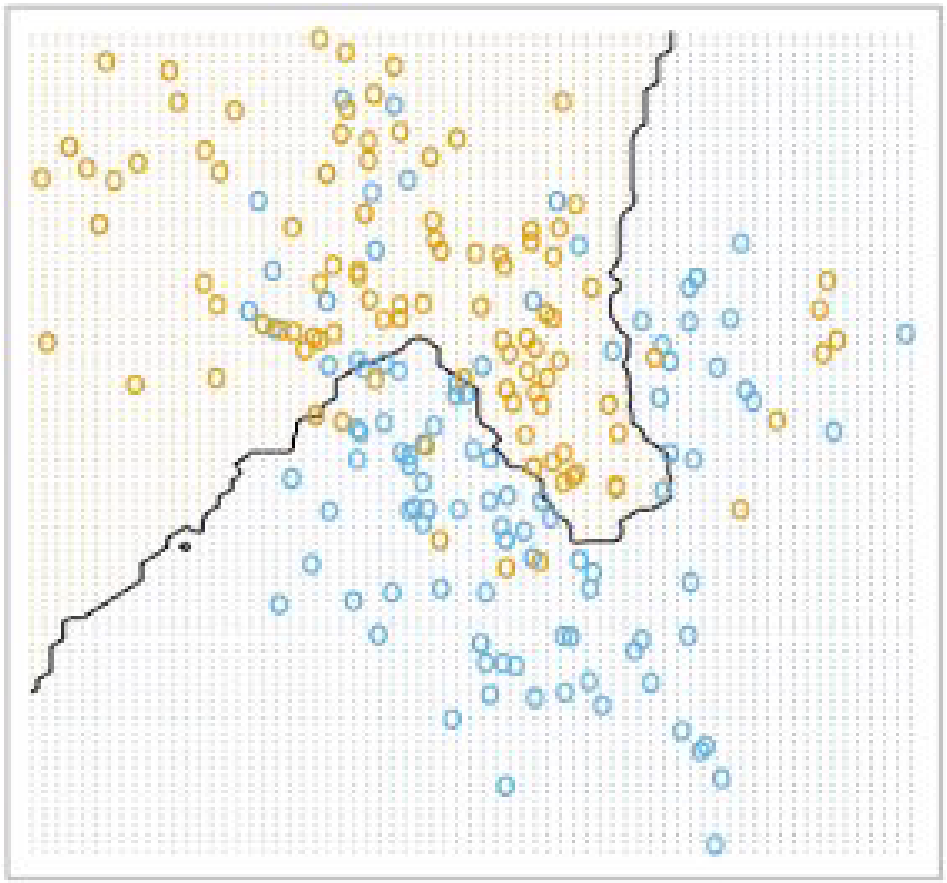
Start with $k = 15$.

Around each point $\mathbf{x}$ in $\mathbf{X}$-space find the 15 closest points.

Classify this $\mathbf{x}$-point (orange=1, blue=0):

$$\hat{Y}(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i$$

$$G = \begin{cases} 1 & \hat{Y} > 0.5 \\ 0 & \hat{Y} \leq 0.5 \end{cases}$$



The border is now the black line, discriminating orange from blue.

This classification is cruder than the previous one but less crude than the first one (the line).

Example (cont.): Nearest Neighbours

Above we used the simple euclidean distance but one is free to use any distance function.

At first sight this model has only 1 parameter, k, compared to the p parameters of the vector $\boldsymbol{\beta}$.

But the effective number of parameters are N/k.

To see this note that if we have exactly N/k distinct neighbourhoods then we would fit one mean per neighbourhood.

Usually $N/k > p$.

Conclusion: Regression and Neighbour

Linear regression

gives a border that do not change much with new observations.

constructs a stable border: Low variance and potential high bias

Nearest Neigbours

gives, for small values of k, borders that are highly dependent on the training set.

gives an unstable border: High variance and low bias.