Data Analysis: A Bird's Eye View

A lecture for SCSIS Birkbeck research students by

Prof. Boris Mirkin

06/11/07 SCSIS Birkbeck

(in My academic interests on my website)

Outline — Data analysis: statistics, machine learning, data mining, knowledge discovery

- Data and feature scales; pre-processing
- Two main problems in data analysis
  - Summarisation
  - Correlation
- Four approaches:
  - statistics
  - machine learning
  - data mining
  - knowledge discovery
- **Examples** (through): Mean, Correlation coefficient
- Methods: global, local, nature-inspired

A Data Analysis Application Involves

- **A. Developing a data set**
- **B. Data pre-processing**
- **C. Applying a method**
- **D. Interpreting results**
- **E. Drawing conclusions**

C is the only item that is treated scientifically,

B and **D** sometimes relate to **C**

A generic data format

Table 0.1. Companies characterized by mixed scale features; the first three companies making product A, the next three making product B, and the last two product C.

Company name	Income, $mln	SharP $	NSup	EC	Sector
Aversiona	19.0	43.7	2	No	Utility
Antyops	29.4	36.0	3	No	Utility
Astonite	23.9	38.0	3	No	Industrial
Bayermart	18.4	27.9	2	Yes	Utility
Breaktops	25.7	22.3	3	Yes	Industrial
Bumchista	12.1	16.9	2	Yes	Industrial
Civiok	23.9	30.2	4	Yes	Retail
Cyberdam	27.2	58.0	5	Yes	Retail

Feature scales

- Quantitative (admits averaging)
- Ordinal (admits ordering)
- Nominal (partition)
- Categorical (not quantitative)

Terminology:
Entity = Instance = Case = Object = Observation
Feature = Attribute = Character = Variable
Value = Score = Grade = Category=State

Two types of data analysis tasks

(Why TWO? - later)

- **Correlation**
  - ***Given:*** (sets of) features **X** (input) and **Y** (target/output)
  - ***Find:*** association **D** between **X** and **Y**
- **Summarisation**
  - ***Given:*** set of features **X**
  - ***Find:*** (a smaller set of factors) **F** representing **X**

Data analysis objectives for the example

- **How to meaningfully map companies to the screen? (Summarization)**
- **Would clustering of companies reflect the product? What features to be involved? (Summarization)**
- **Can rules be derived to make an attribution of the product for an outside company? (Correlation)**
- **The structural features and market related features: any relation? (Correlation.)**

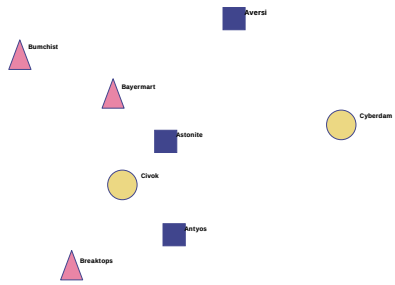
First stage: Preprocessing to quantitative format

Code	Income	SharP	NSup	EC	Util	Indu	Retail
1	19.0	43.7	2	0	1	0	0
2	29.4	36.0	3	0	1	0	0
3	23.9	38.0	3	0	0	1	0
4	18.4	27.9	2	1	1	0	0
5	25.7	22.3	3	1	0	1	0
6	12.1	16.9	2	1	0	1	0
7	23.9	30.2	4	1	0	0	1
8	27.2	58.0	5	1	0	0	1

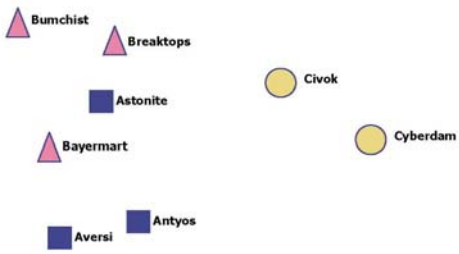
Second stage: Feature standardisation

- $Y_{ik} = (X_{ik} - A_k) / B_k$
  - **X** - original data
  - **Y** – standardised data
  - **i** – entities
  - **k** – features
  - A_k – shift of the origin
  - B_k – rescaling factor
- **Statistics**: **z-scoring** - **A**-mean, **B**- standard deviation
- **SVM**: **A** –midrange, **B** – half-range
- **Clustering: A** - mean, **B** - half-range (Mirkin 2005)

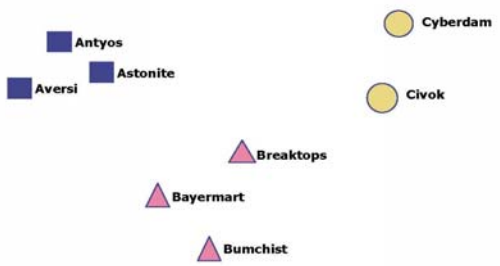
Standardisation effect I: No normalization (B=1)



Standardisation effect II: Z-Scoring



Standardisation effect III: cluster-wise (Mirkin 2005)



Here: both visualization and clustering are ok

Data mining perspective

- **Goal: F**inding Patterns and Regularities within the Data (not quite a scientific perspective)
- **Principles:**
  - **Heuristics (**properties of methods to explore)
  - **Data recovery**
- **User-friendly:** No need to understand methods, only patterns matter
- **Re-sampling based confidence:**
  - Bootstrapping
  - Cross-validation

The mean in data mining perspective

- Heuristic: just a central value
  - Given real $x_1, x_2, ..., x_N$, compute

$$c = \sum_{i=1}^{N} x_i / N$$

- Data recovery:
  - Given real $x_1, x_2, ..., x_N$, find c such that $x_i = c + e_i$ $(i=1,..., N)$ to minimize the sum of squares, $e_1^2 + e_2^2 ... + e_N^2$?

Correlation coefficient in data mining perspective

- Heuristic: just **cosine** between z-scored feature vectors
  - Given **z-scored** pairs $(x_1,y_1), (x_2,y_2)...,(x_N, y_N)$ compute the average product

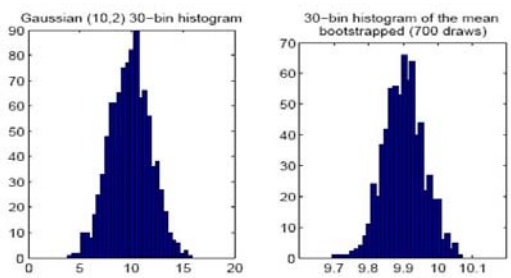
$$r = \sum_{i=1}^{N} x_i y_i / N$$

- Data recovery:
  - Given pairs $(x_1,y_1), (x_2,y_2)...,(x_N, y_N)$ find a, b such that $\boldsymbol{y_i = ax_i + b + e_i}$ $(i=1,..., N)$ to minimize the sum of squares, $\boldsymbol{e_1^2 + e_2^2 ... + e_N^2}$.
  - Property: **Residual y-variance = = (1-r²)*y-variance**

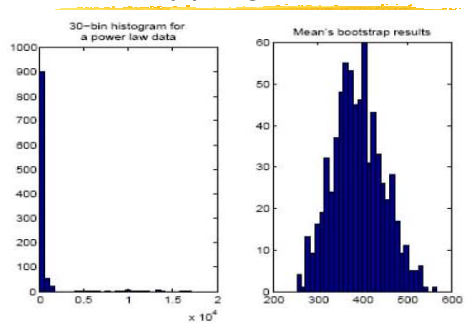
Validity of the mean with Boot-strapping I

- **Bootstrap:**
  - Trial: select randomly, with replacement, N entities, and compute the mean on this sample
  - Consider distribution of the means over 500 or 5000 trials
  - Find boundaries for the mean with a pre-specified confidence level, say 95%

Validity of the mean with Boot-strapping II



Validity of the mean with Boot-strapping III



Classical statistics perspective 1

- Data from a probability distribution
- Goal: Estimate its properties or parameters

- The user must know the model and methods

- Questions: "What regression curve is?"
  "How many clusters are out there?",
  "How data should be pre-processed?"
  are well substantiated

Mean

Given: x_1=1.2, x_2=1.3, x_3=1, x_4=1.1, x_5=1.4, x_6=1.2,…

Model: $\mathbf{x=a+e,\ e \sim N(0, \sigma),\ p(u, 0,\sigma)= c^* \exp\{-u^2/\sigma^2\}}$

Method: Max likelihood $\mathbf{\Pi_i \exp\{-(x_i-a)^2/\sigma^2\} \longrightarrow max}$
Least Squares $\mathbf{\Sigma_i (x_i-a)^2 \longrightarrow min}$
Solution: **a** the average

Statistical properties for testing statistical hypotheses

Data pre-processing: **z**-scoring, $\mathbf{z=(x-a)/\sigma}$

Correlation coefficient

Given: number pairs $\mathbf{w_i=(x_i, y_i),\ \ i=1,…, N}$
Model: $\mathbf{w = m + e,\ e \sim N(0, \Sigma),}$ density $\mathbf{p(u, 0,\Sigma)= c^* \exp\{-u^T \Sigma^{-1} u\}}$

$$\mathbf{\Sigma} = \begin{vmatrix} \sigma_x & \rho \\ \rho & \sigma_y \end{vmatrix}$$

Method: Max likelihood $\mathbf{\Pi_i \exp\{-(w_i-m_i)^T \Sigma^{-1} (w_i-m_i)\} \longrightarrow max}$
Equivalently, Least Squares $\mathbf{\Sigma_i (y_i-ax_i-b)^2 \longrightarrow min\ a, b}$
Solution: $\mathbf{\rho = a\sigma_x / \sigma_y}$ or
$\mathbf{\rho}$ = the average scalar product of **x** and **y**, **z**-scored

Statistical properties for testing statistical hypotheses

Data pre-processing: **z**-scoring, **z=(var - av) / std**

Classical statistics perspective 2

- **Great Distribution Based Principles: Hypothesis Testing & Statistical Confidence**
- **Great Induction Principles:**
  - Maximum Likelihood
    (Least-Squares, Least- Moduli)
  - Minimum Description Length
- **Troubles:** when data relate to a phenomenon of which not much is known. Response: Non-parametric concepts. Challenge: The existence of a distribution is questionable. Response: Test against disorder – non-satisfactory.

Machine learning perspective

- **Goal: Deriving Prediction Rule from Data**
- **Principles:**
  - **Statistics**
  - **Minimising structural risk (SVM)**
- **Re-sampling (**to substantiate methods)
  - **K-fold cross-validation**
  - **Jack-knife**

Knowledge discovery perspective :

- **Goal: Improving knowledge**
- **What is Knowledge**: Structurally,

Categories and **Statements** relating them

  - **Summarisation: Generating new categories (factors, clusters)**
  - **Correlation: Generating new relations (decision rules, regression)**

Methods

- **Global**: Linear - PCA or Simple Combinatorial – MST **(Rarely work)**
- **Local**: Hill-climbing (NN Back Propagation), Alternating Optimisation (Expectation-Maximisation), Neighbourhood Search Heuristics **(No deep minima)**
- **Nature Inspired** - Population Driven
  - Genetic Algorithms
  - Evolutionary Algorithms
  - Particle Swarm Algorithms

(Great expectations)

Conclusion. Bird's Eye View of DA:

- **Generic data types: feature table, network graph (as well as signal, image, video) – pre-processing is a must**
- **Two major problems:**
  - Summarisation
  - Correlation
- **Different perspectives and validation:**
  - classical statistics
  - machine learning
  - data mining
  - knowledge discovery
- **Methods:**
  - Global
  - Local
  - Nature Inspired