Intelligent data Data Processing

Maria-Amparo Vila

vila@decsai.ugr.es

Investigaci'on group Databases and Intelligent Systems Informaci'on https://idbis.ugr.es/

Department of Sciences of the computation and artificial intelligence

University of Granada

Master in Ingenier'ıa

Introduction to the topic

Presentaci' structure on

one. Introduction b'asicas ideas about data

two. Type of data

3. Quality problems

Four. Exploraci'on data

- 4.1 exploration Estad'ıstica
- 4.2 Visualizaci'on data

5. Data Transformations

6. Reducci'on problems of variables.

- 6.1 Selecci'on of variables
- 6.2 Change of coordinates: main components

7. Scaling problems.

Introduction- data

Input data

M'as data structure common to work with DM is the

Dataset

items variables	V_{one}	V_{two}	. . . . .	V_N
or_{one}	d_{eleven}	d_{12}	. . . . .	$d_{one\ N}$
⋮	⋮	⋮	. . . . .	⋮
⋮	⋮	⋮	. . . . .	⋮
or_M	$d_{M\,one}$ $d_{M\,two}$		. . . . .	d_{MN}

Introduction- data

Input data

M'as data structure common to work with DM is the

Dataset

items variables	V_{one}	V_{two}	. . . . .	V_N
or_{one}	d_{eleven}	d_{12}	. . . . .	$d_{one\ N}$
⋮	⋮	⋮	. . . . .	⋮
⋮	⋮	⋮	. . . . .	⋮
or_M	$d_{M\,one}$ $d_{M\,two}$		. . . . .	d_{MN}

- items represent cases, objects etc.
- Variables can be of many types. Also they called factors
- There may be missing data

Introduction- data

Input data

The data set may be obtained from previous data, through transformations, res' umenes etc. In some cases

This is a key point (selecci'on of factors, text mining etc.)

Introduction- data

Input data

The data set may be obtained from previous data, through transformations, res' umenes etc. In some cases

This is a key point (selecci'on of factors, text mining etc.)

There are problems that the structure of data set is not suitable:

- transactional structures
- Miner'ıa graph (structural patterns are sought)
- Miner'ıa sequence (Biocomputaci'on)

Introduction- data

Input data

The data set may be obtained from previous data, through transformations, res' umenes etc. In some cases
This is a key point (selecci'on of factors, text mining etc.)

There are problems that the structure of data set is not suitable:

- transactional structures
- Miner'ıa graph (structural patterns are sought)
- Miner'ıa sequence (Biocomputaci'on)

In mayor'ıa cases they can be transformed into another one representations in order to apply the appropriate t'ecnica

From now on, except indicaci'on against, we will focus on the structure of data set

Type of data

num'ericos attributes

Num'erico has a domain which allows aritm'eticas operations.

You can classify in:

- Discretos :

  They are integers or natural numbers. Usually
  results count. Num'ericos computations allow,
  seg' SU domain.
  Not to be confused with categorical attributes changed.
  Game 1 level 2 or 3 It is not one attribute num'erico

Type of data

num'ericos attributes

- continued :

  correspond An ' real numbers.

  Advanced computations allowed statisticians ym'etodos M'as Sometimes they are too detailed and may have rounding problems.

  always they have a starting point (zero) and a scale factor. Seg' one that can be clas fi in:

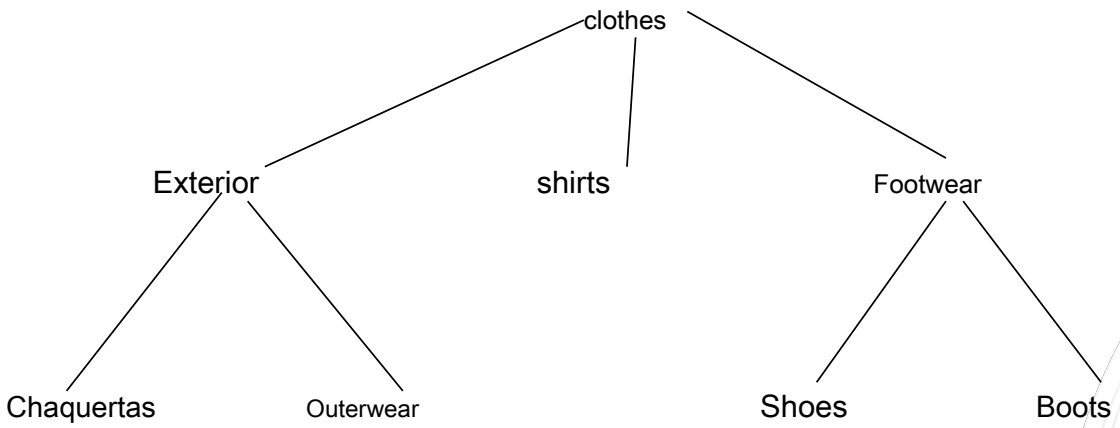
  - "Interval" Zero and scale are arbitrary. (Time in milliseconds and arbitrary starting point, the temperature in Celsius and Fahrenheit etc.)
  - "Ratio" Zero is not chosen but if the scale factor. (Distances, height, weight, volume etc.) have sense proporci'on
  - "Absolute" Both the zero and the scale factor is determined. (Any form of percentage, frequency etc.)

Type of data

simb' attributes *olicos, categ'oricos or nominal*

A discrete domain have no num'ericos values. Do not allow aritm'eticas operations. In principle equality can only support a jer'aquica structure. with different granularity levels

Example entities simb'olicas



Type of data

simb' attributes *olicos, categ'oricos or nominal*

Some domains simb'olicos est'an ordered attributes (acad'emicos courses pron'ostico a disease etc.). They are called
ordinal attributes , and support comparaci'on operators. Binary attributes (presence / absence) are a form of attribute ordinal

M'as clear example of ordinal attribute is the date. Supporting Granularity

Example

DATE - → MONTH - → TRIMESTER - → TO ~

DO NOT

Data Quality

Accuracy (correcci' on and precisi' on)

Accuracy *Similarity between the data value and the true value of the attribute* .

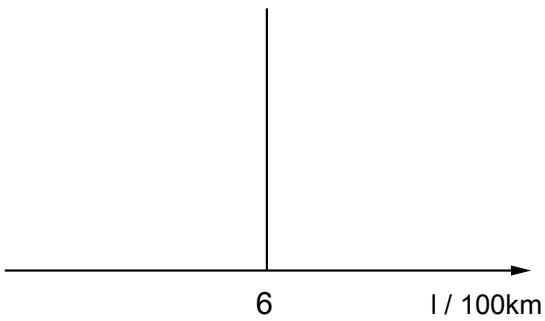
- In the case of num'ericos attributes:
  - They can exisitir rounding errors and must unify precisi'on.
  - intervalares or diffuse: Inaccurate valuations may exist. Habr'a to treat them with proper tools: similarities, fuzzy clustering, fuzzy rules asociaci'on etc.

- In the case of attributes simb'olicos .errors are detected in the data
  - There sint'actica detecci'on
  - There semantics detecci'on

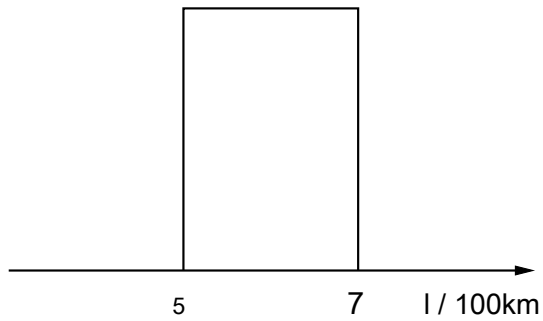
Data Quality

Accuracy (correcci' on and precisi' on)

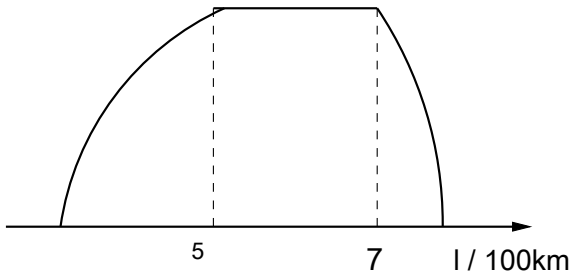
Example inaccurate values



consumption 6



Consumption 5 to 7



moderate consumption

Data Quality

Completeness (Full details)

Completeness Security there is suf fi cient data and not missing any value of an attribute

- Missing Values in the attributes:
  . In the case of num'ericos there t'ecnicas estad'ısticas attributes we will see at the end

  . In the case of simb'olicos attributes may have prior knowledge that can be used. (Functional dependencies etc.)
- Missing items:

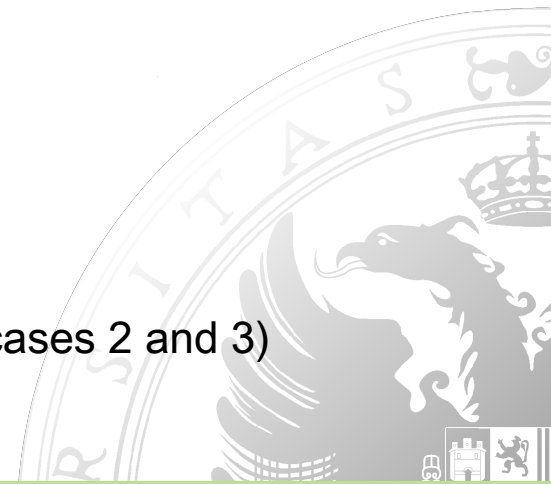
  one. Lost records or tuples

  two. Informaci'on biased

  3. scattered data

  . In this case it is dif fi cult to ensure the quality of the results.

  . A lot of data does not ensure quality domain t'erminos items (cases 2 and 3)



Data Quality

Other problems with data quality

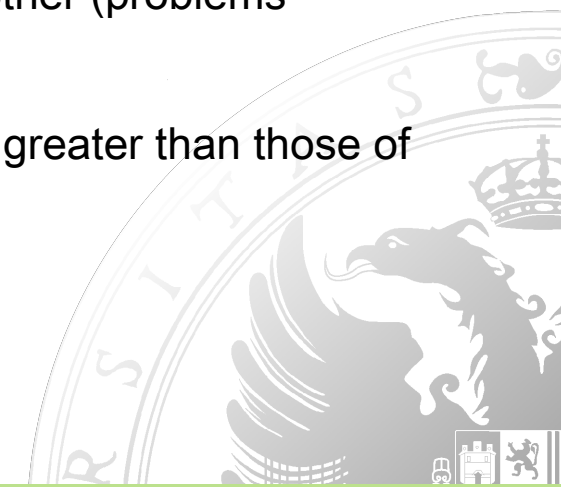
Anomalies (Outliers) These are items that really do not belong
the group to be studied and distorting regularity is sought

time lag (timeliness) Refers to the fact that the data or
of them do not have the same "now" than others.

unbalanced data It can occur at two levels:

- There are many items of a kind and very few other (problems classi fi caci'on / predicci'on)
- The values of an attribute num'erico are much greater than those of others which has much weight M'as

Duplicates data They appear when databases are merged.
Habr'a to clean the data previously



Exploraci' Data on

Why and for what

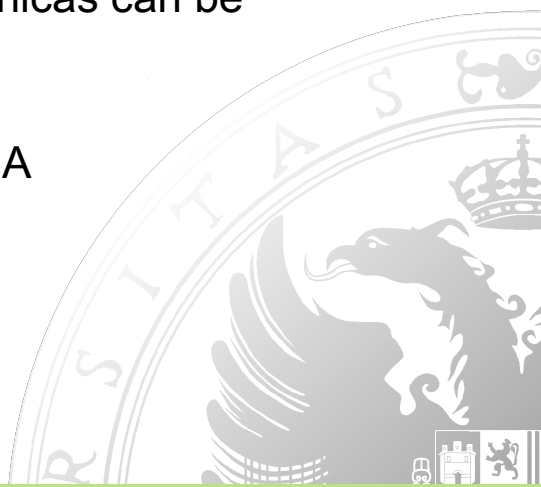
Motivations to explore data

- Helps you choose the best tools to preprocess and analyze
- Lets make initial hip'otesis on patterns extracted as the ability of human being is exploited to recognize patterns

Exploratory Data an'alisis (EDA) 1977

- It is because Tuckey
- The EDA est'a focused on visualizaci'on. Assumes adequate t'ecnicas can be extracted direct knowledge.
- To Tuckey the Clustering and detecci'on of anomal'ıas part of EDA

- More information
  http://www.itl.nist.gov/div898/handbook/eda/eda.htm



Exploraci' Data on

Exploraci' based on descriptive Estad'ıstica

Frequency Distribuci'on They can be considered:

- absolute frequencies : $F(d_i)$ n' umber of times featured a determined value d_i

  Computation of the absolute frequencies

  - discrete attributes: Categ'oricos and whole and fi nite num'ericos (with not many domain data). Simple counting on domain values D

  - continuous attributes. the dicretizaci'on is imposed, the domain becomes discrete intervals through. Usually. Yes $D = [A, B]$ and we want m intervals are chosen equal amplitude

$$[to_{one}, to_{two}], \ldots [to_{m-one}, to_m], to_1 = A, to_i = to_{i-1} + (B - A.M \; \forall i = two, .. m$$

The problem of discretization can be complex for visualizaci'on and asociaci'on issues. In some cases it is a matter of preprocessing.

Exploraci' Data on

Exploraci' based on descriptive Estad'ıstica

Frequency Distribuci'on They can be considered:

- relative frequencies : $f(d_i)$ raz'on between absolute frequency and n' UMBER total of items $f(d_i) = F(d_i) / M$, M It is the n' UMBER total items. Tambi'en may be given in percentages ($f(x_i) * 100$)
- cumulative frequencies I s'olo are defined for sorted data. Suppose the set of values $D = \{d_{one}, .. d_m\}$ Y $d_{one} \leq d_{two} \leq ... \leq d_n$, it is defined:

$$F(d_i) = \sum_{j=one}^{j=i} f(d_j)$$

When percentages are used $F(d_i)$ It indicates the percentage of the poblaci'on $\leq d_i$ and leads to the concept of:

Exploraci' Data on

Exploraci' based on descriptive Estad'ıstica

Frequency Distribuci'on

- percentile Be *s* a value between 0 and 100. We define:

$$p_s = max \{d \in D / f(d) \leq s\}$$

It is the highest value of the domain having below the *s* percent of the poblaci'on.

When $s = 0$ 25, fifty, 75, 100 They are called

quartiles

Centralizaci'on measures

- For data num'ericos

  Half : $\bar{d} = \frac{\sum_{d \in D} d\, M}{}$

- For all types of data:

  fashion : The common value M'as

  Median : The 50th percentile.

Exploraci' Data on

Exploraci' based on descriptive Estad'ıstica

Measures of dispersion

- For data num'ericos

  variance $s^2 = \frac{\sum_{d \in D} (d - \bar{d})^{two}}{M - one}$, s is the Typical deviation

  Average absolute desviaci'on $AAD = \frac{\sum_{d \in D} |d - \bar{d}|}{M}$

  Median absolute desviaci'on

  $MAD = median \{|d - \bar{d}|; d \in D\}$

- For all data:

  interquartile range $r = p_{75} - p_{25}$

Exploraci' Data on

Exploraci' based on descriptive Estad'ıstica

Exploraci'on of relaci'on between num'ericos attributes

Covariance matrix Be the variables V_j, V_k

$$cov(V_j, V_k) = \frac{\sum_{i=1}^{M} (d_{ij} - d_j)(d_{ik} - d_k)}{M - 1}$$

Correlation of matrix Be the variables V_j, V_k

$$corr(V_j, V_k) = \frac{cov(V_j, V_k)}{s_j s_k}$$

- When $corr(V_j, V_k) \approx 1$ or $corr(V_j, V_k) \approx -1$ a linear relaci'on between attributes and one of them can be expressed in funci'on other. This can serve to reduce variables.

Exploraci' Data on

Visualizaci' on

b'asicas ideas You can be displayed:

Objects an object is represented in a co fi gr'a, as a value one, two or three attributes. , Point clouds appear different colors etc.

attributes Seg' one are:

- categ'oricos attributes : Bar charts, pie charts. Using colors in the representations of objects etc.

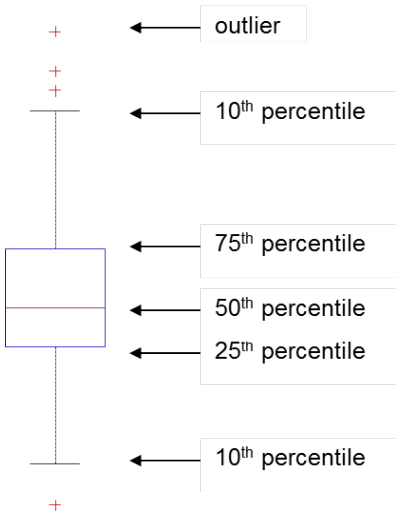
- numeric attributes and ordinal : Histogram of absolute frequencies, relative cumulative, boxplot (box plot)

Relations joint representations of attributes cloud points (scatter plot) dimensional histograms. Joint representation of Cajes diagrams etc. REPRESENTATION Correlation of matrix or covariance function categ'oricos attributes etc.

Exploraci' Data on

The boxplot (Tukey)

It's another way to view data distribuci'on



For num'ericos and continuous attributes tambi'en is usually done using, medium rather than medium and $\pm$ one *s,* two *s,* 3 *s* .. instead of percentiles. Outliers are considered from $\pm$ 3 *s*

Exploraci' Data on

Example

IRIS dataset

You can be obtained from:

http://www.ics.uci.edu/ mLearn / MLRepository.html

- Four numeric attributes: ys'epalos p'etalos length, width ys'epalos p'etalos

- An attribute categ'orico: *setosa, virginica and versicolour*
- 150 items, 50 of each type

Exploraci' Data on

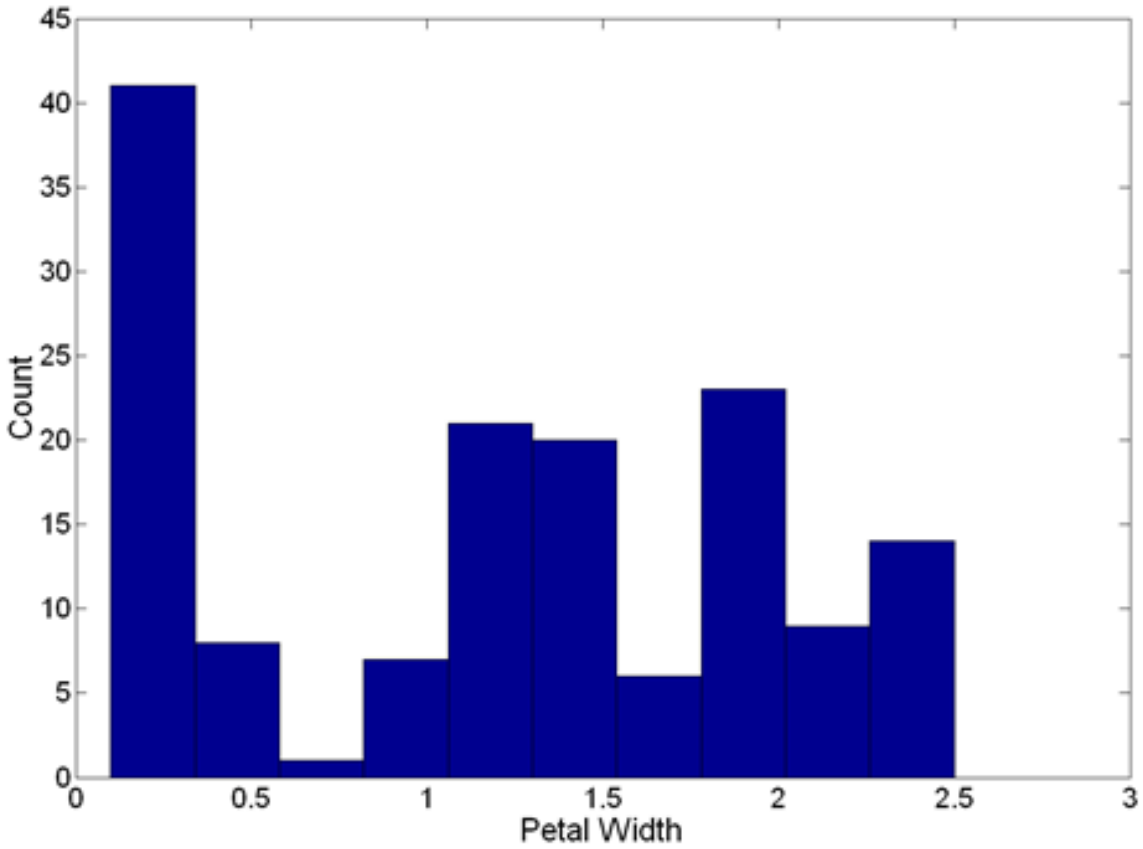
Example: statistical measures

Numeric columns | Nominal columns

Row ID	D sepal l...	D sepal ...	D petal l...	D peta
Minimum	4.3	2	1	0.1
Maximum	7.9	4.4	6.9	2.5
Mean	5.843	3.057	3.758	1.199
Std. deviation	0.828	0.436	1.765	0.762
Variance	0.686	0.19	3.116	0.581
Overall sum	876.5	458.6	563.7	179.9
No. missings	0	0	0	0

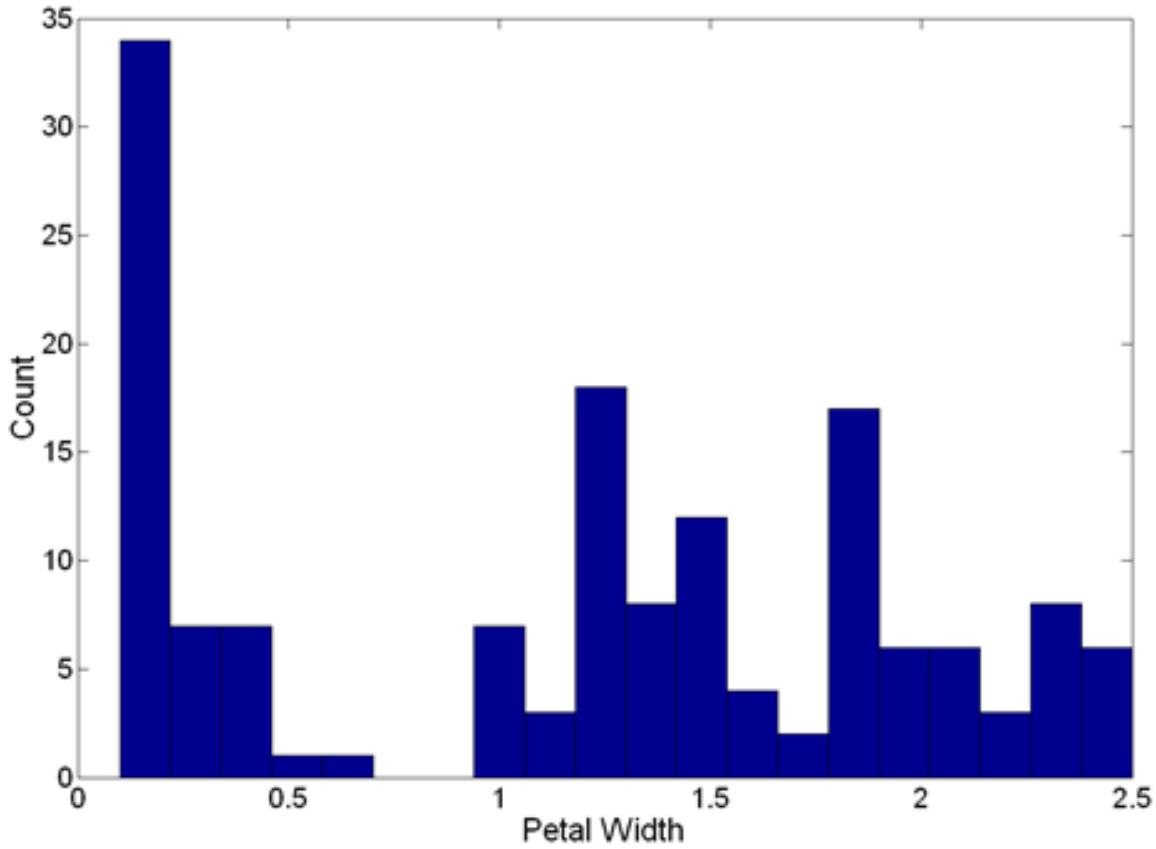
Exploraci' Data on

Example: histogram with 10 intervals



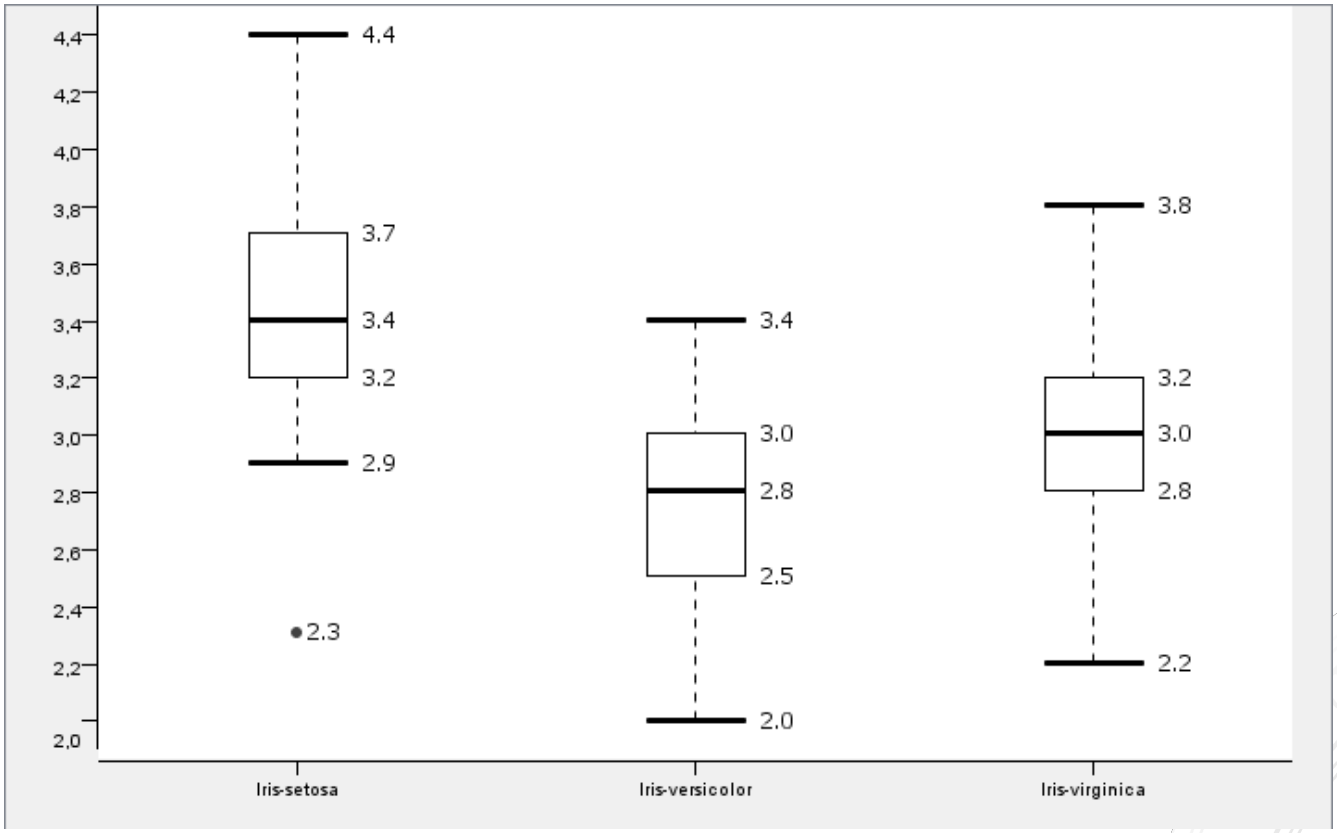
Exploraci' Data on

Example: histograms with 20 intervals



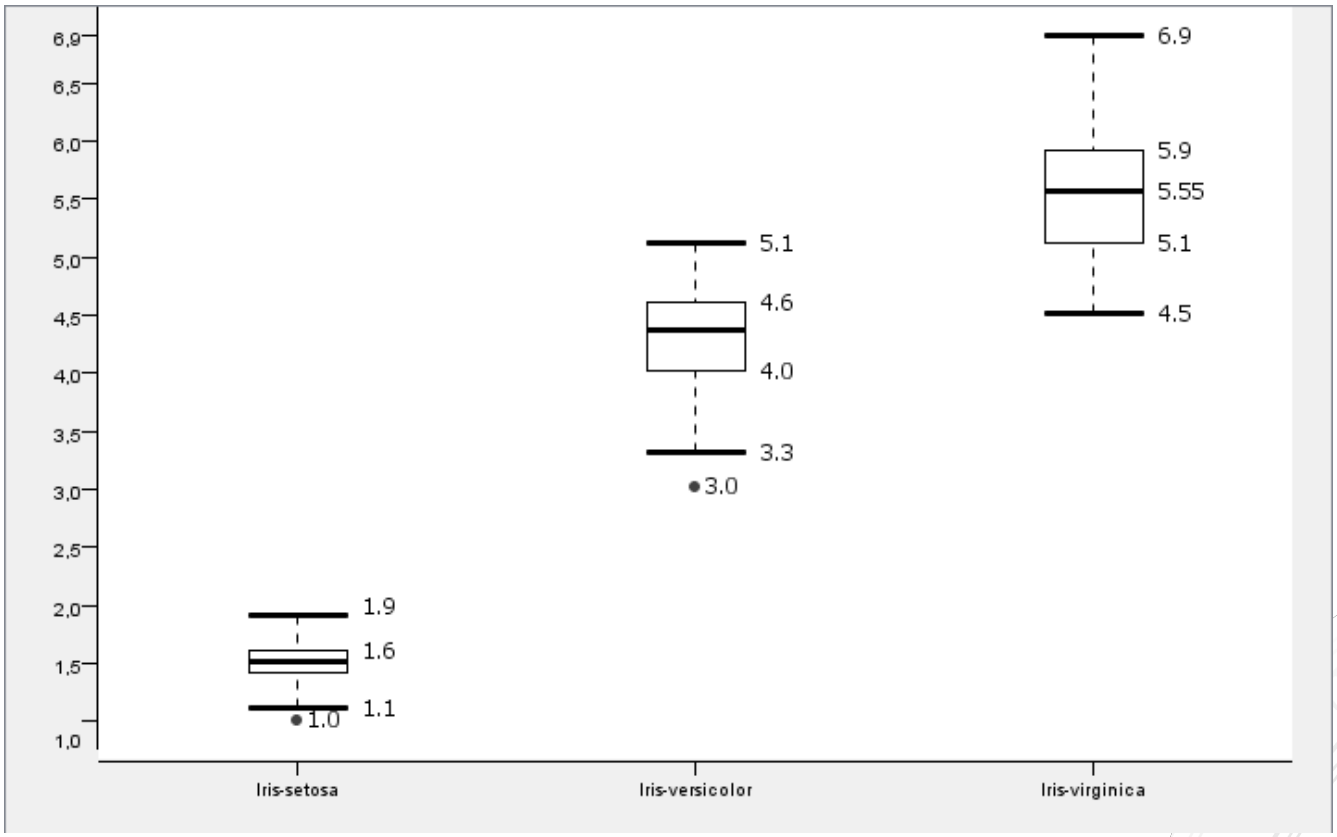
Exploraci' Data on

Example: boxplot (petal length)



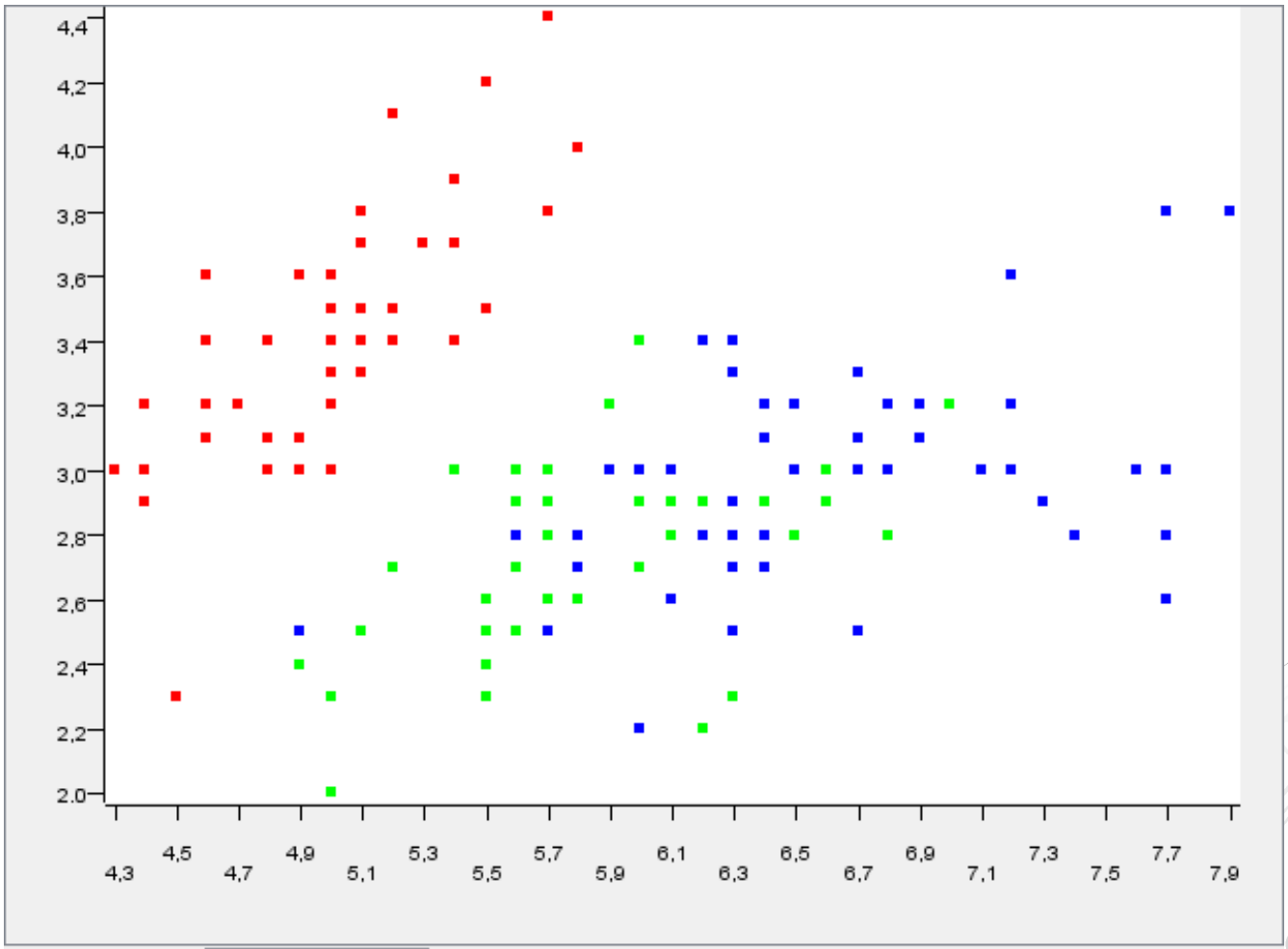
Exploraci' Data on

Example: boxplot (petal width



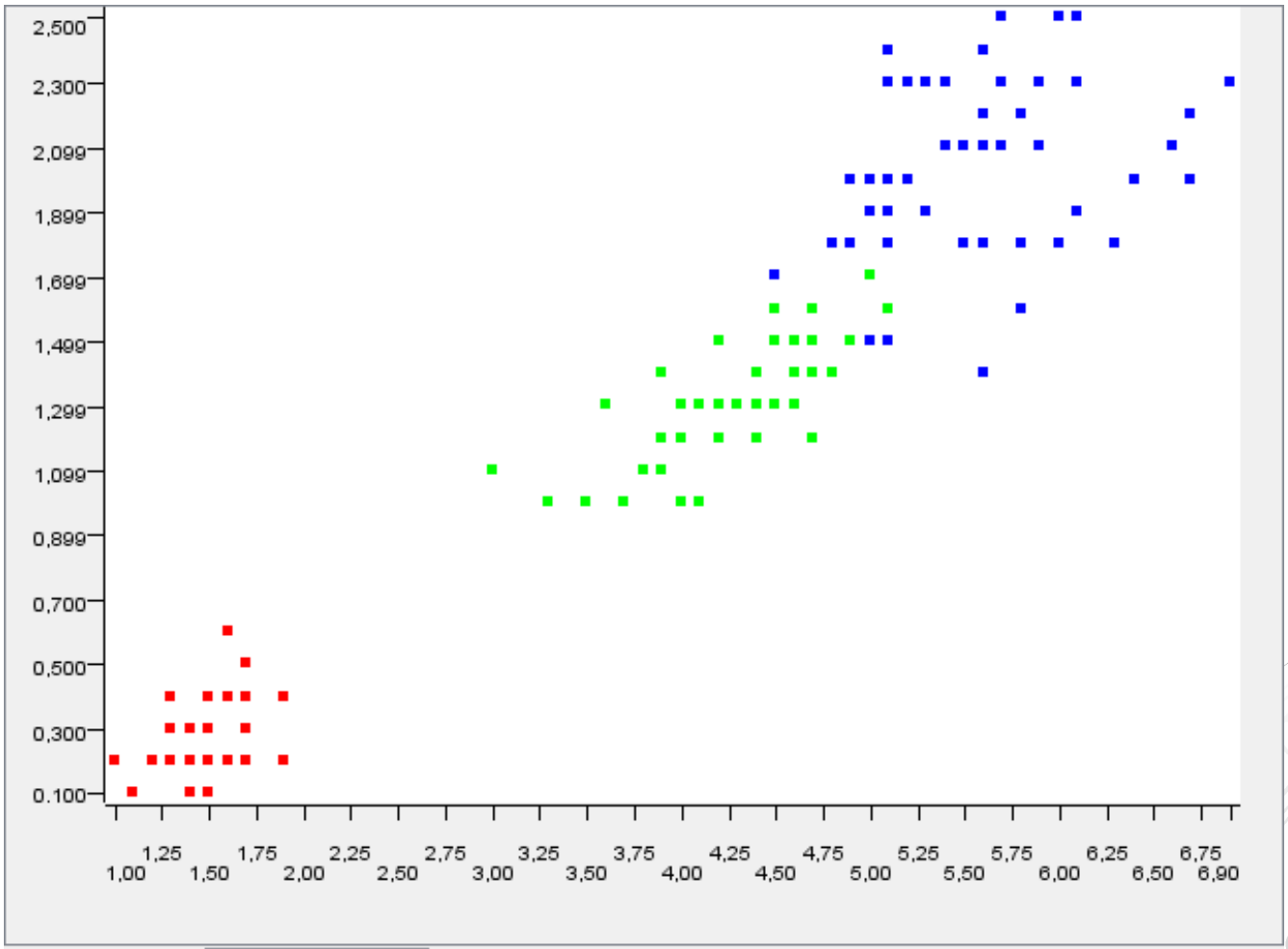
Exploraci' Data on

Example: gr'a fi c points (sepal length / width sepal)



Exploraci' Data on

Example: gr'a fi c points (petal length / width petal)



Data transformations

b'asicas ideas

Before applying t'ecnicas DM in mayor'ıa cases it is necessary to preprocess (transform) data

By qu'e transform data?

- Data needs to be changed because it can be treated directly. (Change of date of birth to age)
- The data are too detailed: add, sum, discretizar, transforming
- There are too many variables: T'ecnicas of reducci'on'o factors selecci'on
- There is much difference between the ranges. T'ecnicas scaling
- There is much lost data T'ecnicas lost data processing

Data transformations

Agregaci' on, short, discretizaci' on

Aggregation

Adding data is to combine various objects to get a new one. In general you need to add when the informaci'on is too detailed. We have:

- Horizontal AGGREGATION (abstract) The data are very detailed level object and must summarize the attributes. Cl'asicas are situations OLAP village-level data are aggregated into zones, etc.

- Vertical AGGREGATION An attribute data are very detailed and you need to add to a higher level. The classic example is the time. Other AGGREGATION the semantics of t'erminos seg' a A ontology

Data transformations

Agregaci' on, short, discretizaci' on

discretization

Discretized data is to substitute a continuous num'erico categ'orico one attribute. It is necessary in extracci'on of asociaci'on rules and classifying certain processes caci'on. Process:

one. K intervals are chosen

two. Each value is associated with the midpoint of the range where est'a and said midpoint located renames. There are several approaches:

- Intervals equally distributed It's the standard
- Equal frequency intervals . Try to have equal n' umber of where each interval.

Data transformations

Agregaci' on, short, discretizaci' on

discretization

- Intervals obtained by grouping . one partitional clustering (K-means) considering only the attribute to discretize applies. Centroid give us the values to be replaced.

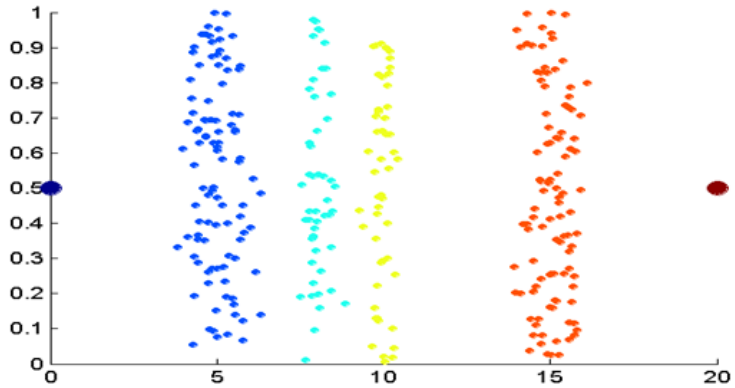
There m'etodos discretization LING based u'ıstica

Fuzzy clustering.

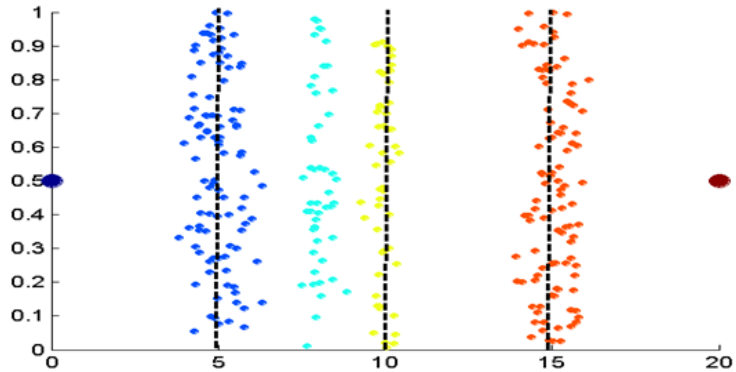
Data transformations

Example discretizaci' on

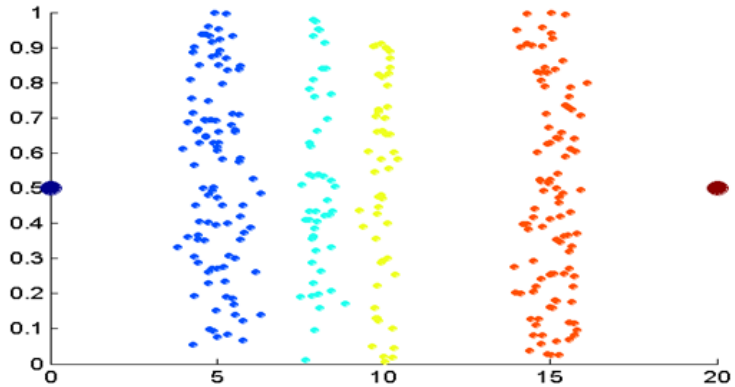
Data



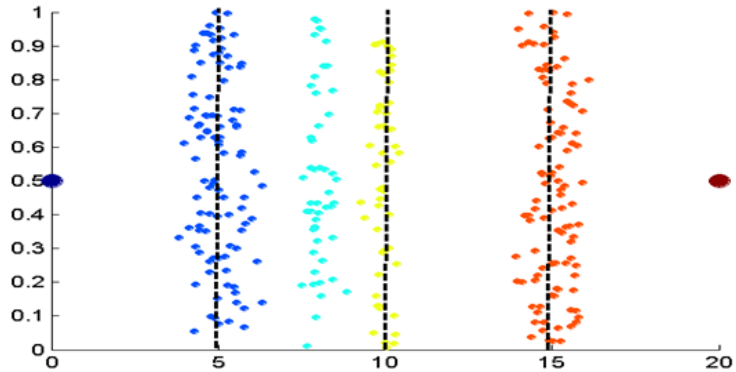
Equal intervals



equal frequency



Grouping



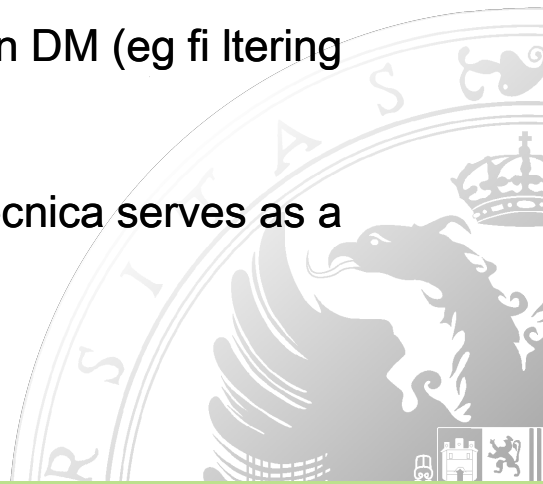
Reducci' on variables

Selecci' on variables

It is dimensi'on reduce the problem by taking a smaller set of variables

T'ecnicas

- Brute force Test Test / Error
- Selecci'on included The DM will t'ecnica choosing M'as meaningful variables (eg spanning trees of decisi'on)
- Filtered out variables are selected prior to the t'ecnicas aplicaci'on DM (eg fi ltering t'erminos in Text Mining)
- Selecci'on hedging The goodness of the result of applying DM t'ecnica serves as a criterion for selecci'on.



Reducci' on variables

Principal component analysis

Problem to be addressed

• *Illustrative example*

Consider a set of students for which the qualifications obtained in five subjects are given, the first two ex'amenes have been made without notes and the other three with notes. You want to instruct students in funci'on their performance.

Solution :

Finding a "normalized linear combination" of scores:

$$x_c = \Sigma\ l_j\, x_j \text{ such that } \Sigma\ l_{\text{two}} \qquad j = \text{one}$$

m'axima to collect the variance because "separar'a" to sort items and f'acil ser'am'as

Reducci' on variables

Principal component analysis

Problem to be addressed

Other examples

- Ordenaci'on banking customers
- Ordenaci'on of seg' items A Website
- Overall obtenci'on of res' umenes attribute

Reducci' on variables

Principal component analysis

The model matem'atico

Consider a data set with real num'ericos data, this can be seen as a variable N dimensional –

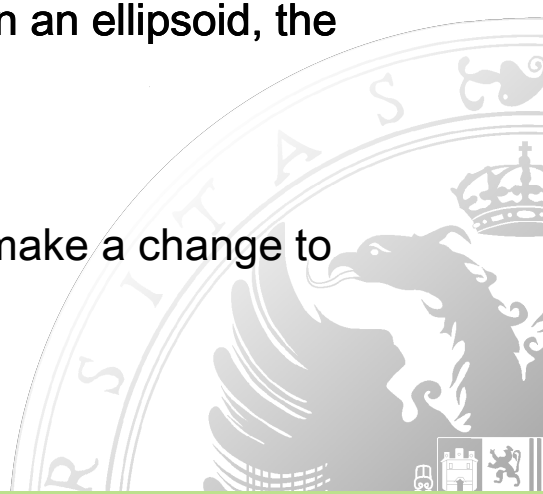
$$x = (x_{\text{one}}, .. x_N) \text{ Y } m \text{ values}$$

the same x_{ij}, $i \in \{\text{one}, , M\}$, $j \in \{\text{one}, .., N\}$, we find standard linear transformations (SLC) that "summarize" the best possible data, capturing most variance thereof.

intuitive idea

If the items are considered as a point cloud R_n, all can be enclosed in an ellipsoid, the average center, whose matrix is the covariance matrix.

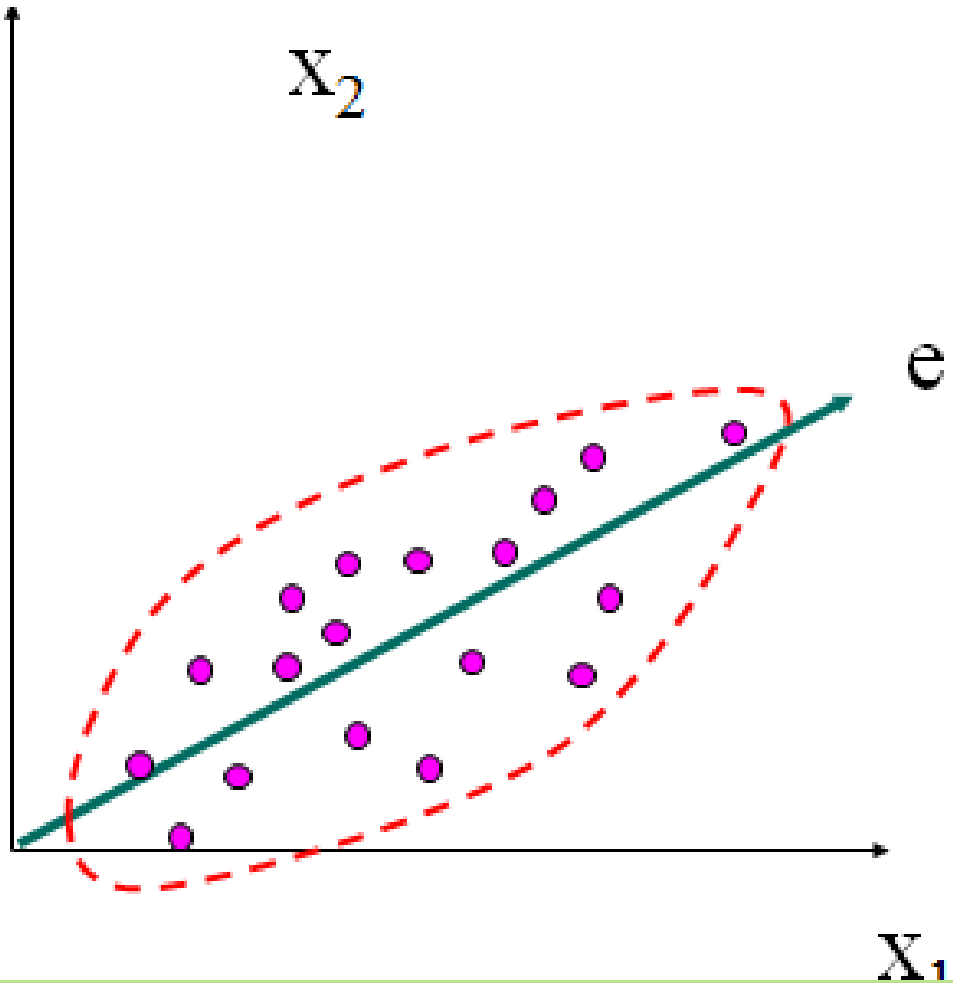
The axes of the ellipsoid are a rectangular coordinate system, if we make a change to this coordinate system, the point spread along axis



Reducci' on variables

Principal component analysis

intuitive idea



Reducci' on variables

Principal component analysis

The model matem'atico

Be μ the average $\bar{x}$ Y Σ its covariance matrix, is find a linear transformaci'on $\bar{y} = \Gamma'(\bar{x} - \bar{\mu})$ such that new coordinate axes are the axes of the ellipsoid. Proves that Γ is a matrix such that:

$$\Gamma' \Sigma \Gamma = \Lambda$$

where,

$$\Lambda = diag(\bar{\lambda}) = \begin{pmatrix} \lambda_{one} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \lambda_n \end{pmatrix}$$

Reducci' on variables

Principal component analysis

The model matem'atico

Vector values $\bar{}$ λ veri fi ed $\lambda_{\text{one}} \geq, .. \geq \lambda_N$ and are the "Eigenvalue" of the covariance matrix and associated with each λ_j there is a "eigenvector" $\bar{}$ γ_j that is the jth column matrix Γ and verify c'andose:

$$\forall j \in \{\text{one}, , N\} \text{ and } j = \bar{} \quad \gamma_j \prime (\bar{} x - \bar{} \mu)$$

Y_j It is called jth principal component.

$$\forall k, j \in \{\text{one}, , N\} \; Cov(\text{and}_j, Y_k) = 0 \; Var(\text{and}_j) = \lambda_j$$

$$Var(\text{and}_{\text{one}}) \geq \geq .. \; Var(\text{and}_n)$$

and given $k \leq N$ there is no SLC which is independent of k first main component and having a greater variance than the $k +$ one main component.

Reducci' on variables

Principal component analysis

The model matem'atico

• *Proporci'on explained variance*

? The proporci'on of variance explained by *k* factors is

($\lambda 1 + .. + \lambda k) / \lambda 1 + .. + \lambda n)$ and it allows us to reduce the dimensionality of space. That is express fen'omeno less variables.

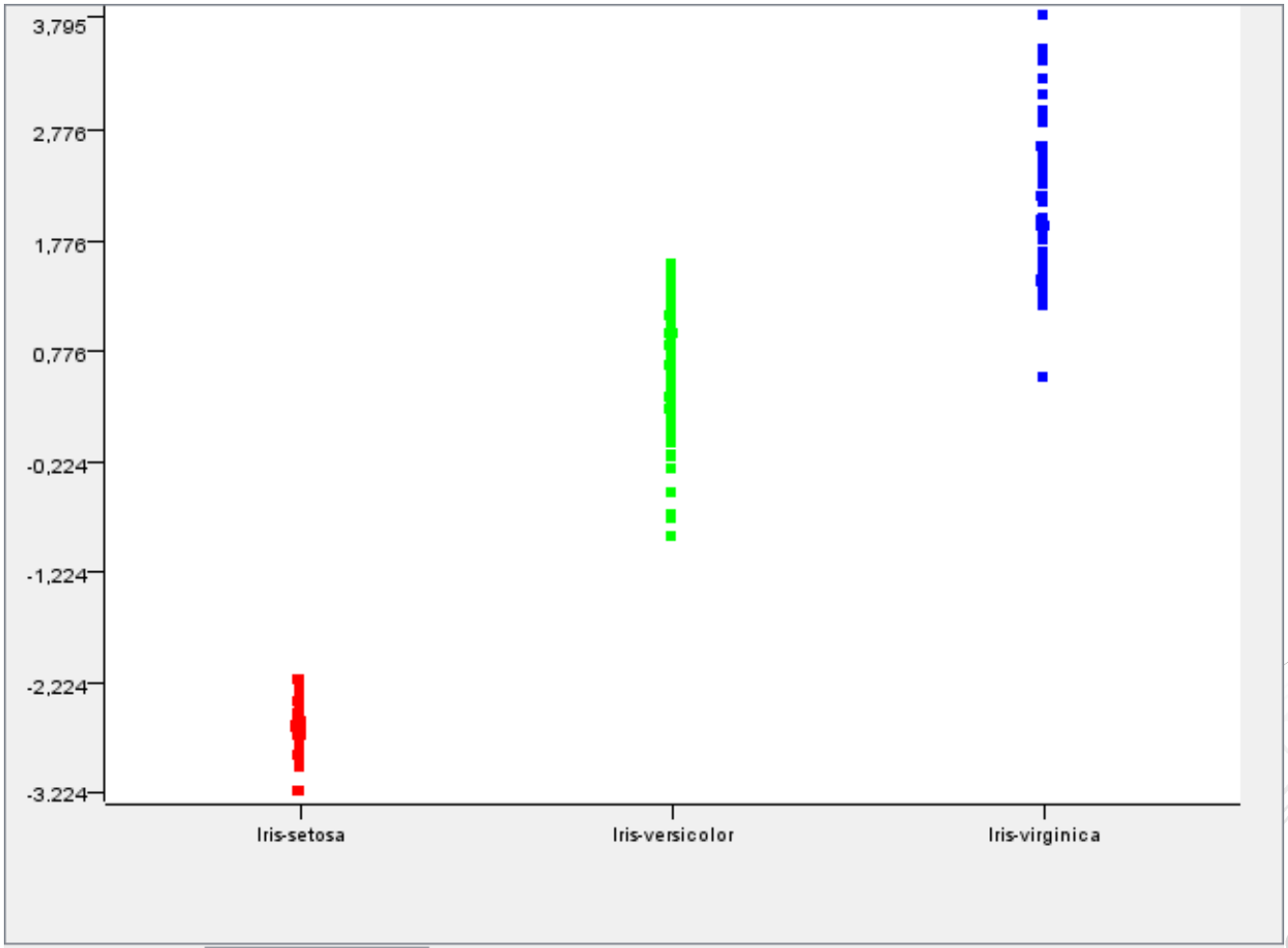
How many components make ?:

- At least 90% of variance explained
- All eigenvalues that are greater than the average of the same. If the matrix is used instead Correlation of the covariance matrix eigenvalues greater than 1.

The proporci'on of variaci'on explained variable j by k component allows identifying the components and identify one semantics for them.

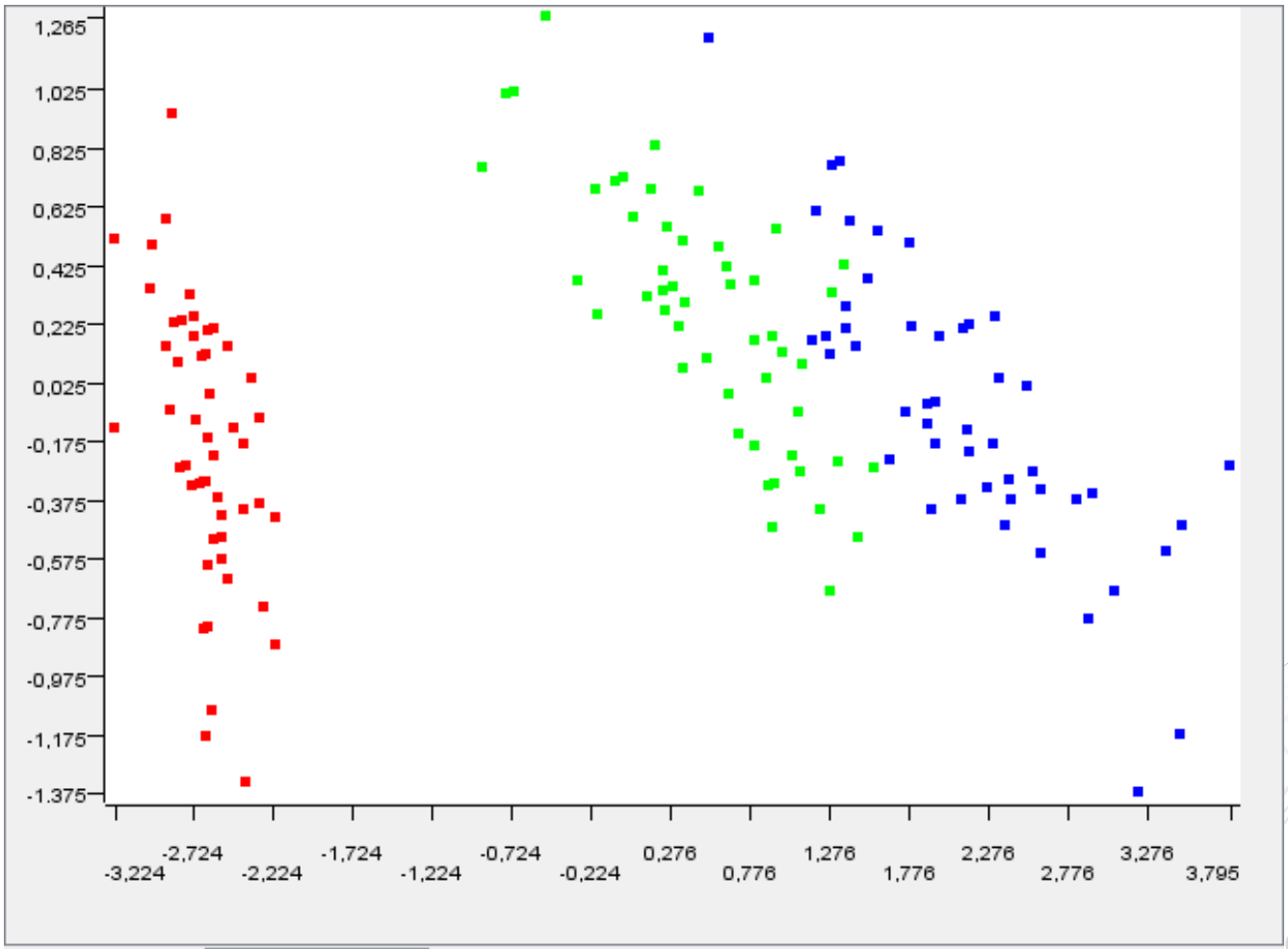
Main components

Example: Iris (classes / first factor)



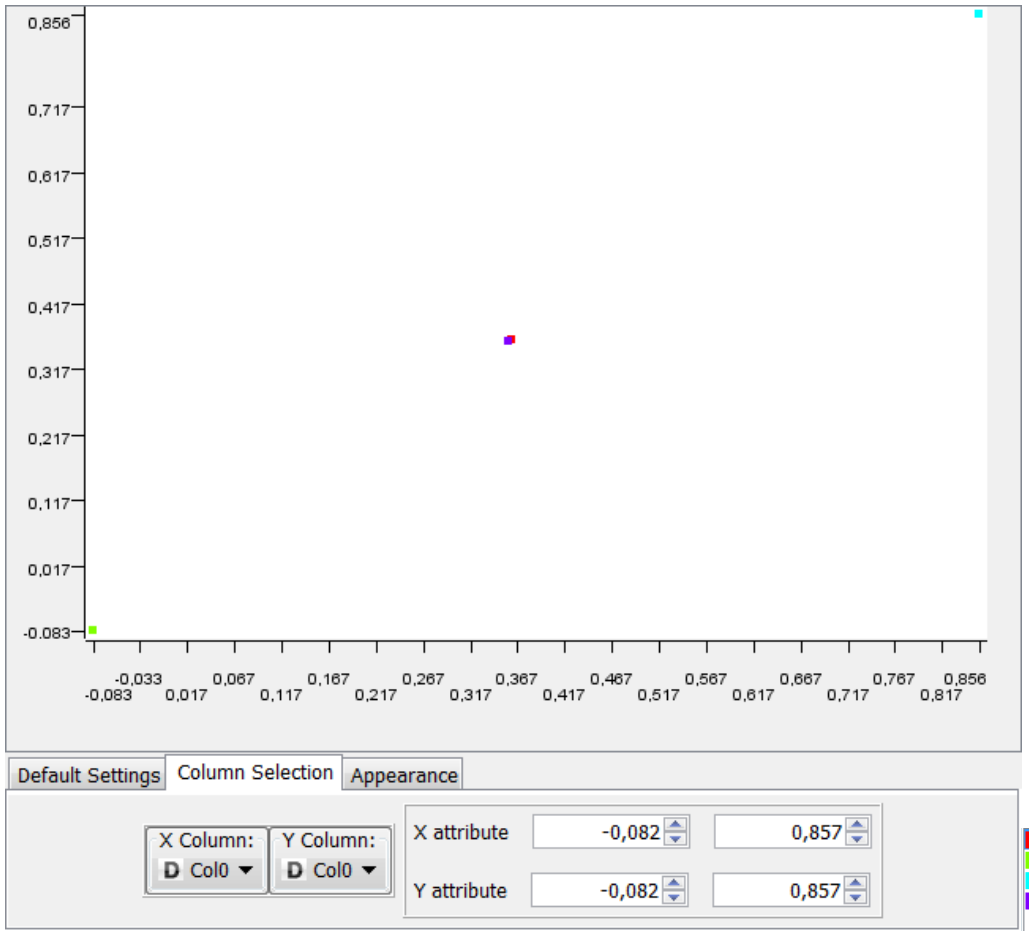
Main components

Example: Iris (first factor / second factor)



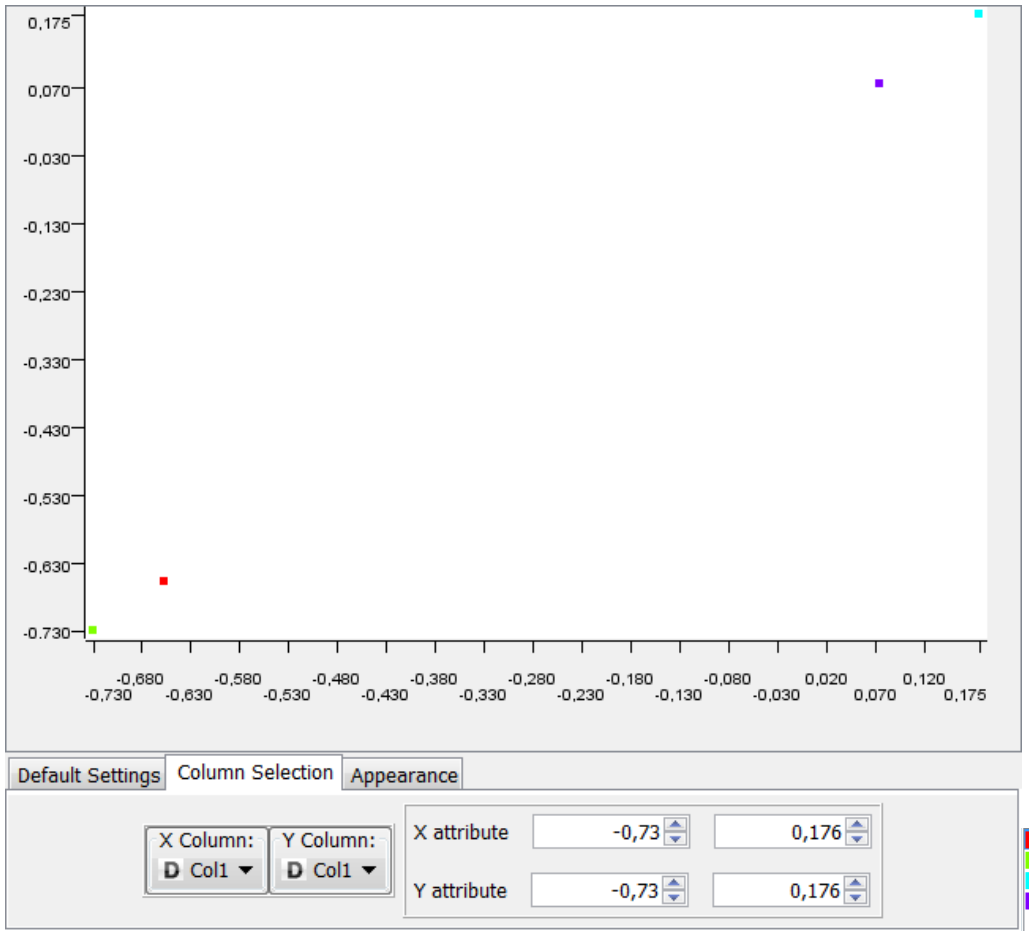
Main components

Example: Iris (variables / first factor



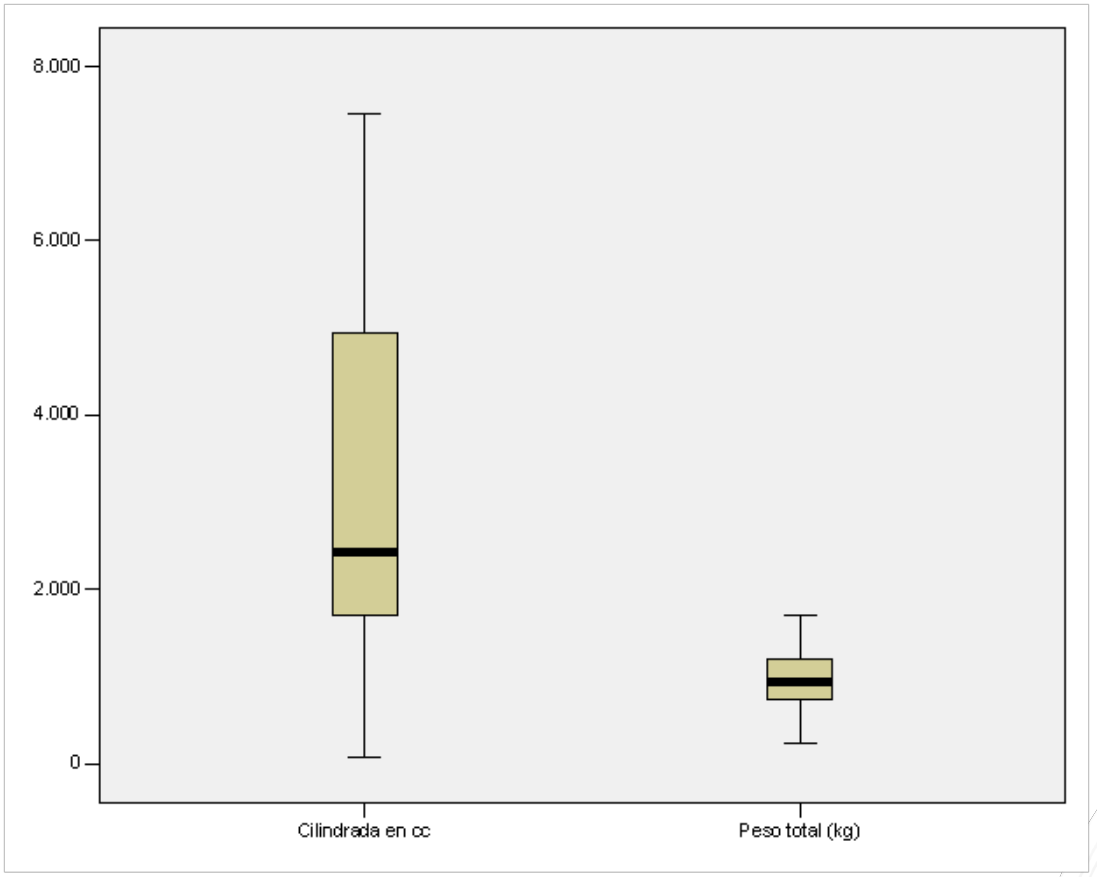
Main components

Example: Iris (variables / second factor)



Scaling problems

Motivaci' on



Scaling problems

b'asicas ideas

Num'ericos the values of two attributes are different scales. This makes can not be treated together on issues such as computation of distances etc.

- *Some expressions to normalize*
  - Normalizaci'on in [0,1]

$$V = \frac{V_{to} - min_{to}}{max_{to} - min_{to}}$$

  - Tipi fi caci'on

$$V = \frac{V_{to} - \bar{d}}{s}$$

  - Tipi Robusta fi caci'on

$$V = \frac{V_{to} - median}{\text{interquartile range}}$$

Problems of missing values

b'asicas ideas

A variable has missing values when its value is not known for a particular data

- *Origin of missing values*
  - Lack of value without random factors. (Someone does not answer something in a survey)
  - Property not applicable. (Hair Color frogs). You can not identify with 0 or NO.
  - Error random source.
    - random completely: it follows that the data distribuci'on (failure of unknown origin from a sensor) (MCAR)
    - conditional random: follows a conditional distribuci'on, failure of a sensor failure when it rains M'as (MAR)

Problems of missing values

To do with missing values

delete records Eliminating lost data

- If it is a totally random situaci'on.
- If the data volume is not seriously altered

Replace Replace the lost data worth

- a new value is generated within a quantum domain. (NS / NC, NO etc.)
- M'as follows using the common value seg' how a class registration.
- If quantitative data is replaced by:
  - The average if the error is completely random
  - The conditioned media to aparici'on error if we have a type MAR
  - Pr'oximos average values (interpolating) if we know that there is a certain temporal or spatial dependence