



**UNIVERSITÀ
DEGLI STUDI
DI FOGGIA**



HR EXCELLENCE IN RESEARCH

Pre-processing

Prof. Crescenzo Gallo

Aggregate Professor of Computer Science and Processing Systems

Department of Clinical and Experimental Medicine

crescenzo.gallo@unifg.it



Pre-processing

Preprocessing consists of a set of strategies and techniques that stand in complex relationships with each other. We will present below some of the most important ideas and approaches, trying to clarify from time to time the relationships that bind them.

We will present the following topics

- *Aggregation*
- *Sampling*
- *Dimensionality Reduction*
- *Feature Subset Selection*
- *Feature Creation (construction)*
- *Discretization and Binarization*
- *Variable Transformation*

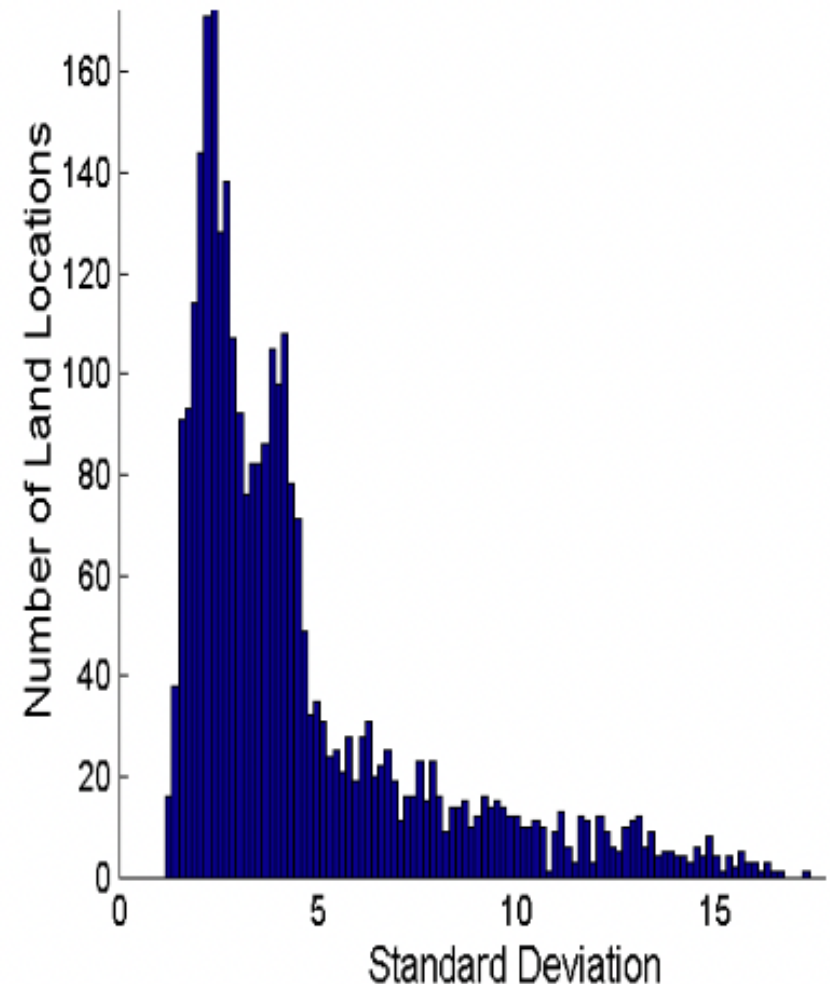
Pre-processing: aggregation

Aggregation

Combining two or more objects (or attributes) into a single object (or attribute)

Goals

- *Data reduction*: reduce the number of attributes associated with objects.
- *Change of scale*: aggregate cities into regions, states, ...
- *More stable data*: aggregate data tends to have less variability.



Pre-processing: **sampling**

Sampling

Main technique to select observations, employed both in the preliminary and in the final phases of a Data Mining analysis.

Goals

- *To reduce economic effort:* we sample to save money.
- *Reduce computational effort:* sampling to save time.

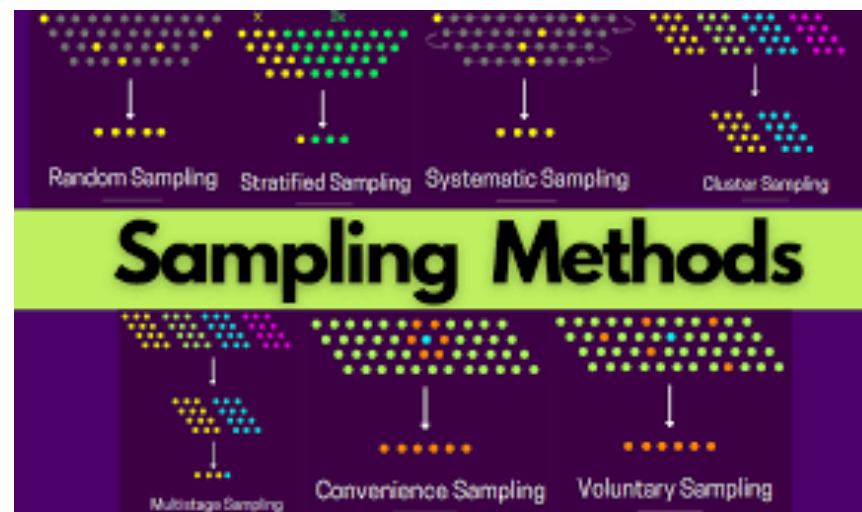
The basic principle on which sampling is based is as follows:

- *Using a sample, in lieu of the entire data set, leads to results comparable to those that would be obtained if the entire data set were used, provided that the sample drawn is representative of the entire data set.*
- *A sample is representative if it enjoys approximately the same properties (of interest) as the entire data set.*

Pre-processing: sampling

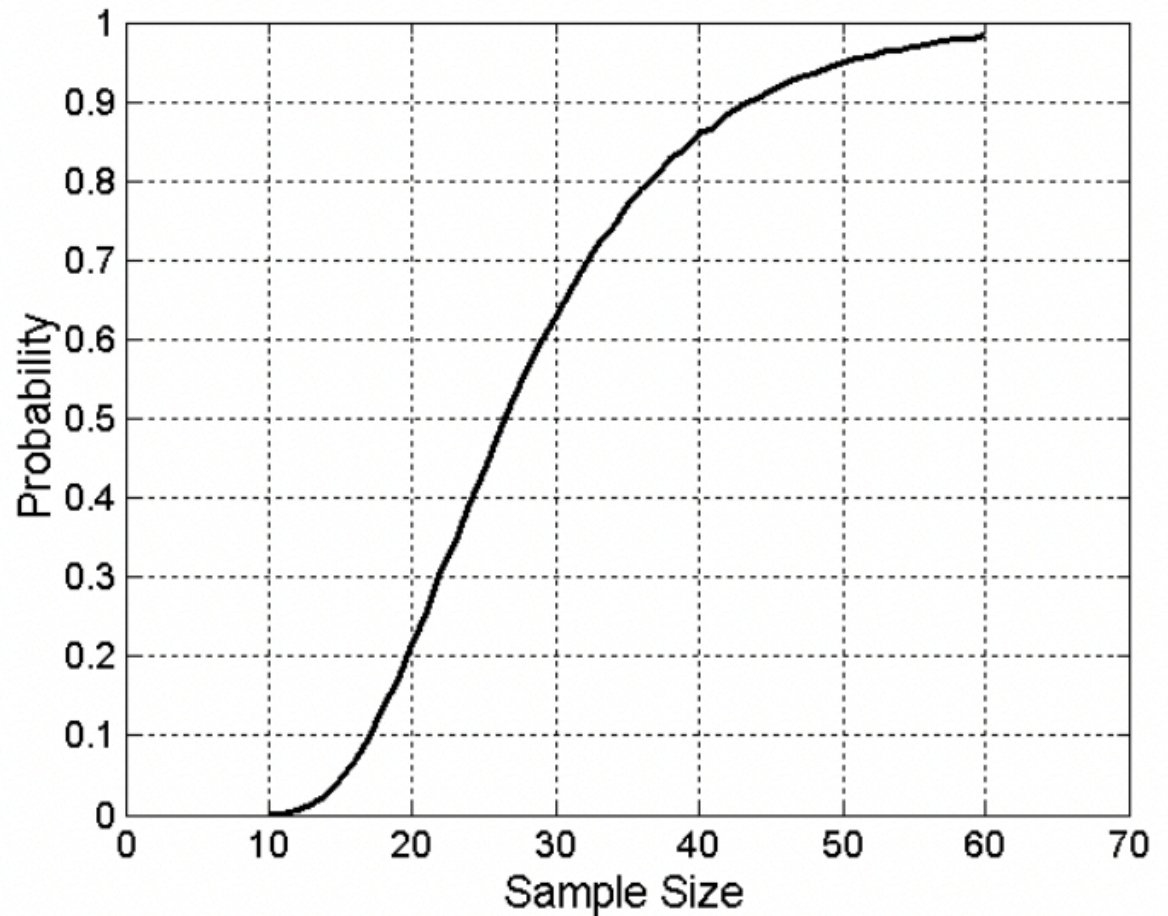
Types of Sampling

- *simple random sampling*: equal probability of selecting each observation from the entire data set.
- *sampling without repacking*: each selected observation is removed from the entire data set, it cannot be selected a second time.
- *sampling with repacking*: each observation can be selected multiple times.
- *stratified sampling*: the entire data set is partitioned into disjointed subsets; samples are drawn from each subset of the partition.



Pre-processing: **sampling**

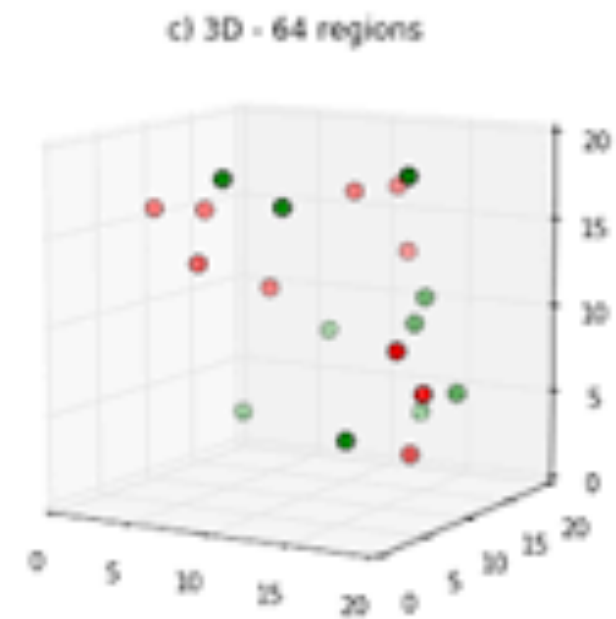
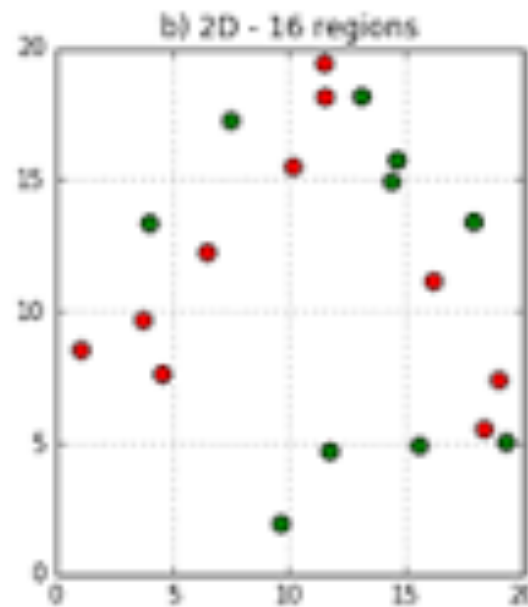
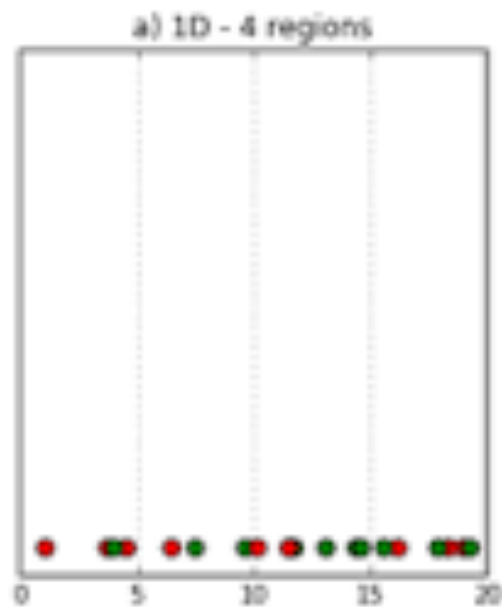
What must be the size of a sample for you to be able to get at least one item from each of 10 groups?



Pre-processing: curse of dimensionality

As dimensionality increases, data are increasingly sparse when evaluated in terms of the input space in which they are located.

The definitions of density and distance between points, which are critical for clustering and outlier identification tasks, lose their meaning.

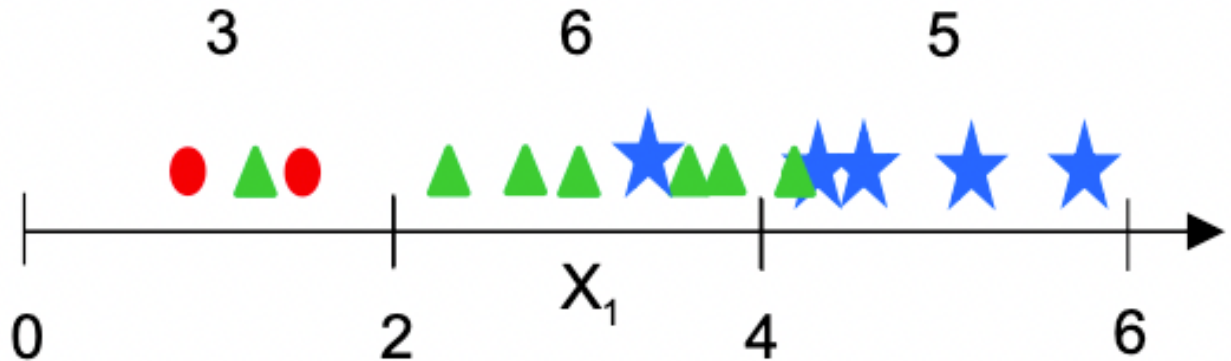


Pre-processing: curse of dimensionality

As dimensionality increases, data are increasingly sparse when evaluated in terms of the input space in which they are located.

The definitions of density and distance between points, which are critical for clustering and outlier identification tasks, lose their meaning.

Classification problem with a class that can take on three different values represented by red circle, green triangle, and blue star.

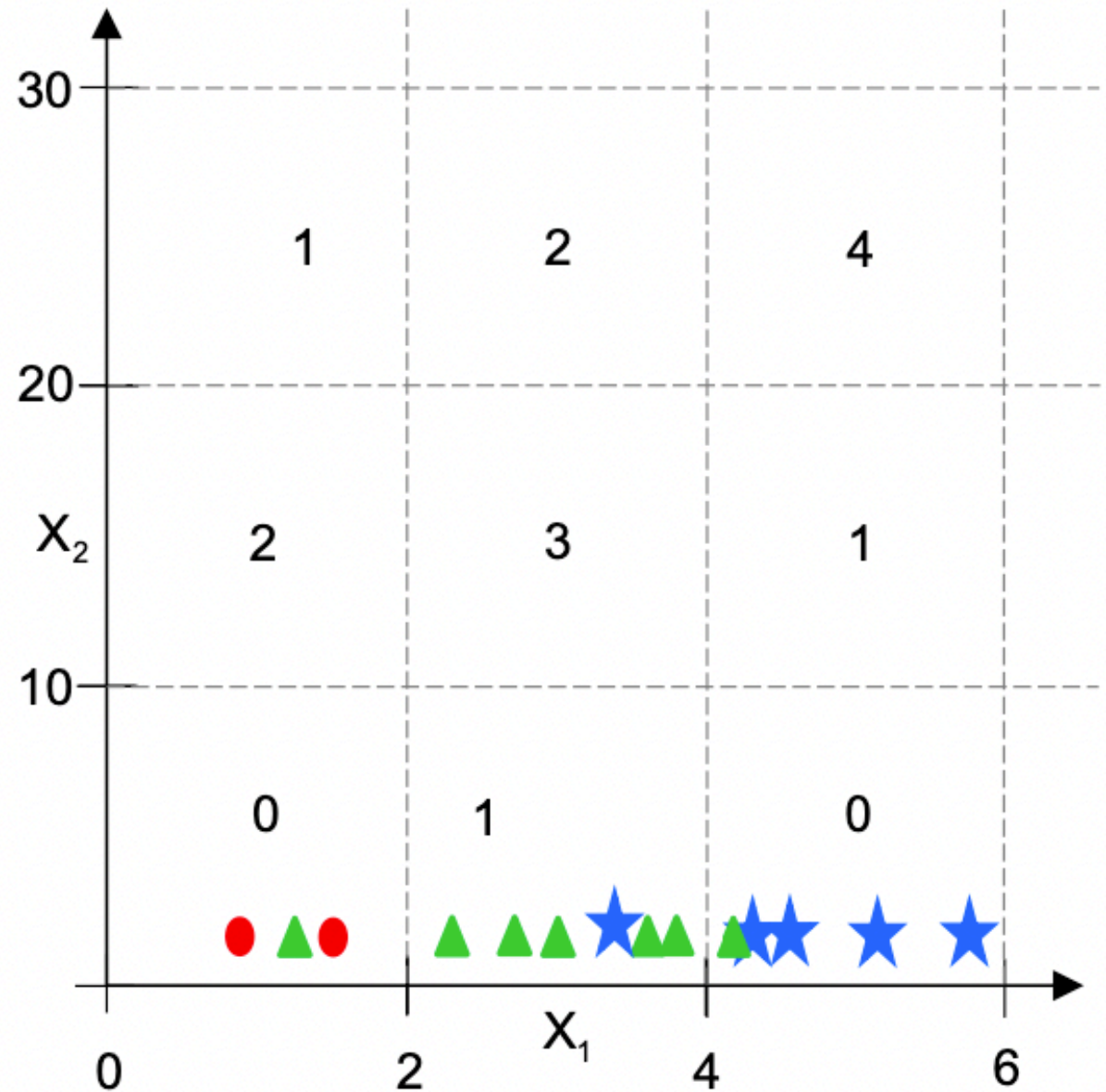


Pre-processing: curse of dimensionality

As dimensionality increases, data are increasingly sparse when evaluated in terms of the input space in which they are located.

The definitions of density and distance between points, which are critical for clustering and outlier identification tasks, lose their meaning.

Considering also the X_2 attribute one can find a decrease in the density of the observations, density in the relative cells.

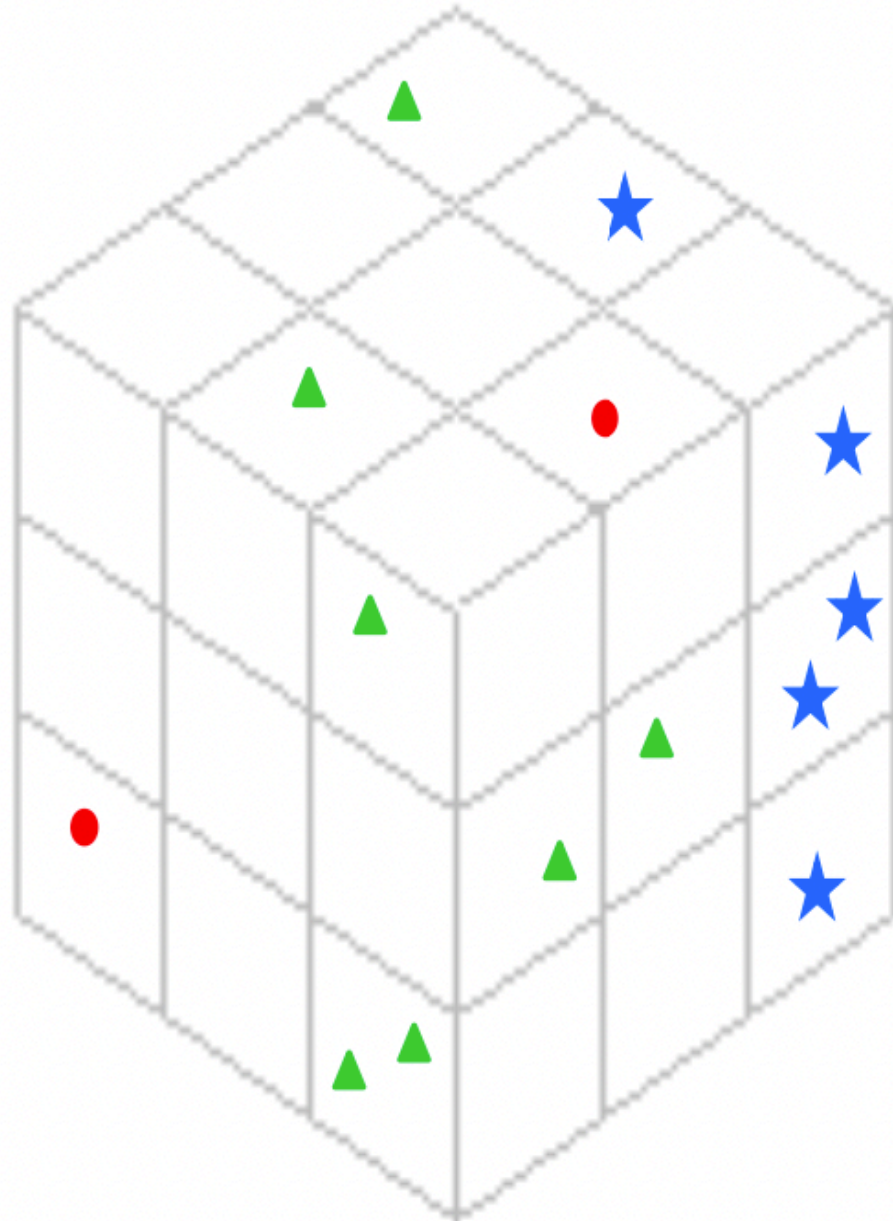


Pre-processing: curse of dimensionality

As dimensionality increases, data are increasingly sparse when evaluated in terms of the input space in which they are located.

The definitions of density and distance between points, which are critical for clustering and outlier identification tasks, lose their meaning.

Considering an additional X_3 attribute only exacerbates the problem of low cell density resulting from the input space partitioning process.



Pre-processing: curse of dimensionality

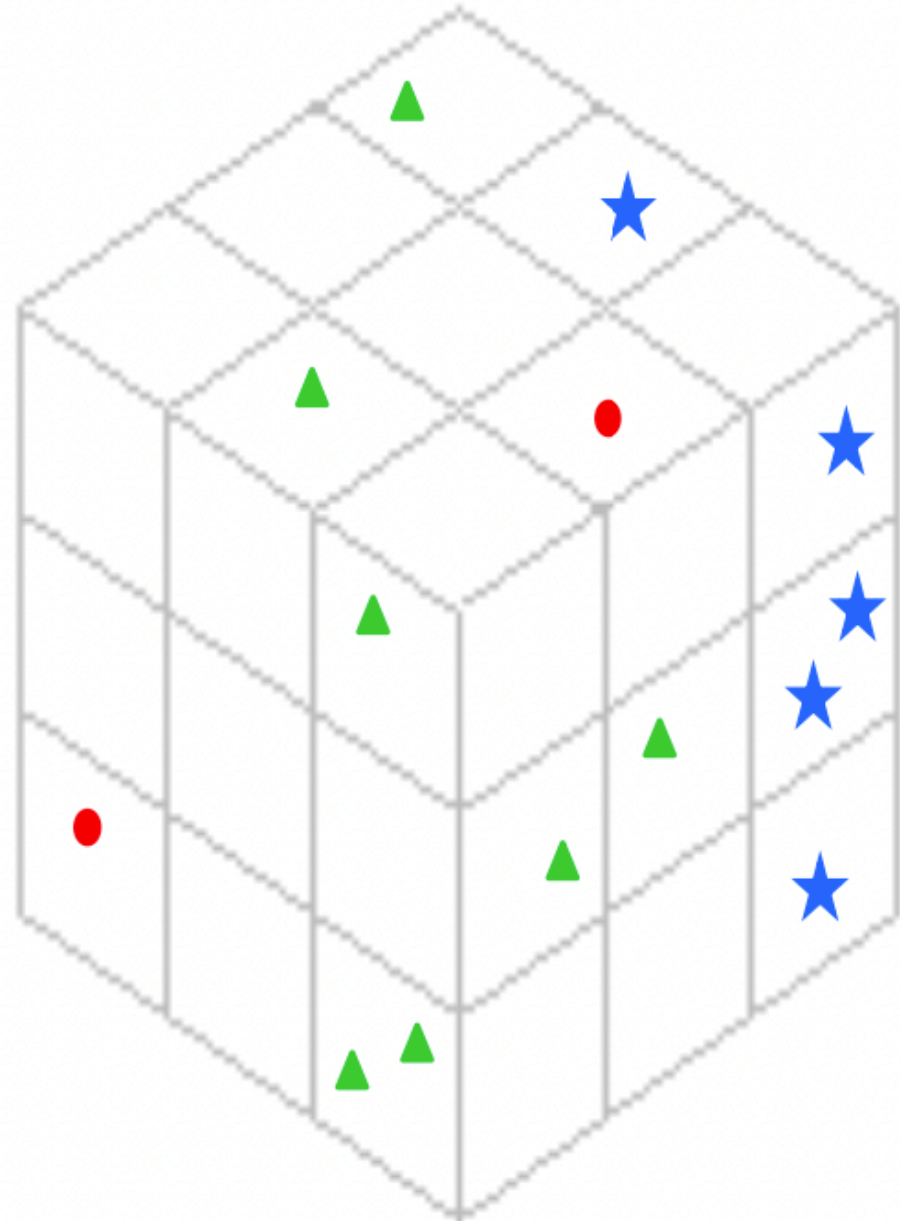
As dimensionality increases, data are increasingly sparse when evaluated in terms of the input space in which they are located.

The definitions of density and distance between points, which are critical for clustering and outlier identification tasks, lose their meaning.

1D: $3 \text{ classes} \times 3 \text{ bins} = 9 \text{ observations}$

2D: $3 \text{ classes} \times 9 \text{ bins} = 27 \text{ observations}$

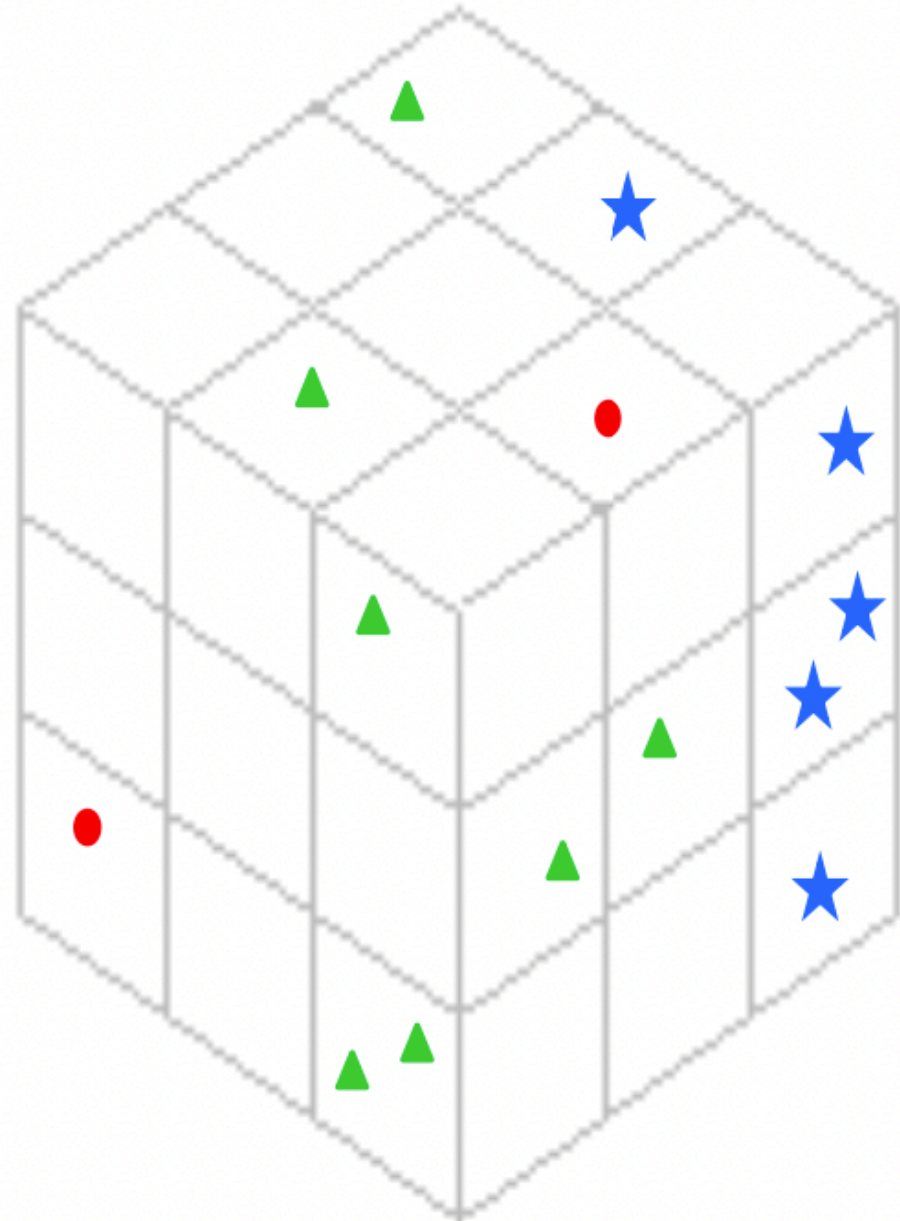
3D: $3 \text{ classes} \times 27 \text{ bins} = 81 \text{ observations}$



Pre-processing: curse of dimensionality

Alternatives

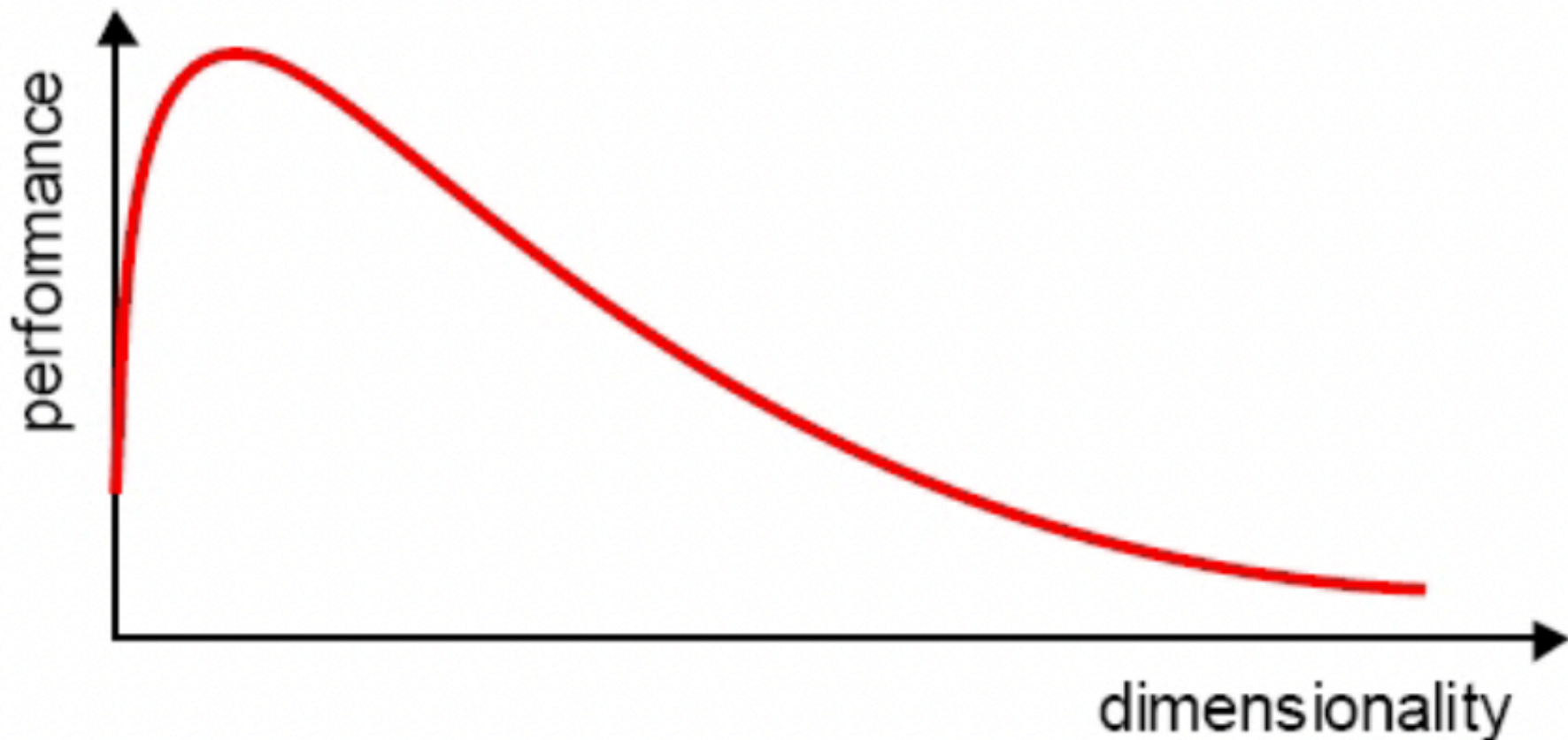
- *Use a priori information to exclude portions of the input space*
- *Reduce the number of intervals, use larger intervals*
- *Reduce the number of attributes used to describe the dataset being analyzed.*



Pre-processing: curse of dimensionality

Summarizing

The phenomenon of the curse of dimensionality means that given a sample with fixed size (size of the sample) there is a number of attributes that can be taken into account jointly beyond which the performance obtainable by a classification model inevitably begin to degrade.



Pre-processing: downsizing

Objectives

- To avoid incurring in the phenomenon of "curse of dimensionality".
- Reduce time and memory needed for Data Mining algorithms
- Enable easier data visualization
- Identify irrelevant attributes or help reduce the level of noise in the data

Techniques

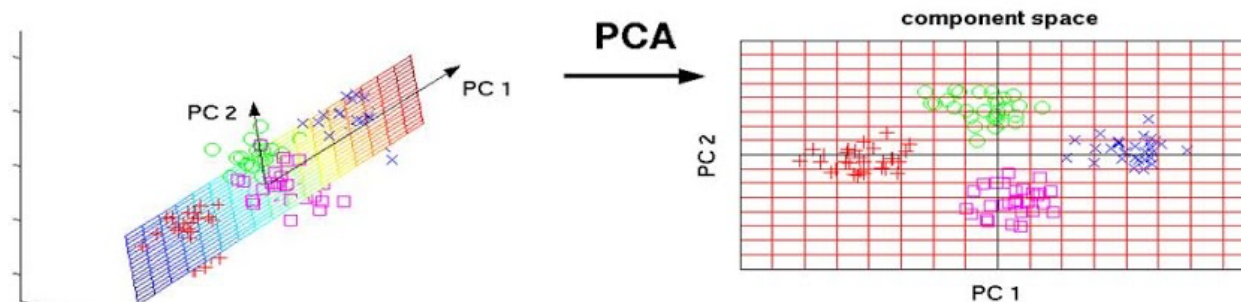
- *Principal Component Analysis (PCA)*
- *Singular Value Decomposition (SVD)*
- *Other: supervised and unsupervised techniques*

Pre-processing: downsizing

Principal Component Analysis (PCA) aims to identify patterns in data, to express data in such a way as to highlight similarities and differences.

Since patterns in data can be particularly difficult to find, in high dimensional spaces where the luxury of graphical representation is not allowed, PCA is a very powerful tool for data analysis.

Another of the primary advantages of PCA is that once patterns are identified in the data, the data is compressed, reducing the number of dimensions, without excessive loss of information.

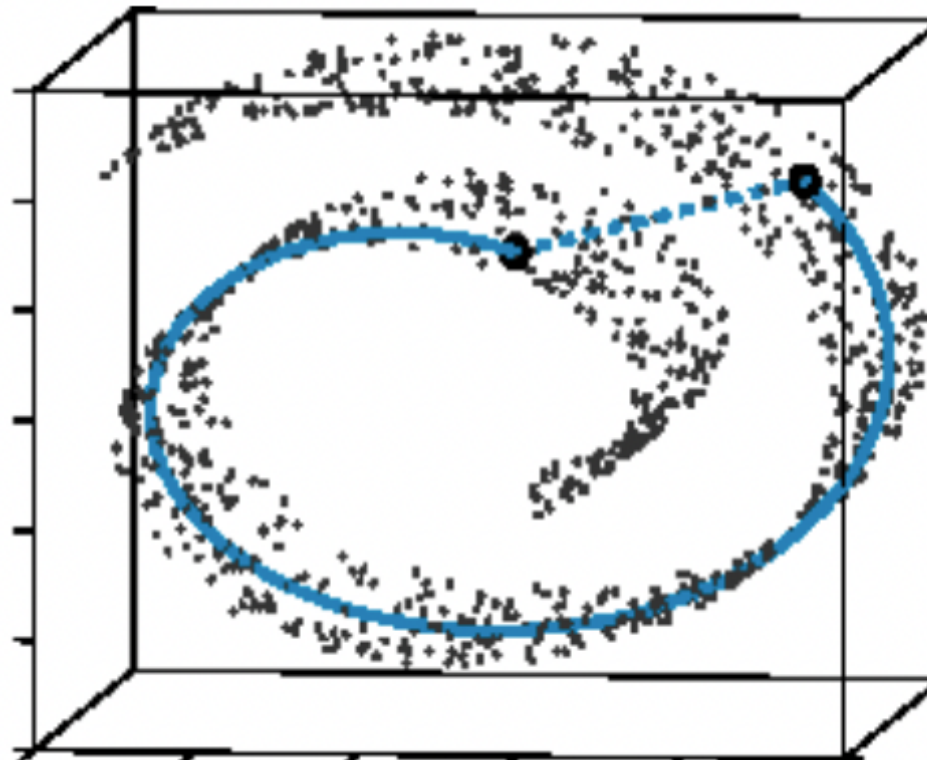


Principal Components Analysis (PCA)

Pre-processing: downsizing

Another possibility is offered by the algorithm called **ISOMAP**, which involves the following main steps:

- Constructing a proximity graph of the observations in the dataset
- Computation, for each pair of points of the proximity graph, of the relative path at minimum cost (geodesic distance)



Pre-processing: feature subset selection

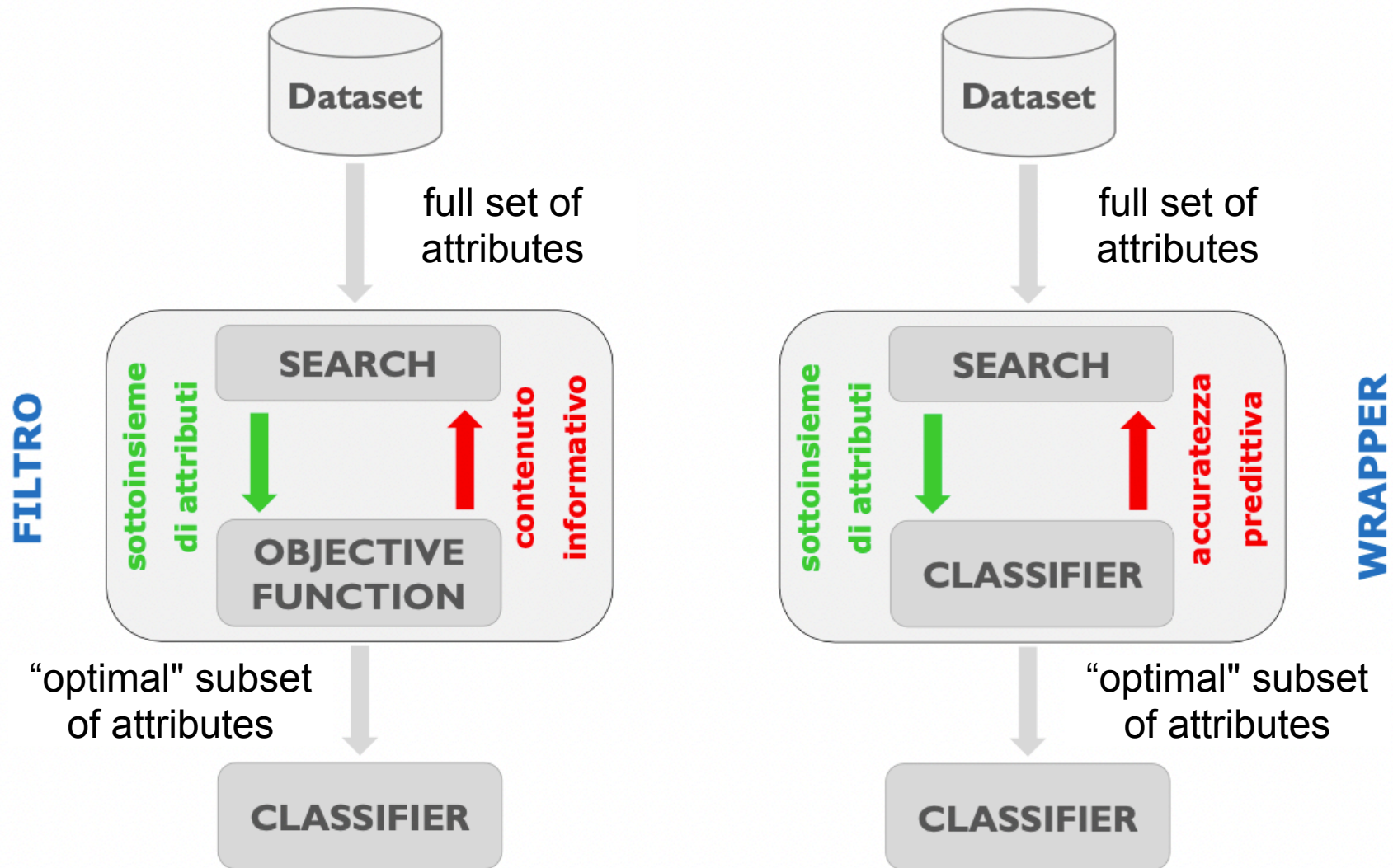
It offers another way to reduce the dimensionality of the data to be analyzed. The attributes (features) that describe the observations could be:

- **Redundant:** carry duplicate (redundant) information as it is contained in one or more of the remaining attributes, (purchase price of an asset and cost of the related tax, ...)
- **Irrelevant:** they do not contain useful information for the task of Data Mining that must be faced, (identifier of a client, ...)

Available approaches

- **Brute-force:** try all possible subsets of attributes as input to DM algorithms
- **Embedded:** the FSS is obtained as part of the execution of the DM algorithm
- **Filters:** attributes are selected before the DM algorithm is used
- **Wrapper:** use the DM algorithm as a black-box to find the best subset of attributes

Pre-processing: feature subset selection



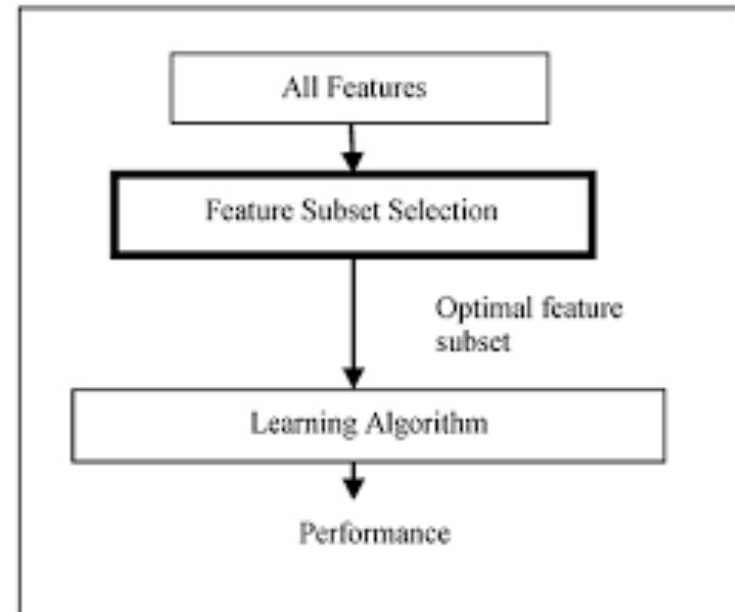
Pre-processing: feature subset selection

Key Benefits

- *Reduced cost of data acquisition*
- *Decreased time required for inference*
- *Increased comprehensibility*
- *Increased accuracy*

Goals

- *To avoid the phenomenon of the overfitting*
- *To obtain a more fast and cost-effective model*
- *Deepened understanding of the data generating process*



Pre-processing: feature subset selection

FILTERS

- **Univariate**

- » *We define a measure of association between attribute X and class Y , Mutual Information*
- » *Sort the attributes according to the chosen association measure*
- » *The first " r " attributes of the sorting are selected*
- » *Relevant attributes are detected but redundant attributes are not eliminated*

- **Multivariate**

- » *Attempt to identify relevant and redundant attributes simultaneously*
- » *Select a subset of attributes associated with the class but not related to each other with each other*
- » *Symmetrical measures of uncertainty or correlation-based measures*

Pre-processing: feature subset selection

Comparison between Univariate and Multivariate approach

	Advantages	Disadvantages
Univariate	<ul style="list-style-type: none">• <i>speed</i>• <i>scalability</i>• <i>classifier independent</i>	<ul style="list-style-type: none">• <i>ignore dependencies between attributes</i>• <i>ignore interactions with the classifier</i>
Multivariate	<ul style="list-style-type: none">• <i>models the dependency between attributes</i>• <i>classifier independent</i>• <i>favorable computational cost compared to wrappers</i>	<ul style="list-style-type: none">• <i>slower than univariate techniques</i>• <i>less scalable than univariate techniques</i>• <i>ignores classifier interactions</i>

Pre-processing: feature subset selection

Comparison between Univariate and Multivariate approach

Univariate	Multivariate
<i>Parametric</i>	Correlation Feature Selection (CFS) Relief Blanket
t-test ANOVA Mutual Information	
<i>Non-Parametric</i>	
Mann-Whitney Kruskal-Wallis Permutation test	

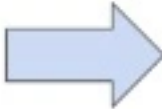
Pre-processing: feature creation

Generate new attributes that are able to effectively represent important information present in the dataset, which can provide information more efficiently than the original attributes can.

METHODOLOGIES

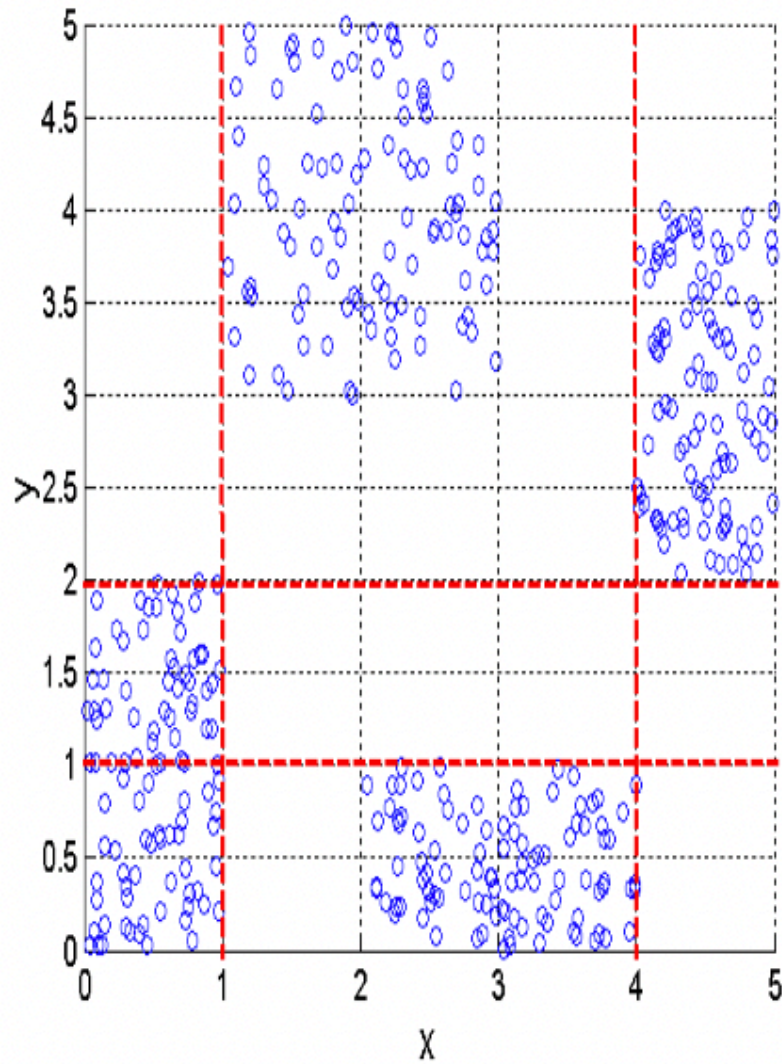
- *Feature Extraction*: specification of the particular domain being analyzed
- *Mapping data into a new space*
- *Feature Construction*: combining existing attributes together

Product	Color
1	Red
2	Green
3	Blue
4	Red
5	Blue

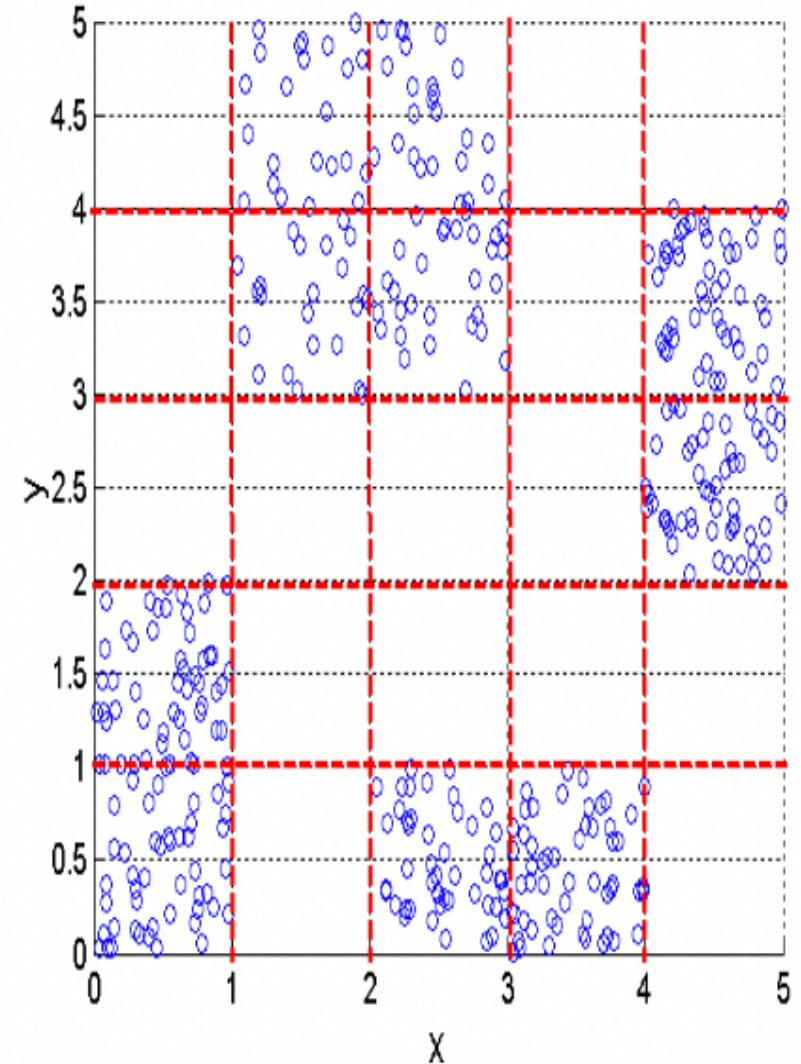


Product	Red	Green	Blue
1	1	0	0
2	0	1	0
3	0	0	1
4	1	0	0
5	0	0	1

Pre-processing: discretization

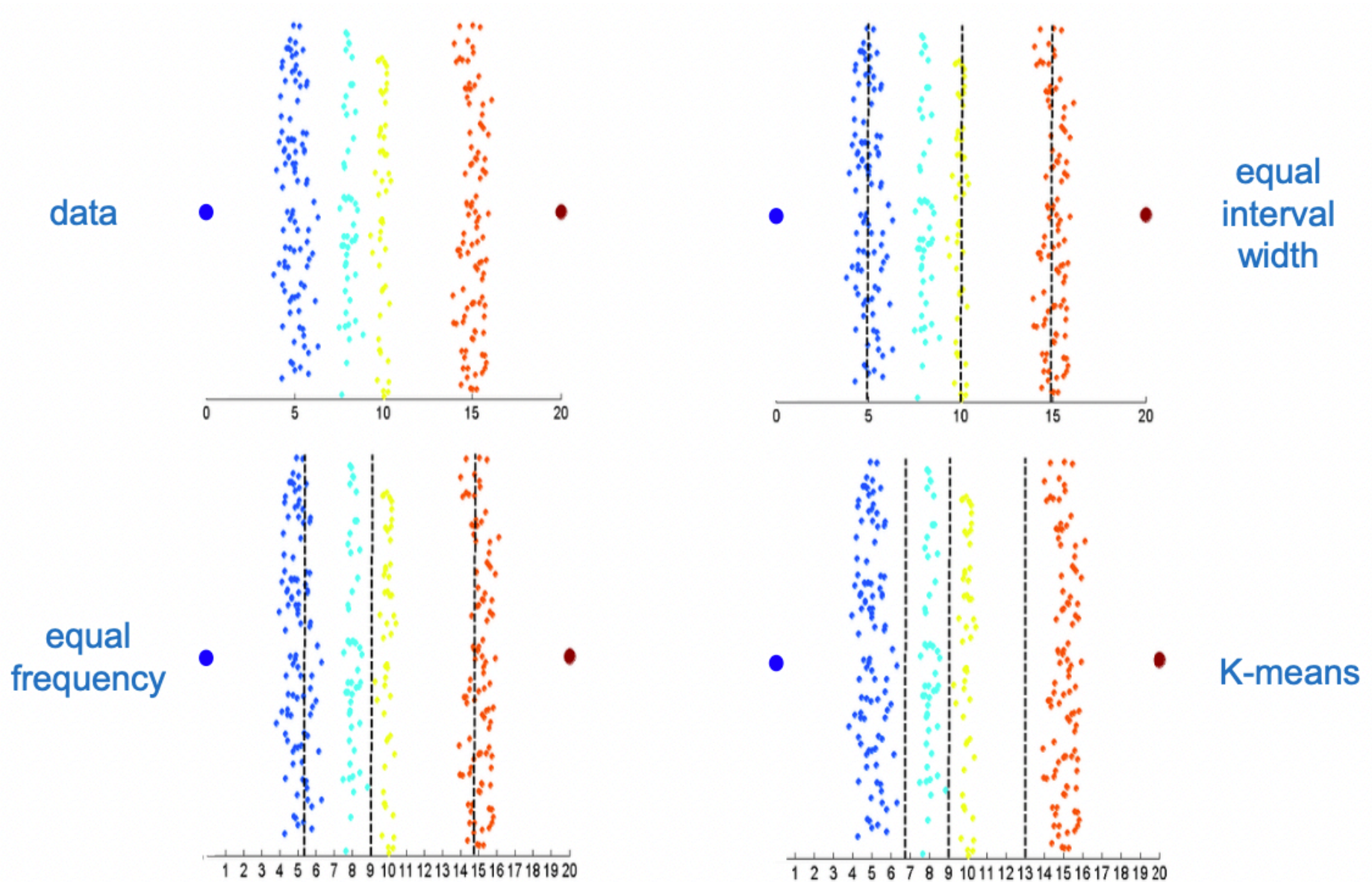


3 bins for X and Y



5 bins for X and Y

Pre-processing: discretization



Pre-processing: binarization

Attribute that can take " k " distinct values

We associate each value of the medium with an integer in the range $[0, k-1]$.

Taste				
<i>categories</i>	<i>integers</i>	X_1	X_2	X_3
very bad	0	0	0	0
bad	1	0	0	1
discrete	2	0	1	0
good	3	0	1	1
excellent	4	1	0	0

Pre-processing: binarization

Be careful not to induce fictitious order relationships and/or associations between transformed attributes.

Asymmetric binary attributes must be used (association analysis).

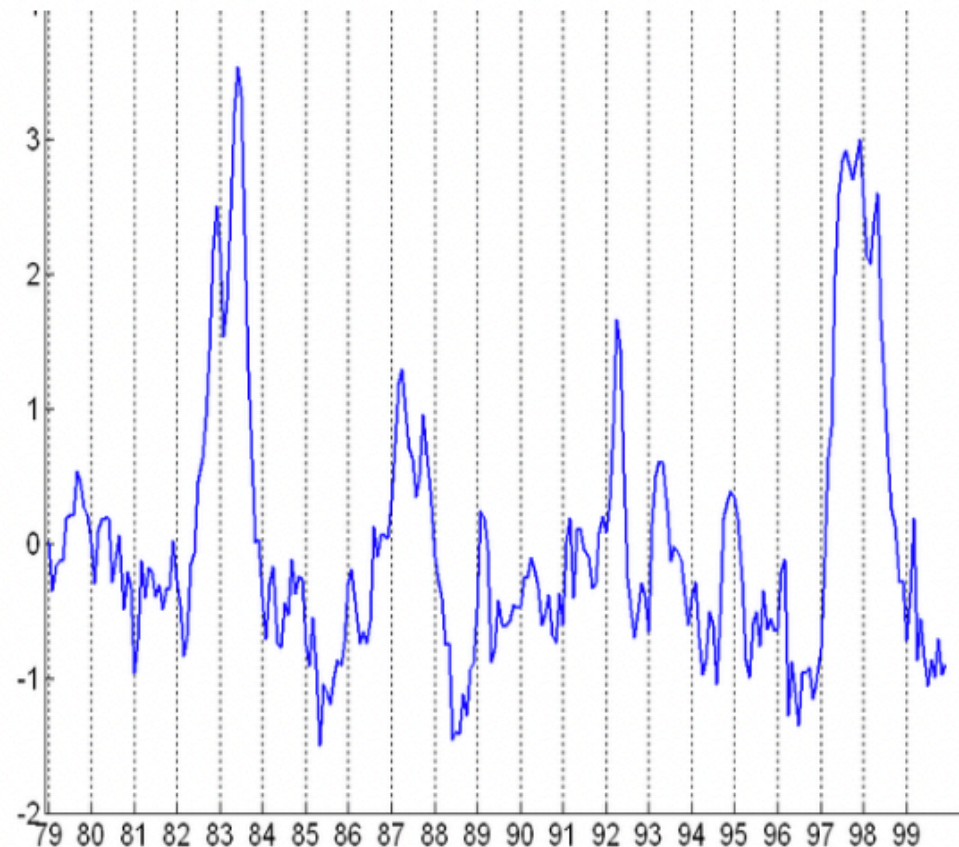
Taste						
<i>categories</i>	<i>integers</i>	X_1	X_2	X_3	X_4	X_5
very bad	0	1	0	0	0	0
bad	1	0	1	0	0	0
discrete	2	0	0	1	0	0
good	3	0	0	0	1	0
excellent	4	0	0	0	0	1

Pre-processing: transformation

A function that maps the entire set of values of a given attribute to a new set of values such that the original values can be identified by one of the new values, such as simple functions of the type:

$$X^k \quad \log(X) \quad \exp(X) \quad |X|$$

- *standardization*
- *normalization*



Pre-processing: missing data

The problem of *incomplete observations* or *missing data* is very relevant especially when dealing with high dimensional databases (high number of attributes).

Some causes of the phenomenon of missing data are:

- *The value of the attribute is not always observable/measurable*
- *The attribute was not considered relevant, so we started to measure and record the values only from a certain point in time.*
- *Trivial misunderstandings about storage guidelines*
- *Malfunctioning of the measuring or storage devices*
- *Removal of values that are inconsistent with the values of other attributes*

The problem of missing data must be taken seriously before starting any data mining study.

The problem of *missing data* should be seriously considered before any data mining study is undertaken.

Pre-processing: missing data

Consider a portion of a database in which blood test values of hospitalized patients are recorded.

ID	Birth date	Gender	Height	Weight	Sampling date	Folic acid	ALT	Ferritin	Eosinophils
122345	27-Aug-65	maschio	182	75	12-May-11	6.0		16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio		91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11		9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

Pre-processing: missing data | skip the record

The main approaches to dealing with missing data are:

- ***Ignore the entire observation (record)***: typically applied if the missing value is relative to the class variable
- ***Manual filling***: very time consuming approach
- ***Use of a global constant***: the missing value is replaced by a dummy value such as dummy value like "Unknown", "999", ...
- ***Mean replacement***: replace the missing value with the average of the attribute calculated on the observations not missing
- ***Conditional mean replacement***: replace the missing value with the average of the attribute of the attribute calculated on the observations not missing and only for those observations that belong to the same class.
- ***Most probable replacement***: we determine the most probable value through regression or regression or Bayesian approaches or decision trees.

Pre-processing: missing data | skip the record

ID	Birth date	Gender	Height	Weight	Sampling date	Folic acid	ALT	Ferritin	Eosinophils
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

The displayed portion of the database consists of 14 observations, cases, or records.

After ignoring records with missing data, 8 observations remain, $6/14 = 0.43$

Pre-processing: missing data | manual filling

ID	Birth date	Gender	Height	Weight	Sampling date	Folic acid	ALT	Ferritin	Eosinophils
122345	27-Aug-65	maschio	182	75	12-May-11	6.0		16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio		91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11		9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

Extremely time-consuming to complete a task in a reliable way.

Pre-processing: missing data | global constant

ID	Birth date	Gender	Height	Weight	Sampling date	Folic acid	ALT	Ferritin	Eosinophils
122345	27-Aug-65	maschio	182	75	12-May-11	6.0	999	16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	999
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio	999	91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11	999	9	63	66
193649	28-Sep-47	maschio	174	999	15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48	unknown	170	61	13-May-11	16.1	13	999	189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

A global variable, possibly attribute dependent, is chosen with which cells associated with missing data are filled.

Pre-processing: missing data | mean replacement

ID	Birth date	Gender	Height	Weight	Sampling date	Folic acid	ALT	Ferritin	Eosinophils
122345	27-Aug-65	maschio	182	75	12-May-11	6.0		16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio		91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11	12	9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

Let's consider the attribute "*Folic acid*". The record with ID = 99372 has a missing value that through mean replacement can be obtained by calculating the average of the values of "*Folic acid*" which is equal to **12**, value that is used to fill the missing data.

Pre-processing: missing data | conditional mean repl.

ID	Birth date	Gender	Height	Weight	Sampling date	Folic acid	ALT	Ferritin	Eosinophils
122345	27-Aug-65	maschio	182	75	12-May-11	6.0		16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio		91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11		9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

Suppose the class variable is *Gender*. The *conditional mean replacement* scheme computes the conditional mean over the various values of the class variable.

Pre-processing: missing data | conditional mean repl.

ID	Birth date	Gender	Height	Weight	Sampling date	Folic acid	ALT	Ferritin	Eosinophils
122345	27-Aug-65	maschio	182	75	12-May-11	6.0	25.7	16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio		91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11		9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

The average of the variable *ALT* for "*gender = male*" is equal to 25.7. This value is used to fill in the missing data for the record with ID=122345.

Pre-processing: missing data | most probable repl.

ID	Birth date	Gender	Height	Weight	Sampling date	Folic acid	ALT	Ferritin	Eosinophils
122345	27-Aug-65	maschio	182	75	12-May-11	6.0		16	177
123467	10-May-82	maschio	177	86	28-Apr-11	4.5	15	66	65
234991	11-Apr-45	maschio	169	78	4-May-11	7.3	30	89	120
34652	5-Jan-81	femmina	165	54	27-Apr-11	13.2	32	139	
23648	10-Feb-77	femmina	172	58	12-May-11	14.5	16	155	230
113904	5-Jul-46	maschio	178.5	91	9-May-11	18.2	19	137	156
384671	31-Mar-82	femmina	181	58	17-May-11	9.7	13	179	120
99372	6-May-78	femmina	162	62	12-May-11		9	63	66
193649	28-Sep-47	maschio	174		15-May-11	10.9	21	72	36
450027	24-Jun-83	femmina	158	47	21-May-11	11.4	27	134	149
33948	30-Jul-79	maschio	184	79	17-May-11	19.3	23	177	230
677309	21-Dec-48		170	61	13-May-11	16.1	13		189
123884	16-Sep-84	maschio	168	59	7-May-11	17.2	32	83	159
64986	22-Oct-80	maschio	173	88	16-May-11	7.2	40	65	189

We construct a simple linear regression model of the type:

$$height = a_0 + a_1 * weight = 158 + 0.222 * weight = 158 + 0.222 * 91 = \mathbf{178.5}$$

Pre-processing: missing data

Methods:

- *Using a global constant*
- *Mean replacement*
- *Conditional mean replacement*
- *Most probable replacement*

introduce bias into the data.

The values imputed in these ways may not be correct.

However, the ***Most Probable Replacement*** method is very often used. It uses information about available data to predict missing values.

