Department of Mathematics
University of Calabria

Data Warehouse and Data Mining

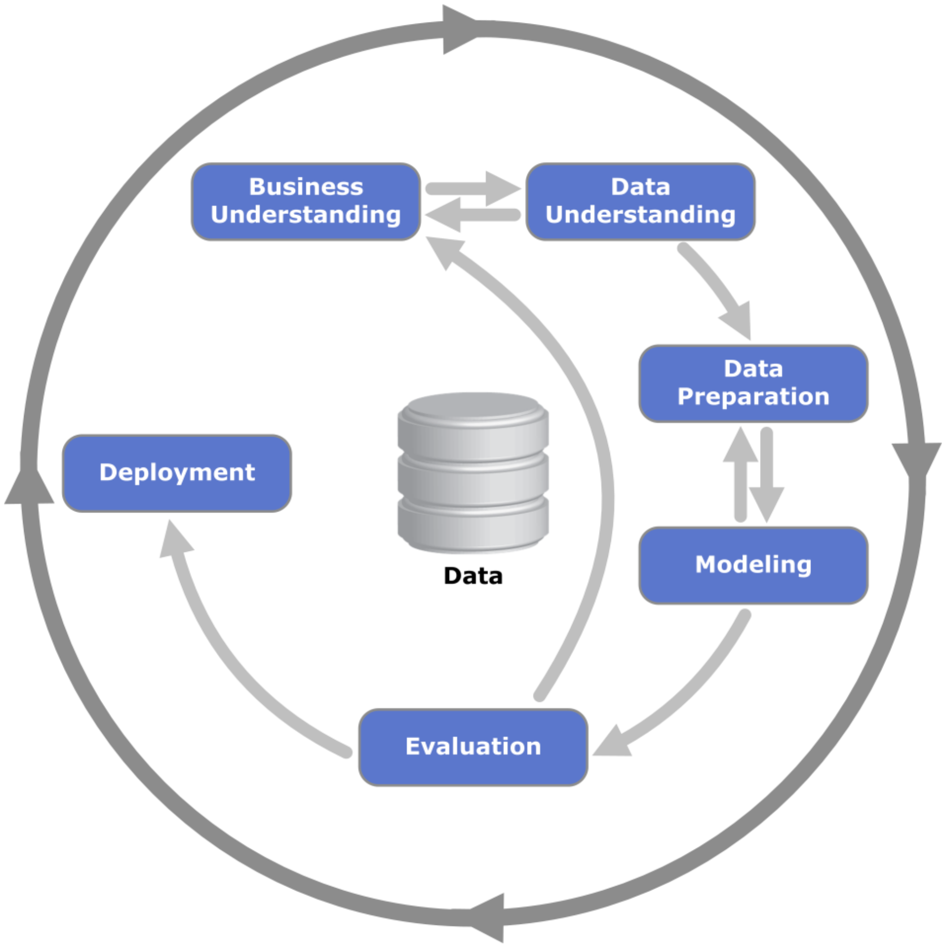
Module II – Data Mining

Data Preparation

Ph.D. Ettore Ritacco

Department of Mathematics
University of Calabria

The Knowledge Discovery Process (CRISP-DM)



About the Lecture

- Main Source:
  - Tan, Steinbach, Kumar “Introduction to Data Mining”

Data Preparation

Data cleaning	• Fill in missing values, smooth noisy data, identify or remove outliers, remove redundant values or values with too high (or too low) variability, resolve inconsistencies
Data integration	• Integration of multiple databases, data cubes, or files
Data transformation	• Normalization and aggregation
Data reduction	• Obtain a reduced representation in volume (which produces similar statistical results)
Data discretization	• Sometimes required by the classification algorithms or computational cost issues

Data Quality

Data in the real world are dirty

- *incomplete*: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data
- *noisy*: containing errors or outliers
- *inconsistent*: containing discrepancies in codes or names

No quality in data? The no quality in mining results!

BAD INPUT = WORSE OUTPUT

Data Cleaning

- Identify outliers and smooth out noisy data
- Fill in missing values
- Correct inconsistent data
- Remove irrelevant data

Data Cleaning – Noisy Data

- Noise: random error or variance in a measured variable

Incorrect attribute values

- faulty data collection instruments
- data entry problems
- data transmission problems
- technology limitation
- inconsistency in naming convention

Data Cleaning - Missing Data

Data is not always available

- E.g., many tuples have no recorded value for several attributes, such as customer income in sales data

Missing data may be due to

- Instrument malfunction
- Inconsistency with other recorded data and thus deleted
- Data not entered due to misunderstanding
- Information without importance at the time of entry
- Not register history or changes of the data

Missing data may need to be inferred.

How to Handle Missing Data?

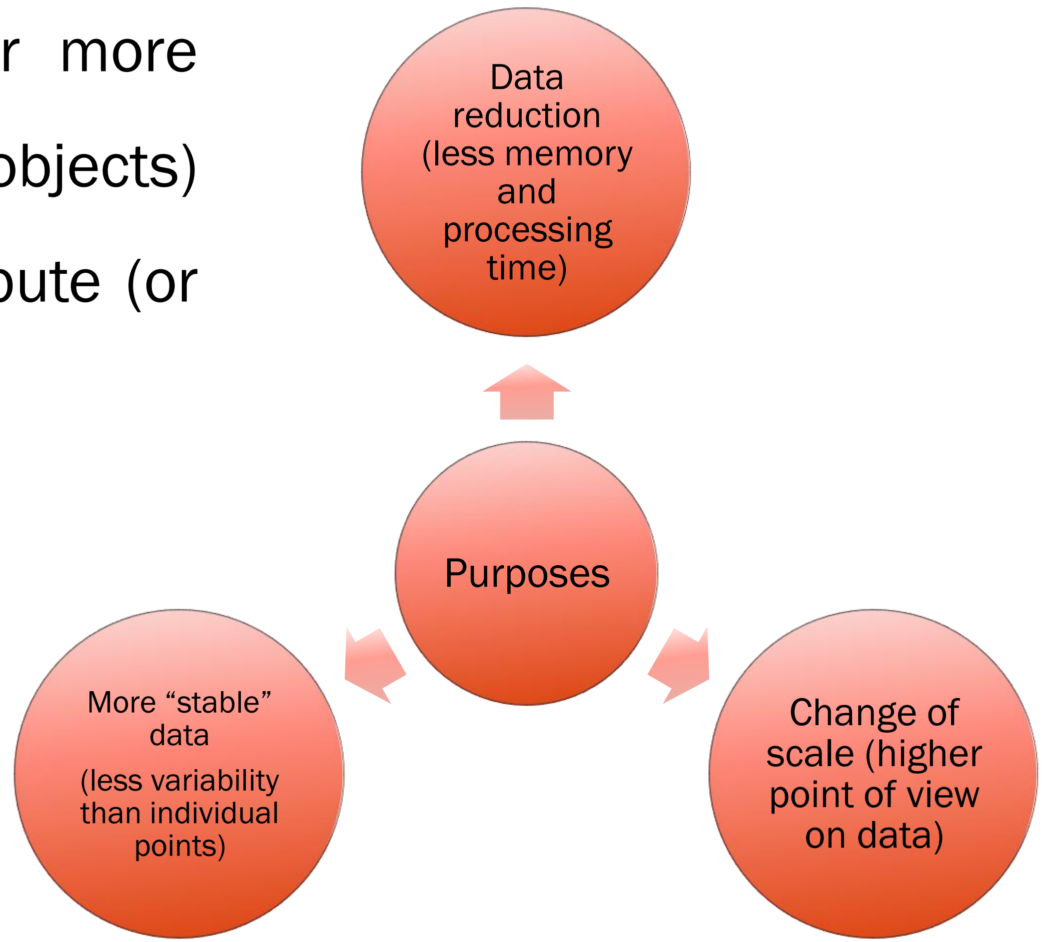
- Ignore the Missing Value During Analysis
- Ignore the tuple (usually done when class label is missing)
- Use a global constant to fill in the missing value (e.g. "unknown", a new class value)
- Estimate Missing Values
  - Use the attribute mean/mode to fill in the missing value
  - Use the most probable value to fill in the missing value: inference-based such as Bayesian formula or decision tree
  - Replace with all possible values (weighted by their probabilities)

Data Preparation

- **Aggregation**
- Sampling
- Dimensionality Reduction
  - Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Aggregation

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



Disadvantages: Potential loss of interesting details

Data Aggregation – Example

Transaction ID	Item	Store Location	Date	Price	...
⋮	⋮	⋮	⋮	⋮	
101123	Watch	Chicago	09/06/04	$25.99	...
101123	Battery	Chicago	09/06/04	$5.99	...
101124	Shoes	Minneapolis	09/06/04	$75.00	...
⋮	⋮	⋮	⋮	⋮	

- Transactions of a single store can be replaced by a single storewide transaction
- Aggregation operation depends on the type of the attribute (i.g. price can be averaged, summed...)

Data Preparation

- Aggregation
- **Sampling**
- Dimensionality Reduction
  - Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Data Sampling

- Horizontal selection: instances removal.
- To process all the data is often too expensive or time consuming
- Often used for both the preliminary investigation of the data and the final data analysis
- *Representative Sample*
  - A sample is representative if it holds (“almost”) the same statistical properties of the original set.
  - Using a representative sample will work almost as well as using the entire data sets

Sampling Tecniques

Simple Random Sampling

- There is an equal probability of selecting any particular item

Sampling without replacement

- As each item is selected, it is removed from the population

Sampling with replacement

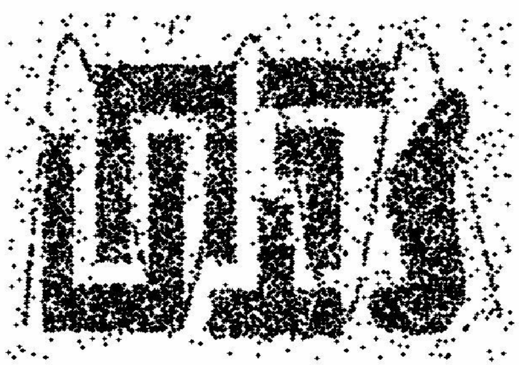
- Objects are not removed from the population as they are selected for the sample.
- The same object can be picked up more than once

Stratified sampling

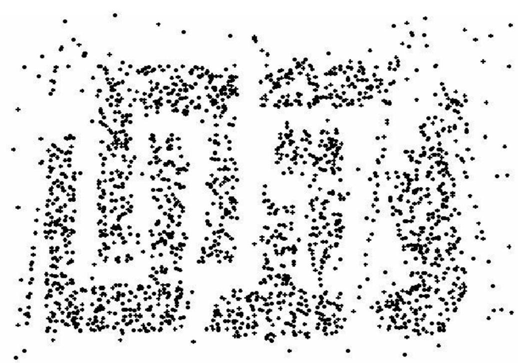
- Split the data into several partitions; then draw random samples from each partition
- Fixed size/percentage for ech class

Sampling – Loss of information

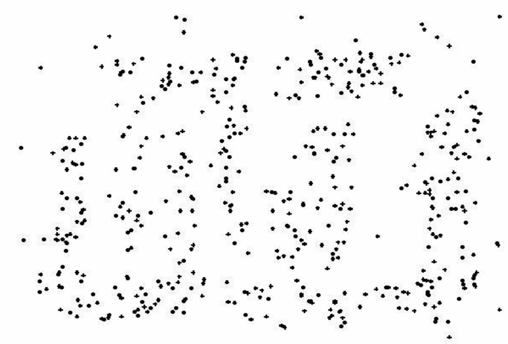
- Which is the most appropriate sample size?



8000 points



2000 Points



500 Points

- If the sample is ***too big***, there is no advantage of sampling
- If the sample is ***small***, erroneus patterns can be detected!

Progressive Sampling

- Start with a small sample size
- Increase the sample size until a sufficient sample size is reached
- Example:
  - Extimate the increase of the accuracy of a predictive model when increasing sample size.
  - Stop when the increase in the accuray levels off.

Data Preparation

- Aggregation
- Sampling
- **Dimensionality Reduction**
  - Feature subset selection
- Feature creation
- Discretization and Binarization
- Attribute Transformation

Dimensionality Reduction

The curse of dimensionality

- As the dimensionality increases, data become **sparse**
- **Risk:** to produce **low quality** results with high dimensional data

Dimensionality Reduction:

- Obtaining a reduced representation of the data set that is much smaller in volume but yet produce the **good analytical results**

Dimensionality Reduction

- Avoid curse of dimensionality
- Reduce amount of time and memory required by data mining algorithms
- Produce more understandable models
- Allow data to be more easily visualized
- Eliminate irrelevant features or reduce noise

Dimensionality Reduction

- Feature Subset Selection
  - Embedded approaches
  - Filter approaches
  - Wrapper approaches
- Linear Algebra Techniques
  - Principle Component Analysis
  - Singular Value Decomposition

Department of Mathematics
University of Calabria

Feature Subset Selection

Irrelevant and redundant features can be removed

Some features can be immediately removed by using common sense or domain knowledge.

A more sistematic approach is required to select the best subset of features (Note that the possibile subsets are 2^n).

Features Subset Selection

Embedded approaches

- Features selection if often performed as a part of tha data mining algorithm

Filter approaches

- Features selection can be performed before the data mining algorithm is applied
- Using a specific and indepedent approach

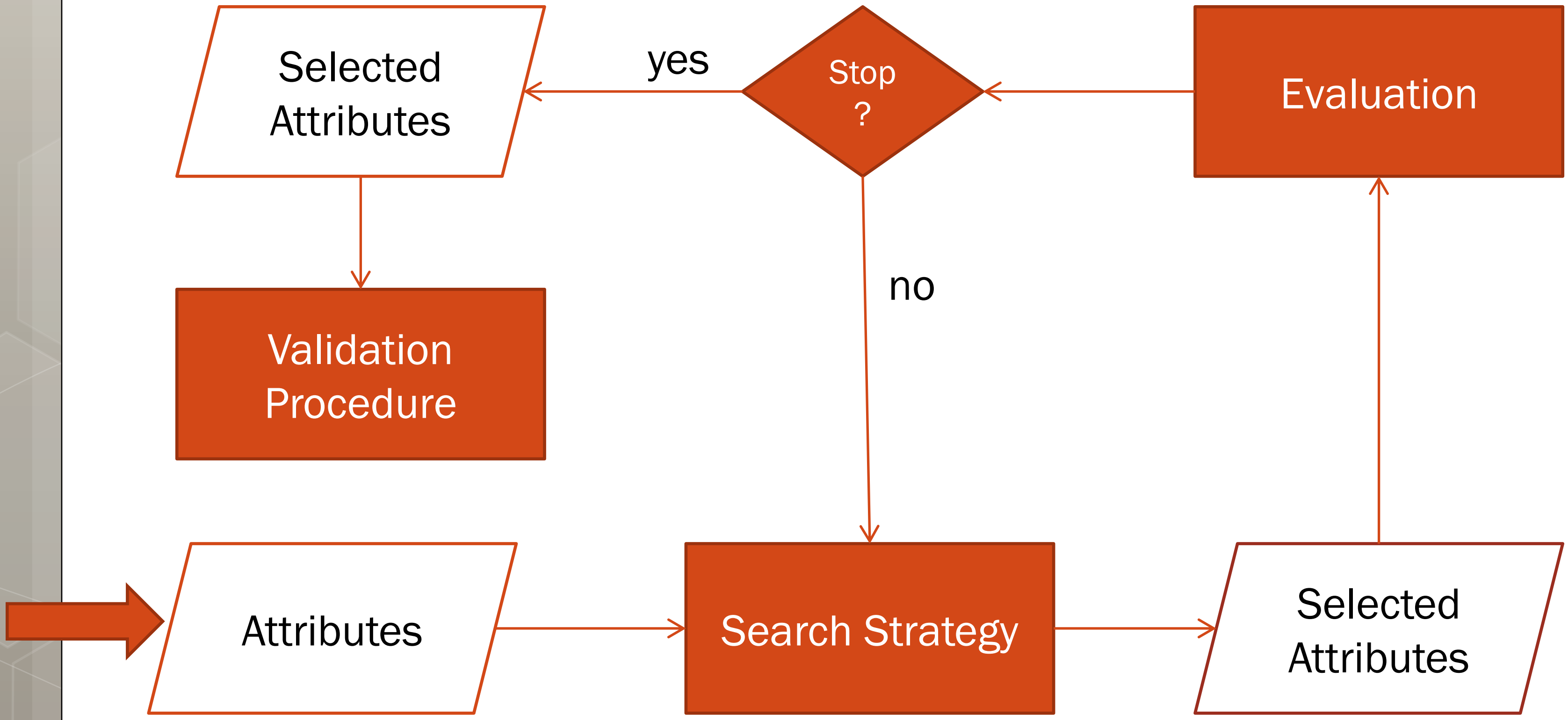
Wrapper approaches

- A data mining algorithm is used as a black box to select the most relevant features

Features Selection – The process

- It consists of 4 main components:
  - A quality measure (to evaluate the features subsets)
  - A search strategy (tha approach to be used)
  - A stopping criterion
  - A validation procedure
- Filter methods and wrapper ones differ only in the way they evaluate the subset.

Features Selection – The process



Data Preparation

- Aggregation
- Sampling
- Dimensionality Reduction
  - Feature subset selection
- **Feature creation**
- Discretization and Binarization
- Attribute Transformation

Features Creation

- Create new attributes that can capture the important information in a data set much more efficiently than the original attributes
- Three general methodologies:
  - Feature Extraction (domain-specific)
  - Mapping Data to New Space
  - Feature Construction, by combining features (e.g. density=mass/volume)

Data Preparation

- Aggregation
- Sampling
- Dimensionality Reduction
  - Feature subset selection
- Feature creation
- **Discretization and Binarization**
- Attribute Transformation

Discretization and Binarization

Problems:

- Many classification algorithms require **categorical** attributes
- Algorithms that find patterns often require **binary** attributes

Solutions:

- **Discretization:** Transforming a continous attribute into categorical attribute
- **Binarization:** Transforming continous or discrete attributes into one or more binary attributes

Binarization – Type 1

- cat_attr={awful, poor, OK, good, great}
- Conversion to 3 binary attributes

Cat_attr	Integer value	a	b	c
awful	0	0	0	0
poor	1	0	0	1
OK	2	0	1	0
good	3	0	1	1
great	4	1	0	0

Binarization – Type 1

- Tecnique:
  - Uniquely assign each original value to an integer in [0, m-1]
  - Convert each integer value to a binary number
  - Create $n = [ln(m)]$ binary attribute to replace the original one.
- Disadvantage:
  - Creation of n attributes with an unintended relationships

Department of Mathematics
University of Calabria

Binarization – Type 2

- cat_attr={awful, poor, OK, good, great}

- Introduction to 5 binary attributes:
  - awful={0,1}
  - poor={0,1}
  - OK={0,1}
  - good={0,1}
  - great={0,1}

Binarization – Type 2

- Tecnique:
  - Create m binary attribute to replace the original one.
  - Assign a value=1 to the binary attribute corresponding to the original value
  - Assign 0 to the other attributes
- Disadvantage:
  - The number of new attributes may be very large
- The best discretization depends on the algorithm being used

Discretization

Decisions to take:

- How many categories?
- How to realize the mapping?

Tecnique:

- sort the values
- divide them into *n* intervals (using *n-1* **split points)**
- Map all the values in one interval to the same categorical value

Department of Mathematics
University of Calabria

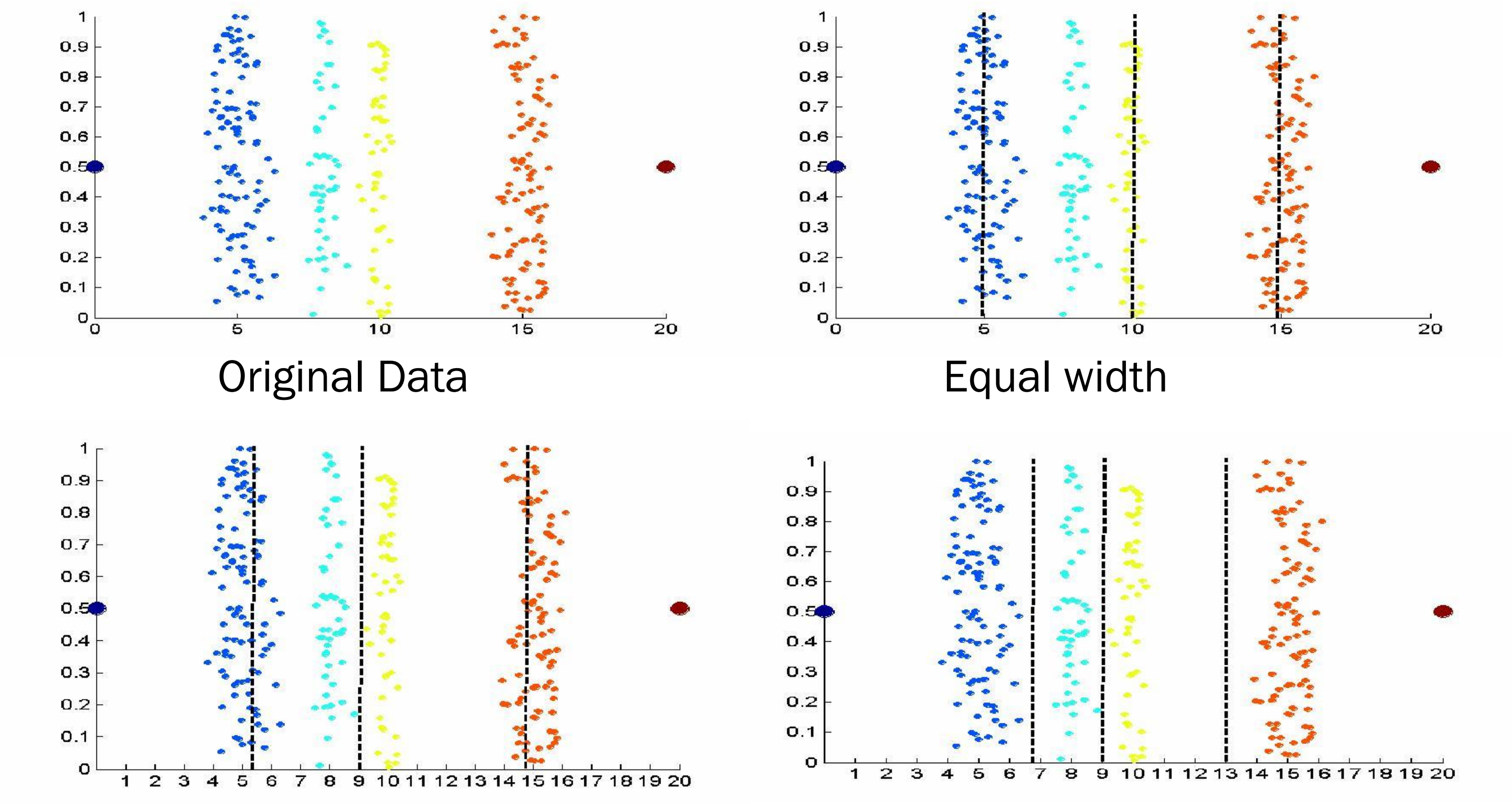
Unsupervised Discretization

- Class information is not used

APPROACHES

Equal width	• Divide the range into **n** intervals with the same width • Can be badly affected by outliers
Equal frequency	• Tries to put the same number of objects into each interval
Clustering aprroach	• Based on the use of any clustering method

Unsupervised Tecniques – Example



Original Data

Equal width

Equal frequency

Clustering based

Department of Mathematics
University of Calabria

Supervised Discretization

- **Problem:** An interval constructed without knowledge about the class distribution often contains a **mixture of classes.**
- Using class information may produce better results
- A **simple approach** is that of choosing the split points in order to maximize the purity of the intervals

Department of Mathematics
University of Calabria

Data Preparation

- Aggregation
- Sampling
- Dimensionality Reduction
  - Feature subset selection
- Feature creation
- Discretization and Binarization
- **Attribute Transformation**

Attribute Transformation

- Transformation of the values of an attribute
- Two main types:
  - Simple functional transformations
  - Normalization

Attribute Transformation – Simple functions

- Application of a function to all the values of an attributes
- Examples of trasformation:
  - logarithm function: used to reduce the range of values;
  - $|x|$function (Absolute value)
  - $\frac{1}{x}$ function: reduces the magnitude of values.

    Note: the order is reverted for values in (0,1) !

Attribute Transformation – Normalization

- Goal: to make an entire set of values to have a specific property.
- Example:
  - $x' = \frac{x-\bar{x}}{s_x}$ $\longrightarrow$ $\begin{cases} s_{x'} = 1 \\ \overline{x'} = 0 \end{cases}$
- Since mean and st.dev. are affected by outliers, their are replaced by be median and the aboslute st.dev.