Data Cleaning

- How to handle anomalous values
- How to handle outliers
- Data Transformations

Anomalous Values

- **Missing values**
  - NULL, ?
- **Unknown Values**
  - Values without a real meaning
- **Not Valid Values**
  - Values not significant

Manage Missing Values

1. Elimination of records
2. Substitution of values

Note: it can influence the original distribution of numerical values

- Use mean/median/mode
- Estimate missing values **using the probability distribution** of existing values
- Data Segmentation and using mean/mode/median of each **segment**
- Data Segmentation and using **the probability distribution within the segment**
- Build a model of **classification/regression** for computing missing values

Discretization

- Discretization is the process of converting a continuous attribute into an ordinal attribute
  - A potentially infinite number of values are mapped into a small number of categories
  - Discretization is commonly used in classification
  - Many classification algorithms work best if both the independent and dependent variables have only a few values

Discretization: Advantages

- Hard to understand the optimal discretization
  - We should need the real data distribution
- Original values can be **continuous** and **sparse**
- Discretized data can be **simple** to be interpreted
- Data distribution after discretization can have a **Normal shape**
- Discretized data can be too much **sparse yet**
  - Elimination of the attribute

Unsupervised Discretization

- Characteristics:
  - No label for the instances
  - The number of classes is unknown

- Techniques of *binning*:
  - **Natural binning** → Intervals with the same width
  - **Equal Frequency binning** → Intervals with the same frequency
  - **Statistical binning** → Use statistical information (Mean, variance, Quartile)

Discretization of quantitative attributes

Solution: each value is replaced by the interval to which it belongs.

height: 0-150cm, 151-170cm, 171-180cm, >180c

weight: 0-40kg, 41-60kg, 60-80kg, >80kg

income: 0-10ML, 11-20ML, 20-25ML, 25-30ML, >30ML

CID	**height**	**weight**	**income**
1	151-171	60-80	>30
2	171-180	60-80	20-25
3	171-180	60-80	25-30
4	151-170	60-80	25-30

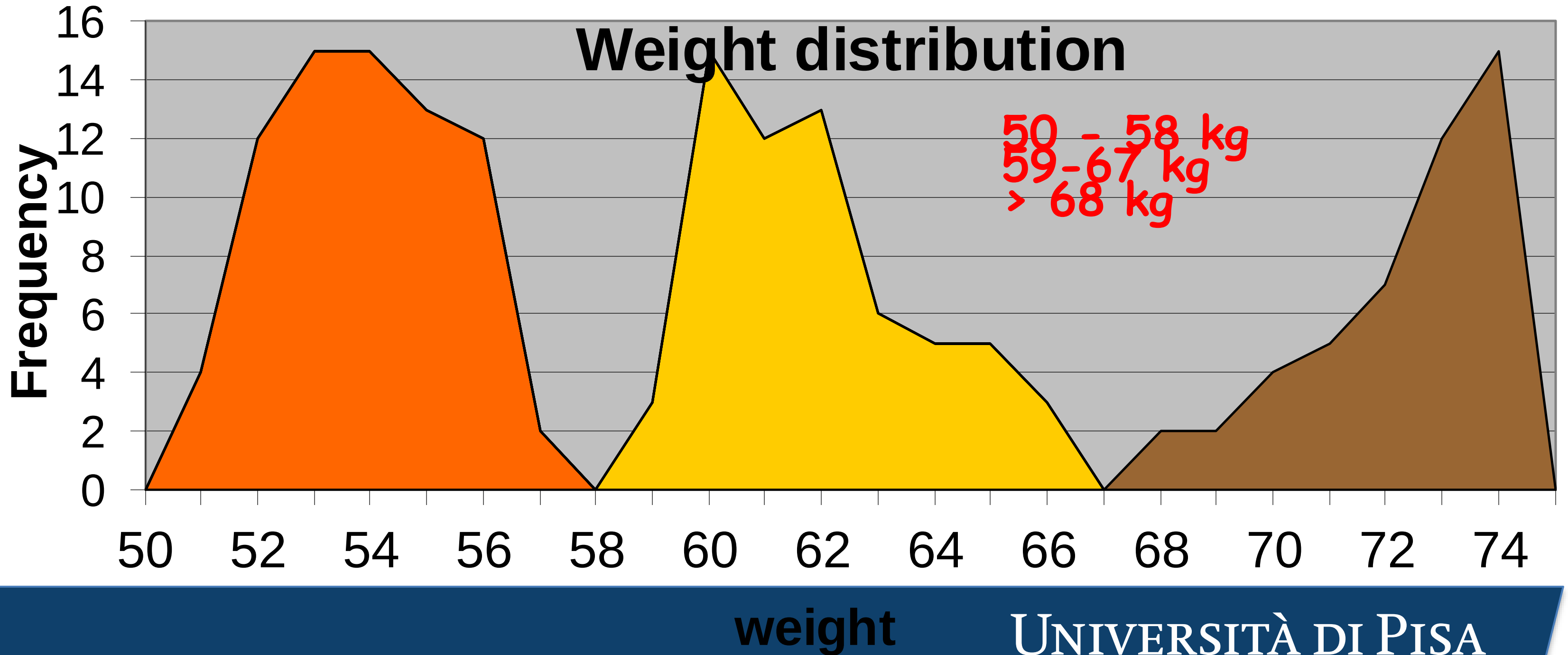
Problem: the discretization may be useless (see **weight**).

How to choose intervals?

1. Interval with a fixed “reasonable” granularity
   Ex. intervals of 10 cm for height.

2. Interval size is defined by some domain dependent criterion
   Ex.: 0-20ML, 21-22ML, 23-24ML, 25-26ML, >26ML

3. Interval size determined by analyzing data, studying the distribution and find breaks or using clustering



Natural Binning

- Simple
- Sort of values, subdivision of the range of values in *k* parts with the same size

$$\delta = \frac{x_{\max} - x_{\min}}{k}$$

- Element x_j belongs to the class i if

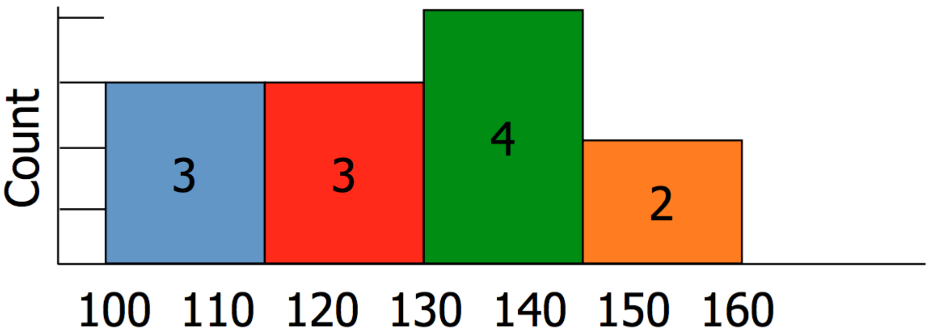
$$x_j \in [x_{min} + i\delta,\ x_{min} + (i+1)\delta)$$

- It can generate distribution very unbalanced

Example

Bar	Beer	Price
A	Bud	100
A	Becks	120
C	Bud	110
D	Bud	130
D	Becks	150
E	Becks	140
E	Bud	120
F	Bud	110
G	Bud	130
H	Bud	125
H	Becks	160
I	Bud	135

- δ =(160-100)/4 = 15
- class 1: [100,115)
- class 2: [115,130)
- class 3: [130,145)
- class 4: [145, 160]



Equal Frequency Binning

- Sort and count the elements, definition of k intervals of f, where:

$$f = \frac{N}{k}$$

(*N* = number of elements of the sample)

- The element x_i belongs to the class j if

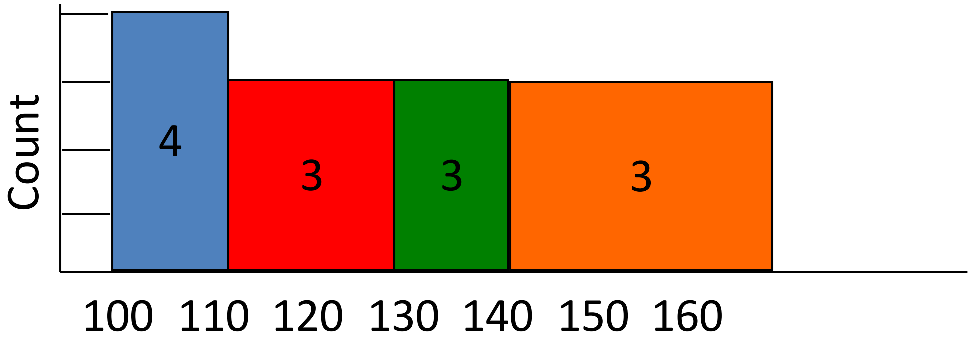
$$j \times f \leq i < (j+1) \times f$$

- It is not always suitable for highlighting interesting correlations

Example

Bar	Beer	Price
A	Bud	100
A	Becks	120
C	Bud	110
D	Bud	130
D	Becks	150
E	Becks	140
E	Bud	120
F	Bud	110
G	Bud	130
H	Bud	125
H	Becks	160
I	Bud	135

- $f = 12/4 = 3$
- class 1: {100,110,110}
- class 2: {120,120,125}
- class 3: {130,130,135}
- class 4: {140,150,160}



How many classes?

- If too few

  $\Rightarrow$ Loss of information on the distribution

- If too many

  => Dispersion of values and does not show the form of distribution

- The optimal number of classes is function of N elements (Sturges, 1929)

$$C = 1 + \frac{10}{3}\log_{10}(N)$$

- The optimal width of the classes depends on the variance and the number of data (Scott, 1979)

$$h = \frac{3,5 \cdot s}{\sqrt{N}}$$

Supervised Discretization

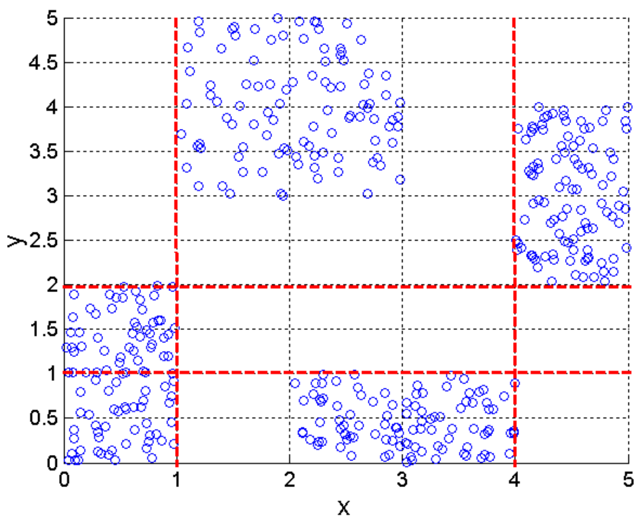
- **Characteristics**:
  - The discretization has a quantifiable goal
  - The number of classes is known
- **Techniques**:
  - discretization based on Entropy
  - discretization based on percentiles

Entropy based approach

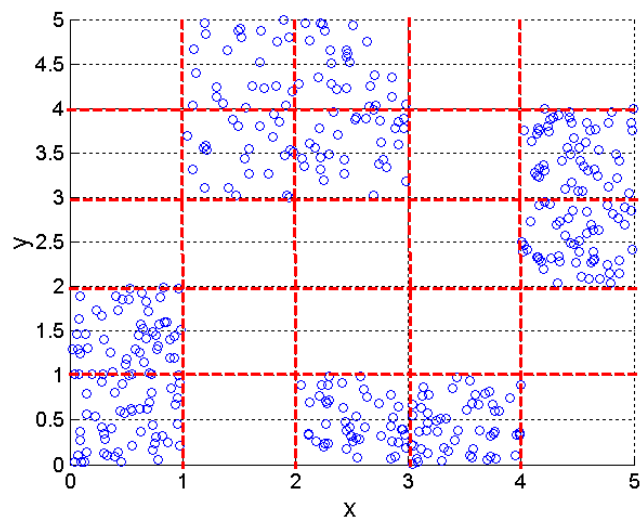
- Minimizes the entropy wrt a label
- **Goal:** maximizes the purity of the intervals
- Decisions about the purity of an interval and the minimum size of an interval
- To overcome such concerns use statistical based approaches:
  - start with each attribute value as a separate interval
  - create larger intervals by merging adjacent intervals that are similar according to a statistical test

A simple approach

- Starts by bisecting the initial values so that the resulting two intervals give minimum entropy.
- The splitting process is then with another interval, typically choosing the interval with the worst (highest) entropy
- Stop when a user-specified number of intervals is reached, or a stopping criterion is satisfied.



3 categories for both x and y



5 categories for both x and y

Binarization

- Binarization maps a continuous or categorical attribute into one or more binary variables
- Typically used for association analysis
- Often convert a continuous attribute to a categorical attribute and then convert a categorical attribute to a set of binary attributes
  - Association analysis needs asymmetric binary attributes
  - Examples: eye color and height measured as {low, medium, high}

Binarization

n = $\log_2(m)$ binary digits are required to represent m integers.

It can generate some correlations

Table 2.5. Conversion of a categorical attribute to three binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3
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

- **One variable for each possible value**
- Only presence or absence
- Association Rules requirements

Table 2.6. Conversion of a categorical attribute to five asymmetric binary attributes.

Categorical Value	Integer Value	x_1	x_2	x_3	x_4	x_5
awful	0	1	0	0	0	0
poor	1	0	1	0	0	0
OK	2	0	0	1	0	0
good	3	0	0	0	1	0
great	4	0	0	0	0	1

Data Transformation: Motivations

- Data with errors and incomplete
- Data not adequately distributed
  - Strong asymmetry in the data
  - Many peaks
- Data transformation can reduce these issues

Attribute Transformation

- An attribute transform is a function that maps the entire set of values of a given attribute to a new set of replacement values such that each old value can be identified with one of the new values
  - Simple functions: x^k, $\log(x)$, e^x, $|x|$
  - Normalization
    - Refers to various techniques to adjust to differences among attributes in terms of frequency of occurrence, mean, variance, range
    - Take out unwanted, common signal, e.g., seasonality
  - In statistics, standardization refers to subtracting off the means and dividing by the standard deviation

Properties of trasformation

- Define a transformation T on the attribute X:

  $$Y = \mathrm{T}(X)$$

  such that :

  - Y preserve the **relevant** information of X
  - Y eliminates at least one of the problems of X
  - Y is more **useful** of X

Transformation Goals

- **Main goals:**
  - stabilize the variances
  - normalize the distributions
  - Make linear relationships among variables

- **Secondary goals:**
  - simplify the elaboration of data containing features you do not like
  - represent data in a scale considered more suitable

Why linear correlation, normal distributions, etc?

- Many statistical methods require
  - linear correlations
  - normal distributions
  - the absence of outliers

- Many data mining algorithms have the ability to automatically treat **non-linearity** and **non-normality**
  - The algorithms work still better if such problems are treated

Normalizations

- min-max normalization

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- z-score normalization

$$v' = \frac{v - mean_A}{stand\_dev_A}$$

- normalization by decimal scaling

$$v' = \frac{v}{10^j}$$ Where j is the smallest integer such that Max($|v'|$)<1

Example of decimal scaling

- Let the input data is: -10, 201, 301, -401, 501, 601, 701
- To normalize the above data,
  - Step 1: Maximum absolute value in given data(m): 701
  - Step 2: Divide the given data by 1000 (i.e j=3)
  - Result: -0.01, 0.201, 0.301, -0.401, 0.501, 0.601, 0.701

Transformation functions

- Exponential transformation

$$T_p(x) = \begin{cases} ax^p + b & (p \neq 0) \\ c \log x + d & (p = 0) \end{cases}$$

- with a,b,c,d and p real values
  - Preserve the order
  - Preserve some basic statistics
  - They are continuous functions
  - They are derivable
  - They are specified by simple functions

Better Interpretation

- Linear Transformation

  1€ = 1936.27 Lit.

  – $p=1$, $a= 1936.27$,$b =0$

$$T_p(x) = \begin{cases} ax^p + b & (p \neq 0) \\ c \log x + d & (p = 0) \end{cases}$$

  °C= 5/9(°F -32)

  – $p = 1$, $a = 5/9$, $b = -160/9$

Stabilizing the Variance

- **Logarithmic Transformation**

$$T(x) = c \log x + d$$

  – Applicable to positive values
  – Makes homogenous the variance in log-normal distributions
    - E.g.: normalize seasonal peaks

Logarithmic Transformation: Example

Bar	***Beer***	***Gain***
A	Bud	20
A	Becks	10000
C	Bud	300
D	Bud	400
D	Becks	5
E	Becks	120
E	Bud	120
F	Bud	11000
G	Bud	1300
H	Bud	3200
H	Becks	1000
I	Bud	135

2481,8182	Mean
4079,0172	Standard Deviation
5	Min
120	1° Quartile
400	Median
2250	3° Quartile
11000	Max

Data are sparse!!!

Logarithmic Transformation: Example

Bar	*Beer*	*Gain (log)*
A	Bud	1,301029996
A	Becks	4
C	Bud	2,477121255
D	Bud	2,602059991
D	Becks	0,698970004
E	Becks	2,079181246
E	Bud	2,079181246
F	Bud	4,041392685
G	Bud	3,113943352
H	Bud	3,505149978
H	Becks	3
I	Bud	2,130333768

Mean	2,595567
Standard Deviation	1,065137
Min	0,69897
First Quartile	2,079181
Median	2,60206
3rd Quartile	3,309547
Max	4,041393

Stabilizing the Variance

$$T(x) = ax^p + b$$

- **Square-root Transformation**
- $p = 1/c$, c integer number
  - To make homogenous the variance of particular distributions e.g., Poisson Distribution
- **Reciprocal Transformation**
  - $p < 0$
  - Suitable for analyzing time series, when the variance increases too much wrt the mean

Outliers on single dimension

- **Interquartile Range for detecting outliers**
  - IQR = Q3-Q1
  - Define a range with lower bound L=Q1-1.5*IQR and upper bound U=Q3+1.5*IQR
  - X is outlier if X > U or X<L
  - For the **substitution** of the outlier X you have two options
    - With L or U
    - Median
- **Z-score based approach:**
  - Standardize the data using z-score
  - When data is regularly distributed, 95% of instances fall between z-scores of $\pm$1.96 and 99% of cases fall between z-scores of $\pm$ 2.58.
  - A z-score of 0 denotes the mean.
  - The usual value for identifying outliers is $\pm$ 3.29: any z-score greater than +3.29 or less than -3.29 is an outlier case
  - **Substitution**: converting X with z-score > 3.29
    - to the value that corresponds with a **z-score of 3.0**. This approach assumes that a normal distribution includes values that fall within 3$\boldsymbol{\sigma}$ above or below a standardized mean score of 0.
    - to the value that corresponds with a z-score of 0 that is the **mean value**