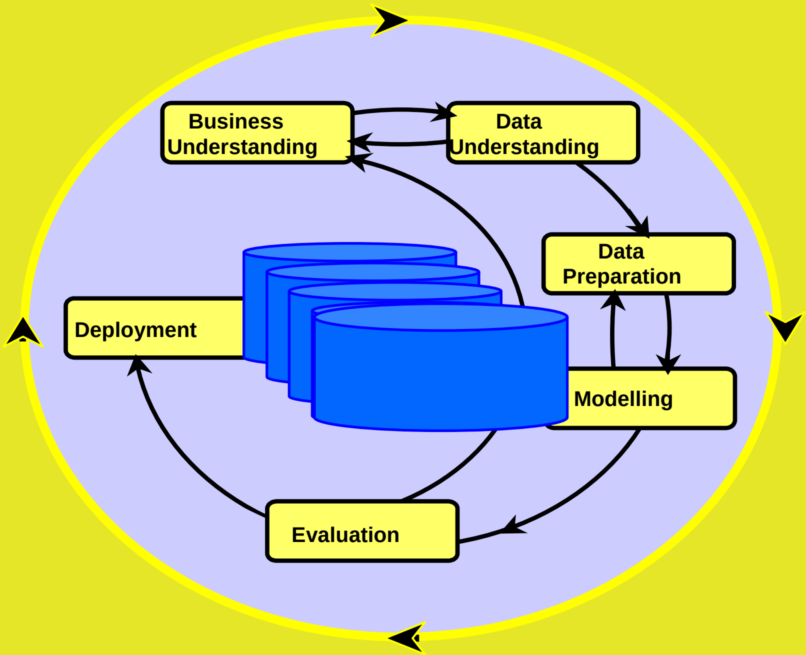
Statistic Methods in Data Mining



Data Mining Process
(Part 2)

Professor Dr. Gholamreza Nakhaeizadeh

Short review of the last lecture

Data Understanding

- **Collect initial data**
  - Can the data be accessed effectively and efficiently ?
  - Is there any restriction in collecting the data ?
  - what are the needed data ? where are the data ?
  - Examples of data sources
  - Data warehouse
- **Describe data**
  - Some of data characterization measures
  - Data Structure

Observation, attribute type (nominal, ordinal, interval, ratio, qualitative, quantitative, discrete)

Data Type: Cross-section data, time series data, panel data, spatial data…

- **Explore data**

- Data exploration Tools

Using descriptive data summarization (mean, median, mode, variance,…)

- Using Visualization
- OLAP

- **Verify data quality**

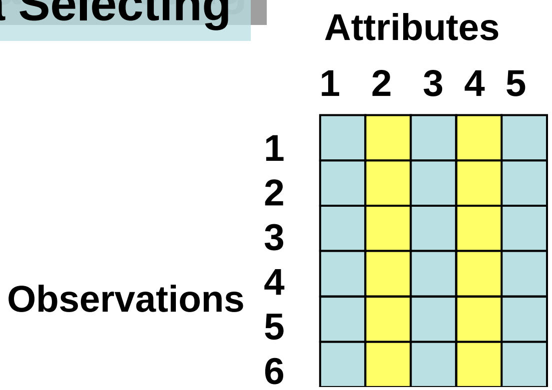
- Are data accurate ? Are data complete ? Are data consistent ?

Data Preprocessing: Select data, Clean data, Transfer data, Integrate data
Select data: Observation reduction, attribute reduction
Observation reduction: Sampling

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting

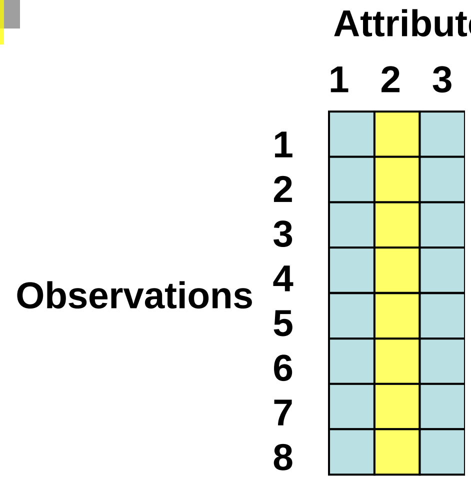


Observation Reduction

- Sampling
- Intelligent Sampling
- Learn to forget

....

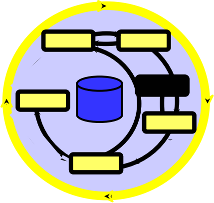
Attribute Reduction



Data Mining Process

CRISP-DM: Data Preparation

Data Selecting



Observation Reduction : Sampling

Statisticians: Sampling because ***obtaining*** the entire dataset (population) is too expensive or time consuming (often they *do not have* the data and start collecting)

Data Miners: Sampling because ***processing*** of the population is too expensive or time consuming (often they *have* the data)

good sample ~ representative sample

↓

has nearly the same property as the population : →

- sample **mean** is very close to population mean
- sample **variance** is very close to population variance
- ....

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting

Observation Reduction : Sampling

Task: **Choose a sampling method that with high probability leads to a representative sample**

- **Choosing the right sampling technique**
- **Choosing the right sample size**

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting

Observation Reduction : Sampling technique

Random sampling: Equal and known probability of being selected for each member of the population

General aspects:

- Sampling without replacement (s.wo.r)
- Sampling with replacement (s.w.r.)

→ During the sampling process the probability of selecting any objects remains constant

→ Analyzing is easier

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting

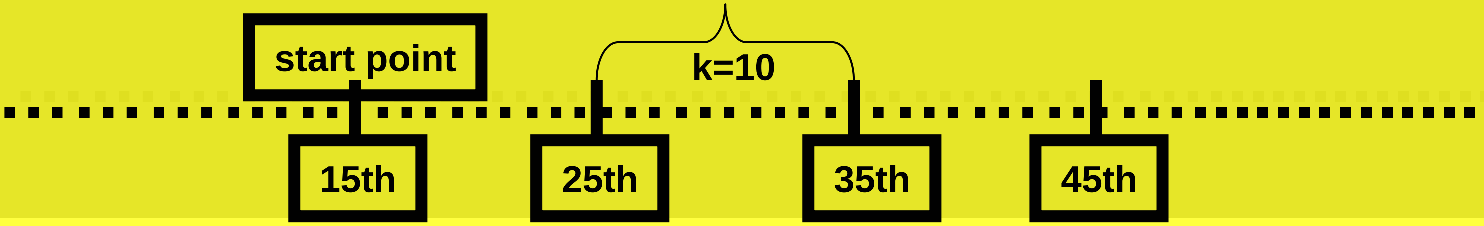
Observation Reduction : Sampling technique

Systematic Sampling (called also kth name selection method)

- Selection of k; k= population size / sample size (k sampling interval)
- Selection of a start point
- Selection of every kth member as sample

Example: Population size = 2000 sample size = 200

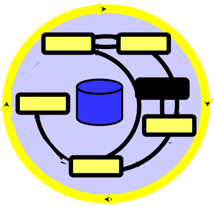
- k=10
- start point = member number 15
- then sample consists of members number 15, 25, 35, 45,…



Data Mining Process

CRISP-DM: Data Preparation

Data Selecting



Observation Reduction : Sampling technique

Stratified Sampling

Population consists of different mutually exclusive subgroups (strata) varying considerably in size.
Examples: (120 men, 30 women), (1900 employment, 100 unemployment), (300 white, 20 black)

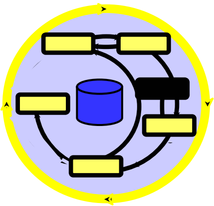
Random sampling can fail to adequately represent the members with low frequency

Solution: Stratified Sampling: Random sampling in each Subgroup (stratum) independently

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting



Observation Reduction : Sampling technique

Stratified Sampling Strategies

Stratified sampling strategies

1. **Number of members drawn from each subgroupa is proportional to the size of that subgroup**

2. **Equal numbers of members are drawn from each subgroup even though the gropus are of different sizes**

Example: Size of population 2000: 1900 employment, 100 unemployment
Size of needed sample: 50

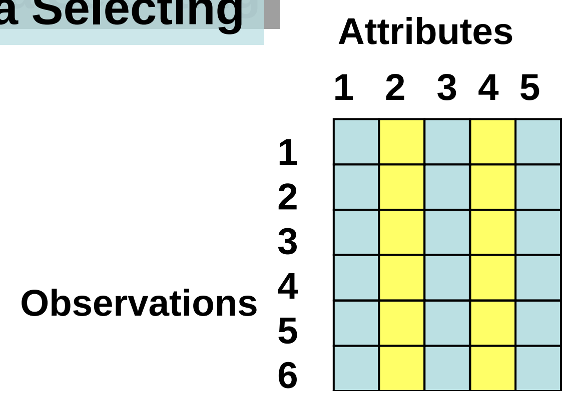
Strategy 1 : 50/2000 = 1/40 1900 * 1/40 = 47,5 100 * 1/40 = 2,5
Sample consists of 47 employment and 3 unemployment

Strategy 2 : Sample consists of 25 employment and 25 unemployment

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting

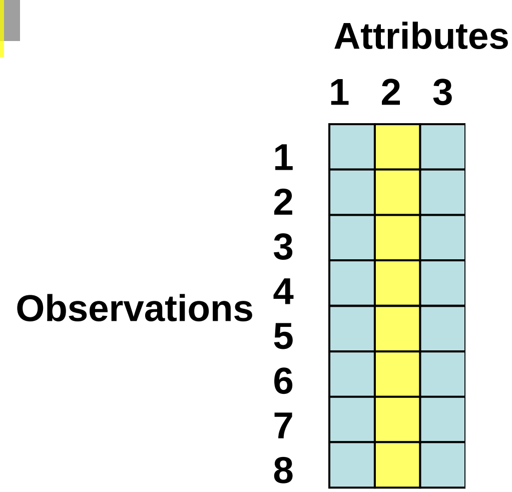


Observation Reduction

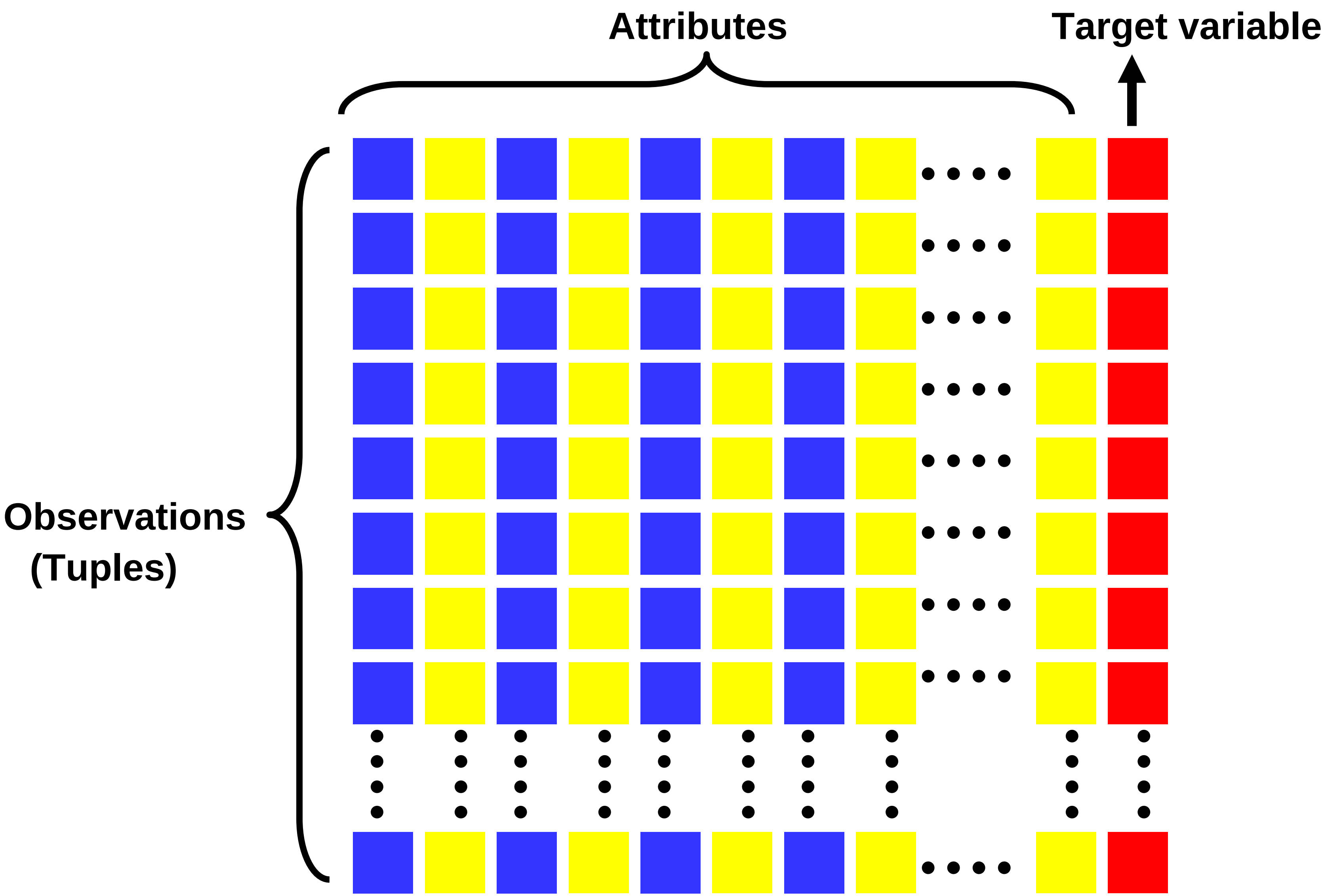
- Sampling
- Intelligent Sampling
- Learn to forget

....

Attribute Reduction



Supervised and unsupervised learning



Supervised Learning

Nr.	A1	A2	A3………		An	T
1	a11	a12	a13		a1n	t1
2	a21	a22	a23		a2n	t2
3	a31	a32	a33		a3n	t3
.						
.						
.	.. ….	….	….	…..	….	
.						
.						
.						
m	am1	am2	am3		amn	tm

Examples for Supervised Learning : Classification, Prediction

Unsupervised Learning

Nr.	A1	A2	A3	………	An
1	a11	a12	a13		a1n
2	a21	a22	a23		a2n
3	a31	a32	a33		a3n
.					
.					
.	.. ….	….	….	…..	….
.					
.					
.					
m	am1	am2	am3		amn

Example for Unsupervised Learning: Clustering

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting

Attribute Reduction General Aspects

Data mining problems that deal with classification and prediction may involve hundreds or even thousands of attributes that can potentially be used as predictors Example: Document classification in Text Mining: *Bag-of-words*: >100000 attributes , fault analysis in the automotive industry,…

Problem: A lot of time and effort may be needed to decide which attribute should be included in the model

Solution: In the last years Statisticians and Data Miners have developed many attribute reduction algorithms

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting

Why we need attribute Reduction ?

- to reduce the effect of the *curse of dimensionality*
- to speed up learning process
- to reduce the amount of memory required
- to improve model interpretability
- to do visualization easier
- to make scalable the datasets with many nominal attributes

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting

Attribute Reduction

curse of dimensionality

↓

As the dimensionality of data increases
often data analysis become harder

classification → reduced classification accuracy

clustering → Poor quality cluster

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting

Attribute Reduction

creating new attributes
(combination of old attribute)
attribute extraction

Selection a subset of old attributes
FSS: feature subset selection
attribute selection

Loss of information ?

no information lost if redundant and irrelevant attributes are present

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting

Attribute Reduction

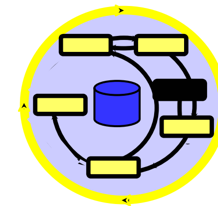
First elementary steps

- Using common sense or domain Knowledge (if available) to select a subset of attributes
- Attribute Screening

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting



Attribute Reduction

First elementary steps

- **Attribute Screening**

removes problematic attributes e.g:

- **attributes with many missing values**
- **attributes with values that have too much or too little variation**

Example
Income of 100 individuals = { 20, 20, 20, 20, ………….20, 20 }

Attribute income is not informative

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting

Attribute Reduction **Attribute Ranking**

Determining attribute importance by criteria like:

- **Information Gain**
- **Gini-Index**
- **Pearson Chi-Square**
- **Correlation coefficient**
- **Akaike information criterion (AIC)**
- **….**

Data Mining Process

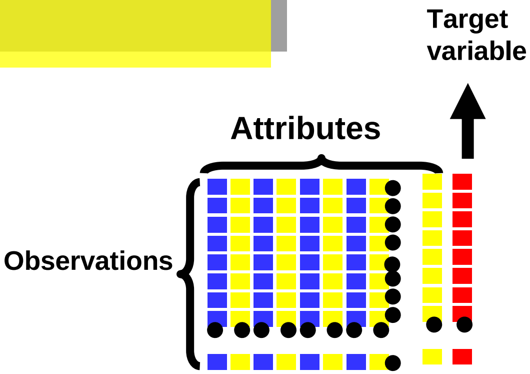
CRISP-DM: Data Preparation

Data Selecting

Attribute Reduction | Attribute Ranking

The ranking criteria mentioned before can be used to measure the correlation between

1. each attribute and the target variable (applicable only to Supervised Learning
2. between two attributes, pairwise



Remarks

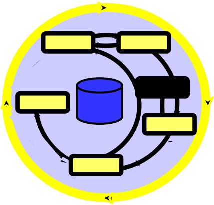
- In case 1, an attribute useless by itself can be useful together with others
- In case 2 attribute selection is independent of the target variable or , generally, independent of the data mining task

Known as Filter Approach

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting



Attribute Reduction

Embedded Methods

- in embedded approaches attribute selection is a part of the training process
- not all Data Mining algorithms have this built-in mechanism to perform attribute selection within the training process
- due to avoiding retraining for different attribute subsets , embedded approaches are more efficient
- Examples: Decision and Regression Trees

Remark

- in some studies, in a first step simple linear embedded systems are use for attribute selection
- later in a second step, the selected attributed are used for training of a more complicated non-linear system

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting

Attribute Reduction

Wrapper Methods

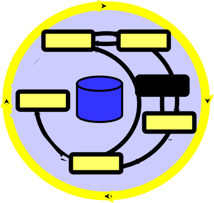
Main Idea :

- Using a given classification or prediction algorithm, evaluate the prediction performance of different subsets of attributes
- Select the subset with highest performance

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting



Attribute Reduction

Wrapper Methods

Main Challenges :

1. **Selecting a search method to find all possible attribute subsets**
2. **Selecting an evaluation approach and an evaluation function to compare the prediction performance of different attribute subsets**

About 1: Total search in the case of too large number of attributes needs massive amounts of computation. Greedy search like forward selection and backward elimination are more appropriate

About 2 : Validation datasets or cross validation as well as evaluation functions (e.g accuracy rate or mean squared error) can be used

Data Mining Process

CRISP-DM: Data Preparation

Data Selecting

Principal Component Analysis (PCA)

Main Idea

Reducing multidimensional data sets to lower dimensions by combination of old attributes

the variance of the observations in original space should be satisfactory covered by the new created dimensions

$b_1 = p_1\, a_1 + p_2\, a_2$

$b_2 = q_1\, a_1 + q_2\, a_2$

Instruments:

- Covariance Matrix
- Eigenvalues
- Eigenvectors

Interpretation ?