

Business Intelligence and Data Management

academic year 2019-2020

Dr. Ekaterini Ioannou, Department of Management, University of Tilburg

Lecture 4

topic: Introduction to Data Mining

material:

- Chapters 1, 2, 3 (book “Data Mining for Business Intelligence”)
- Wirth, R and Hipp, J. *CRISP-DM: towards a standard process model for data mining*. In KDD 2000.

Agenda for remaining weeks

Date		Lecture contents	Lecturer	Lab topics	Test
Jan-27	1	Intro. to BI+ Data Management	Caron		
Jan-30				SQL-1	1
Jan-28	2	Data warehousing	Caron		
Feb-06				SQL-2	2
Feb-03	3	OLAP business databases & dashboard	Caron		
Feb-13				SQL-3 & OLAP	3a & 3b
Feb-10	4	Data mining introduction	Ioannou		
	5	Regression models	Ioannou		
Feb-17	6	Naïve Bayes	Ioannou		
	7	k nearest neighbors	Ioannou		
Feb-20				Bayes & neighbors	4
Feb-27	8	Performance measures	Ioannou		
Mar-02	9	Decision trees	Ioannou		
Mar-05				Dec. trees	5
Mar-09	10	Association rules	Ioannou		
Mar-11,12&13				Ass. Rules	6
Mar-16	11	Clustering (+20 mins exam preparation)	Ioannou		
Mar-19				Clustering	7

Data Mining

Business Intelligence

Definitions discussed in previous lectures:

[Sharda'14] An umbrella term that *combines processes, technologies, and tools* needed to transform

- data into information,
- information into knowledge, and
- knowledge into plans that drive profitable business action

[Sabherwal'11] *Information and knowledge* that enables business decision-making

[Shmueli et al.'19] Data visualization and reporting for understanding “what happened and what is happening”

Data Mining

- Creative process that provides results useful for decision making
- Requires several different skills, e.g.,
 - Statistics
 - Machine-learning
 - Programming
 - Etc.
- Ability to cope with huge amounts of data
→ used frequently for Big Data

Big Data

- Current data is big (by reference to the past)
- Challenges of Big Data is typically characterized by:
 - **Volume**: amount of data
 - **Velocity**: flow rate, i.e., speed at which data is being generated and changed
 - **Variety**: different types of data being generated, i.e., currency, dates, numbers, text, etc.
 - **Veracity**: data is being generated by organic distributed processes, i.e., quality (missing, clean value)
 - **Value**: data have no value for the company unless you turn it into something useful

Methods

- Classification and Prediction
 - Regression
 - k-nearest neighbors
 - Decision trees
- Association
 - Association rules
- Clustering
 - k-means

Topics not covered

- Type
 - Document management
 - Text mining
 - Web mining
- Technology
 - Map Reduce
 - Hadoop
 - Deep Learning
- Some of the above topics are covered in the course
“Business Analytics Emerging Trends”

Why are there so many
different data mining
methods?

Existence of various methods

- Each method has advantages and disadvantages
- These depend on several factors
- Factor examples:
 - Size of the dataset
 - Types of patterns that exist in the data
 - If the data meet some underlying assumptions of the method
 - How noisy is the given data
 - The goal of the analysis

Data

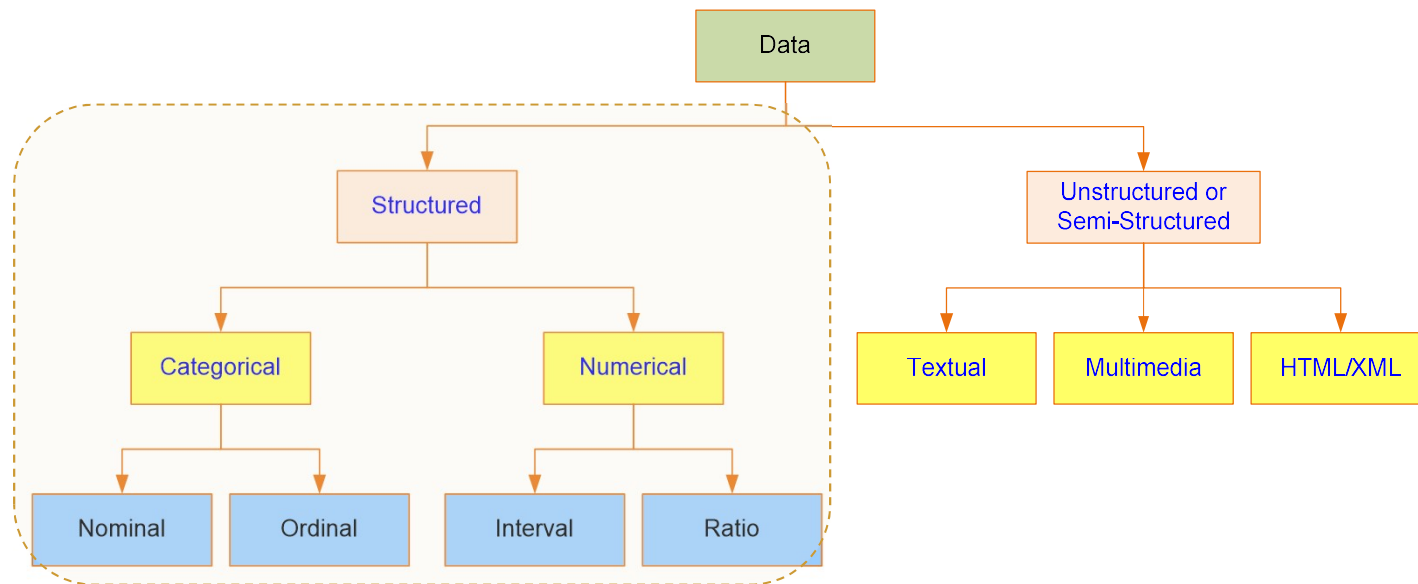
- A **collection of facts** usually obtained as the result of experiences, web page visits, observations, or experiments, extracted from documents
- **Lowest level of abstraction** (from which information and knowledge are derived)
- May consist of numbers, words, images, etc.

Terminology

- **Observations:** the unit of analysis on which the measurements are taken (a customer, a transaction, etc.)
 - Also called instance, sample, example, case, **record**, pattern, or row
 - In spreadsheets:
 - each row typically represents a **record** and
 - each column typically represents a **variable**
- **Algorithm:** a specific procedure used to implement a specific data mining technique
- **Variable:** any measurement on the record values
 - We can have **input variables** and **output variable**

Data & Types of Variables

- Data may consist of numbers, words, images, etc.
- Classification:



Nominal Data

- Values are **distinct symbols**
 - Values themselves serve only as labels or names
 - Nominal comes from the Latin word for name
- Example: attribute “outlook” from weather data
 - Values: “sunny”, “overcast”, and “rainy”
- No relation is implied among nominal values
 - There is no ordering or distance measure
- Only equality tests can be performed

Ordinal

- Impose **order on values**
- But: no distance between values defined
- Example attribute “temperature” in weather data
 - Values: “hot” > “mild” > “cool”
- Note: addition and subtraction don’t make sense
- Example rule *temperature* < “hot” \rightarrow *play* = “yes”
- Distinction between nominal and ordinal not always clear (e.g., attribute “outlook”)

Interval





- Interval quantities are not only **ordered**
But measured in **fixed and equal units**
- Example 1:
 - Attribute “temperature” when this is expressed in degrees Fahrenheit
- Example 2:
 - Attribute “year”
- Difference of two values makes sense
- Sum or product doesn't make sense
 - Zero point is not defined!

Radio

- Quantities for which the measurement scheme defines a zero point
- Example: attribute “distance”
 - Distance between an object and itself is zero
- Ratio quantities are treated as real numbers
 - All mathematical operations are allowed
- In our example:
 - Three times a distance makes sense
 - Multiplying two distances gives an area

Quiz

attributes:

	A. color	B. flying-altitude	C. mood
	red	10	very unhappy
	blue	11	happy
	white	12	unhappy
	black	15	normal

What are the types of the attributes, i.e., color, flying-altitude, mood?

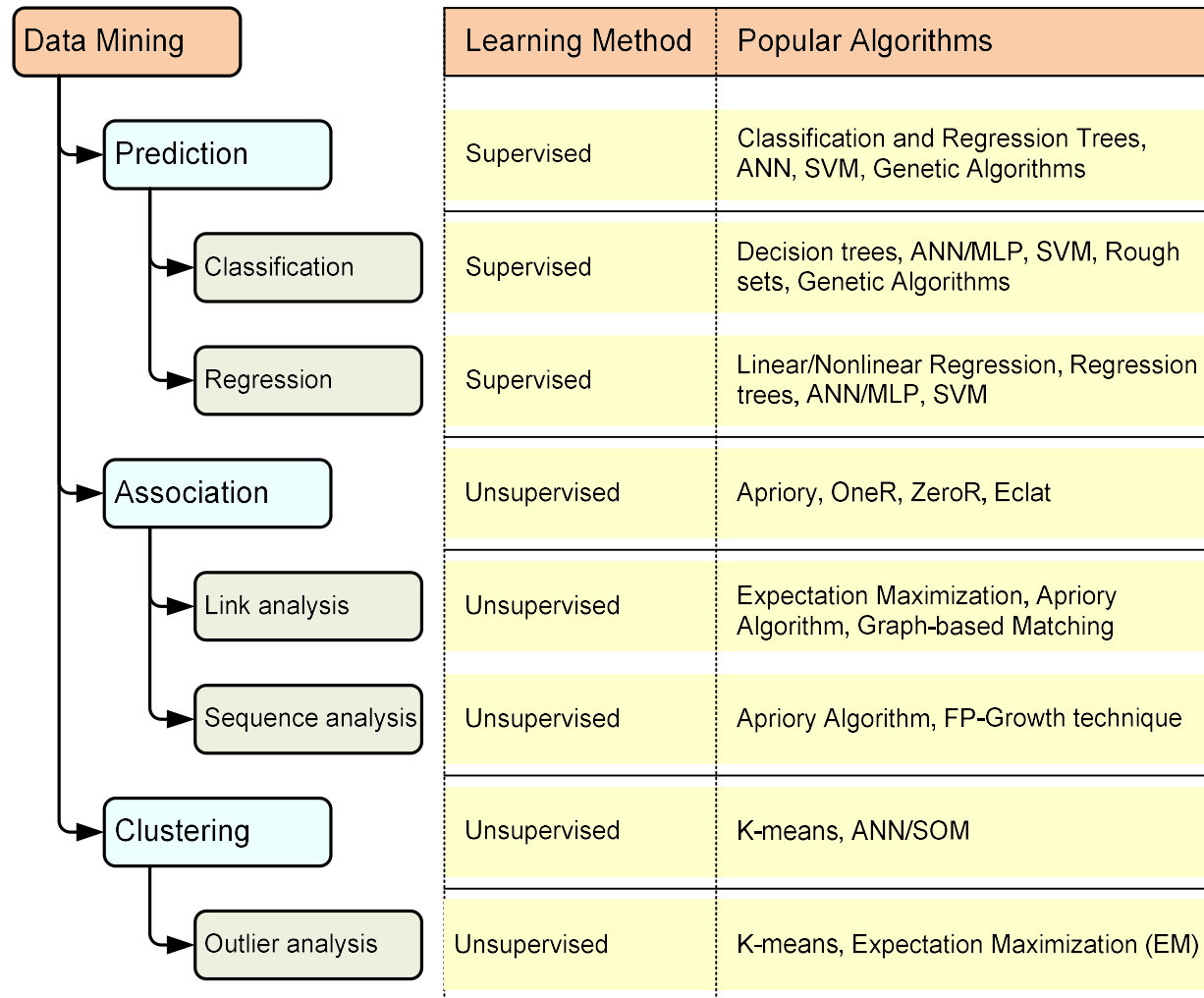


1. Nominal, Interval, Ordinary.
2. Ordinary, Ratio, Nominal.
3. Nominal, Ratio, Ordinary.
4. Ordinary, Interval, Nominal.

Sources for the data

- Internal
 - Transactions
 - Mailing lists
 - Intranet
- External
 - Claims other companies
 - Demographic
 - Internet

A Taxonomy for Data Mining Tasks



Data Mining Process(es)

I.e., overview of the steps
involved in data mining

Steps in a typical data mining effort

- Course focuses on the understanding and using data mining algorithms, however
- Some of the most serious errors in analytics projects result from a poor understanding of the problem (i.e., before we get into the algorithms)

STEP-1: Develop an understanding of the purpose of the data mining project

- How will the stakeholder use the results?
- Who will be affected by the results?
- Will the analysis be a one-shot effort or an ongoing procedure?

STEP-2: Obtain the dataset to be used in the analysis

- Often involves sampling from a large database to capture records to be used in an analysis
- Could also involve pulling together data from different databases or sources

STEP-3: Explore, clean, and preprocess the data

- Verify that the data are in reasonable condition
- How should missing data be handled?
- Are the values in a reasonable range, given what you would expect for each variable?
- Are there obvious outliers?

Define Purpose	Obtain Data	Explore & Clean	Dimensions	Determine DM Task
----------------	-------------	-----------------	------------	-------------------

STEP-4: Reduce the data dimension, if necessary

- Ensure you know what each variable means
- Also if it is sensible to include it in the model
- Transform variables, eliminate unneeded variables, creating new variables

STEP-5: Determine the data mining task

- Translate the general question or problem of STEP-1 into a more specific data mining question

Define Purpose	Obtain Data	Explore & Clean	Dimensions	Determine DM Task	Partition	Choose Methods	Apply & Select
----------------	-------------	-----------------	------------	-------------------	-----------	----------------	----------------

STEP-6: Partition the data (for supervised tasks)

- Randomly partition the dataset into three parts: (i) training, (ii) validation, and (iii) test datasets

STEP-7: Choose the data mining technique(s)

STEP-8: Use algorithms to perform the task

- Iterative process, i.e., trying multiple variants of the algorithm, including different variables or settings within the algorithm
- Use feedback (if appropriate) from performance to refine the settings

Define Purpose	Obtain Data	Explore & Clean	Dimensions	Determine DM Task	Partition	Choose Methods	Apply & Select	Interpret	Deploy
----------------	-------------	-----------------	------------	-------------------	-----------	----------------	----------------	-----------	--------

STEP-9: Interpret the results of the algorithms

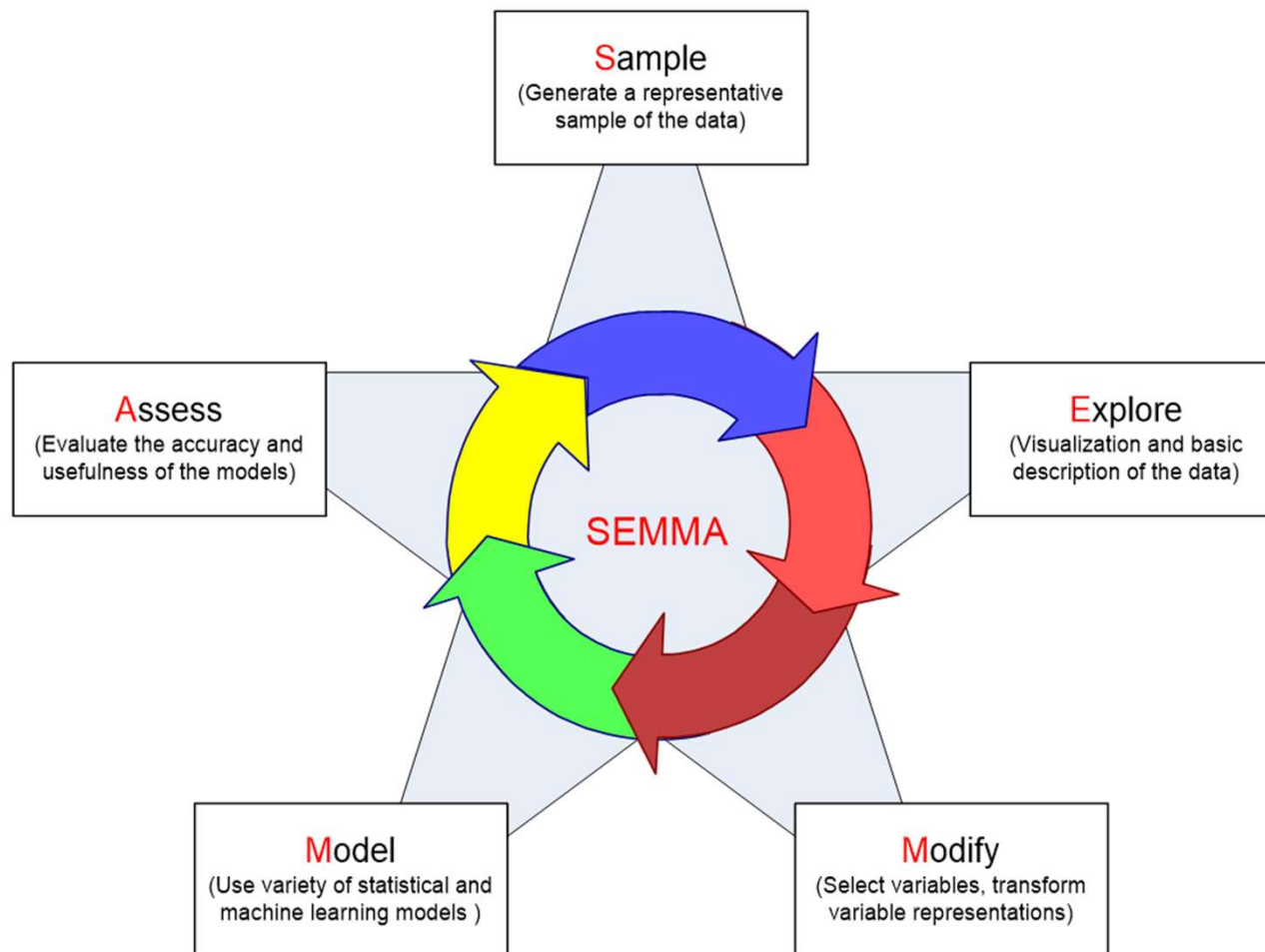
- Use the training, validation, and test datasets
- Make a choice as to the best algorithm to deploy
- Evaluate this on the test data to get an idea as to how well it will perform

STEP-10: Deploy the model

- Integrate the model into operational systems
- Running it on real records to produce decisions or actions

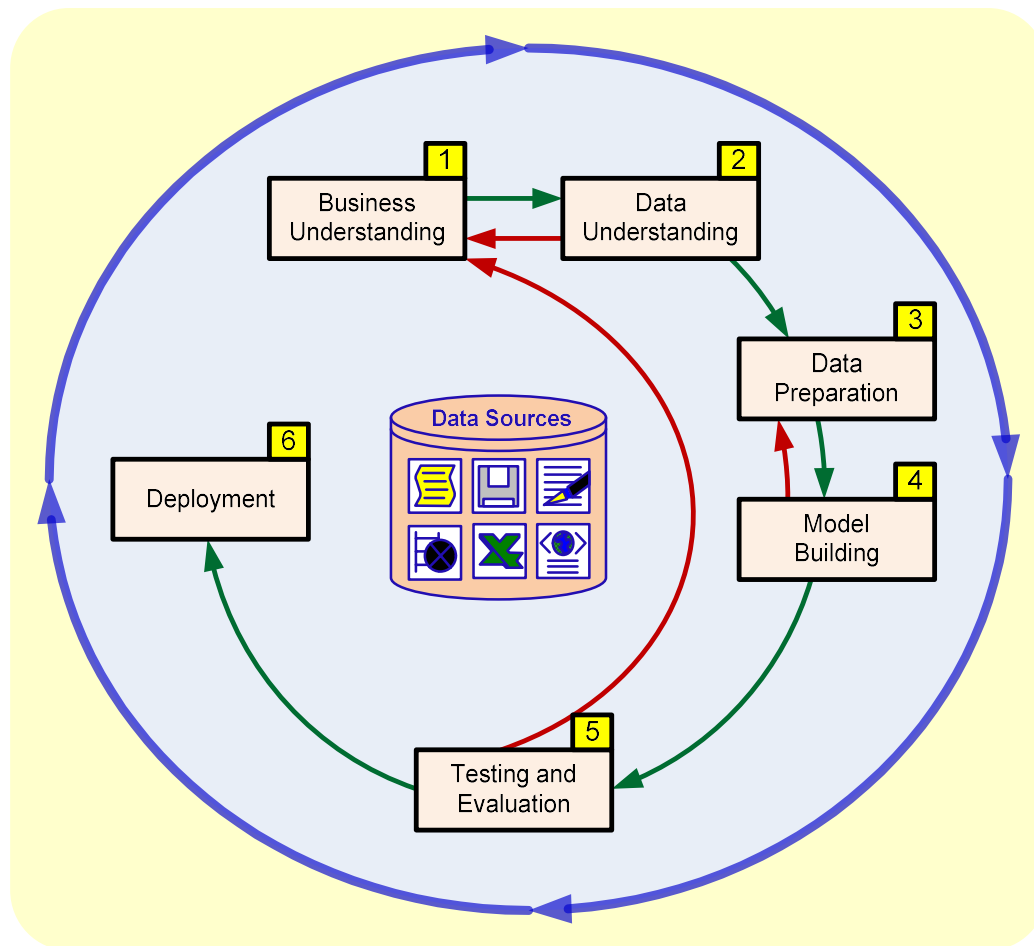
SEMMA methodology

- Developed by the software company SAS
- Include the previous steps



CRISP-DM

- Similar methodology by IBM SPSS Modeler
- **C**Ross **I**ndustry **S**tandard **P**rocess for **D**ata **M**ining



generic CRISP-DM
reference model

CRISP-DM

Step 1: Business Understanding

Step 2: Data Understanding

Step 3: Data Preparation (!)

Step 4: Model Building

Step 5: Testing and Evaluation

Step 6: Deployment

Accounts for
~85% of total
project time

SEMMA

- The process is highly repetitive and experimental (DM: art versus science?)





Data Mining:

- Provides results useful for decision making
- Data types:
 - Nominal → no order, ranking, or sequence (e.g., gender)
 - Ordinary → ordered series of data (e.g., performance)
 - Interval → no zero but sum and difference are meaningful
 - Ratio → interval plus meaningful zero
- Data Mining process involves multiple steps:
 - Business understanding → Data preparation → Model building → Testing & Evaluation → Deployment

Summary for the following lectures

Quiz

- Nominal → no order, ranking, or sequence (e.g., gender)
- Ordinary → ordered series of data (e.g., performance)
- Interval → no zero but sum and difference are meaningful (e.g., temperature in Celsius)
- Ratio → interval plus meaningful zero

	attributes		
	color	flying-altitude	mood
	nominal	ratio	ordinary
	red	10	very unhappy
	blue	11	happy
	white	12	unhappy
	black	15	normal