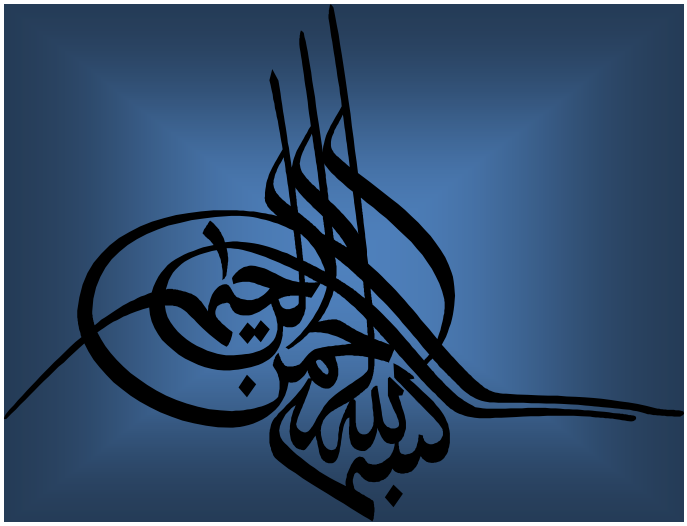


1

داده کاوی

Data Mining

Lecturer: A. Rabiee
azrabiee@gmail.com
Rabiee.iauda.ac.ir

۲

منابع و مراجع

Main Reference:

-Larose, Daniel T. *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons, 2014.

Other References:

- Han, Jiawei, Micheline Kamber, and Jian Pei. *Data mining: concepts and techniques*. Elsevier, 2011.
- Tan, Pang-Ning, Michael Steinbach, and Vipin Kumar. *Introduction to data mining*. Vol. 1. Boston: Pearson Addison Wesley, 2006.
- …

مراجع فارسی

- دکتر مهدی اسماعیلی، مفاهیم و تکنیک‌های داده‌کاوی
- ژیاوی هان، میشلین کامبر و ژان پی، داده کاوی، ترجمه دکتر نسترن حاجی حیدری و بهنام خاکباز
- دکتر محمدحسین ندیمی، مریم تاکی و فاطمه حبیب اللهی، داده کاوی، مفاهیم و کاربردها
- …

Larose's book, table of content

- Chapter 1: Introduction
- Chapter 2: Data Preprocessing
- Chapter 3: Exploratory Data Analysis
- Chapter 4: Statistical Approaches
- Chapter 5: k-Nearest Neighbor algorithm
- Chapter 6: Decision trees
- Chapter 7: Neural Networks
- Chapter 8: Hierarchical k-means Clustering
- Chapter 9: Kohonen Networks
- Chapter 10: Association Rules
- Chapter 11: Model Evaluation Techniques

ارزشیابی درس

Final Exam:	50
Assignments:	20
Final Project + Presentation:	30
Paper (optional):	+15

۴

Chapter 1:

Introduction to Data Mining

6

Table of Contents

7

Definition

- Data mining
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
- Alternative names
  - Knowledge discovery (mining) in data (KDD),
  - knowledge extraction,
  - data/pattern analysis
  - …

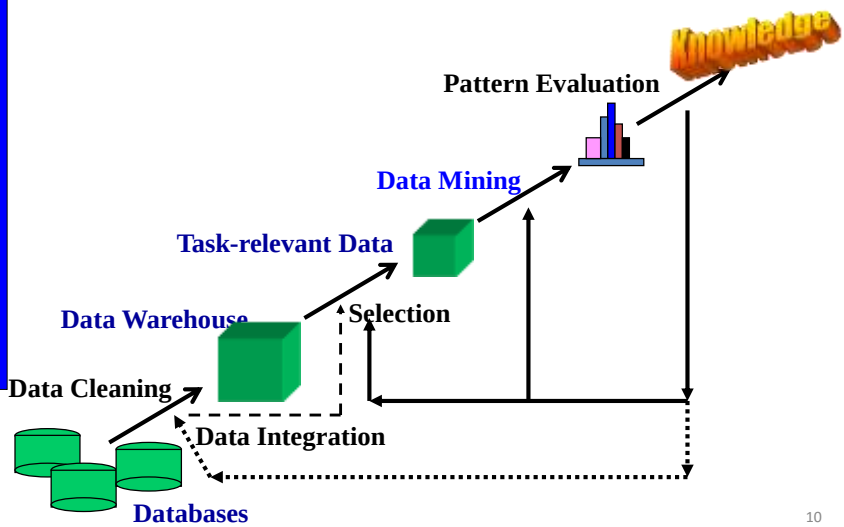
8

Data Mining

- "one of the most revolutionary developments of the next decade," according to the online technology magazine *ZDNET News (2001)*
- the *MIT Technology Review* (2001) chose data mining as one of 10 emerging technologies that will change the world.

9

Data Mining:
Core of knowledge discovery process



10

What is (not) Data Mining?

What is not Data Mining?

– Look up phone number in phone directory

– googling

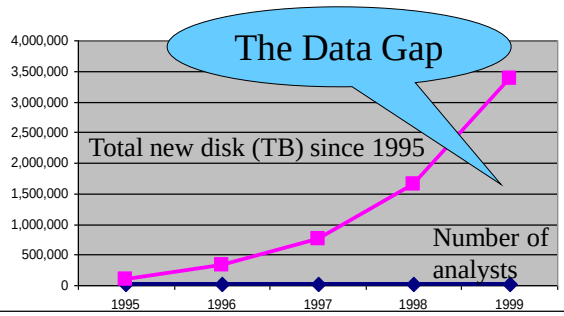
What is Data Mining?

– Certain names are more prevalent in certain country locations

– Group together similar documents, according to their context

Mining Large Data Sets - Motivation

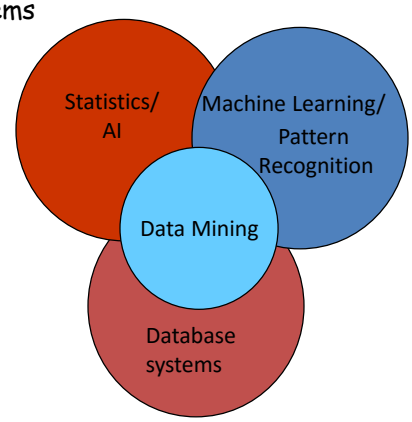
- There is often information "hidden" in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all



From: R. Grossman, C. Kamath, V. Kumar, "Data Mining for Scientific and Engineering Applications"

Origin of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional Techniques may be unsuitable due to
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data



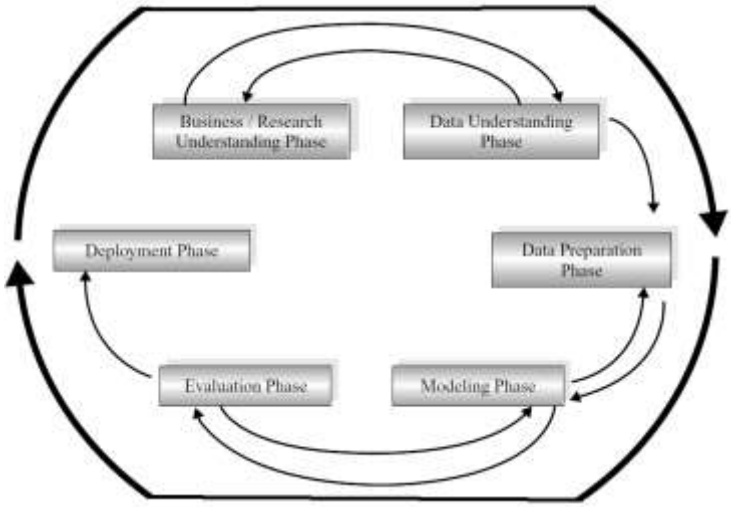
13

CRISP–DM
CRoss-Industry Standard Process

- A standard process for fitting data mining into the general problem-solving strategy of a business or research unit
- 6 phases

14

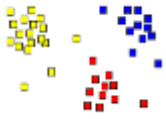
CRISP–DM
CRoss-Industry Standard Process



15

Data Mining Tasks

- Description
- Estimation
- Prediction
- Classification
- Clustering
- Association



Market Basket Analysis

16

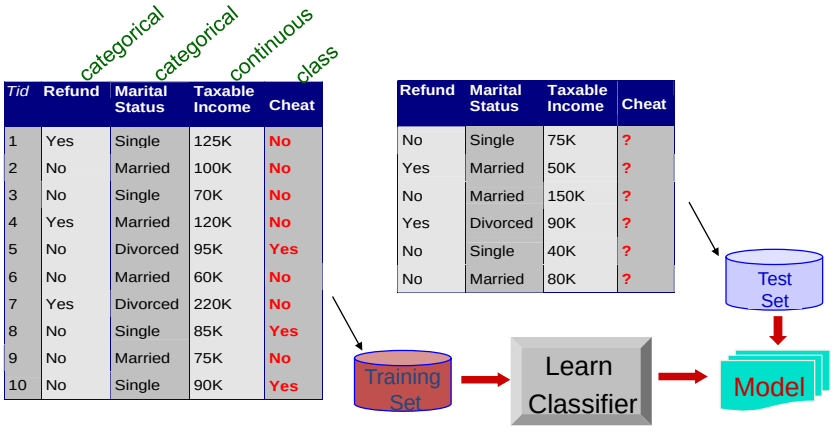
Description

- DM can be used for describing patterns and trends lying within data
  - Example: finding a pattern for distribution of a virus

=> Exploratory data analysis (EDA), chapter 3

17

Classification



18

Classification

- *k*-nearest neighbor (Chapter 5)
- Decision tree (Chapter 6)
- Neural network (Chapter 7)

Income classification

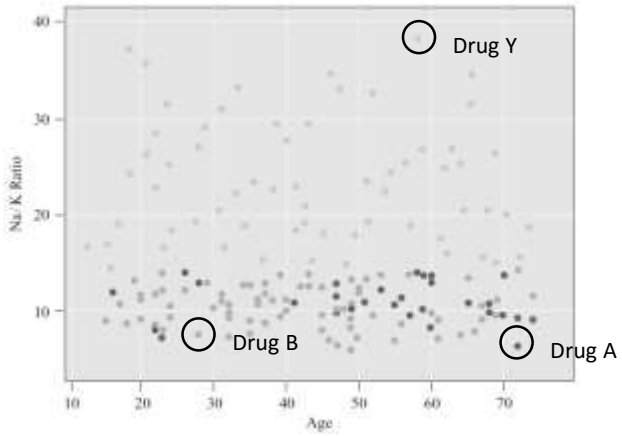
Subject	Age	Gender	Occupation	Income Bracket
001	47	F	Software engineer	High
002	28	M	Marketing consultant	Middle
003	35	M	Unemployed	Low
⋮				

Graphs and plots are helpful for understanding two- and three-dimensional relationships in data.

19

Classification

- Which drug should be prescribed for which type of patient?



20

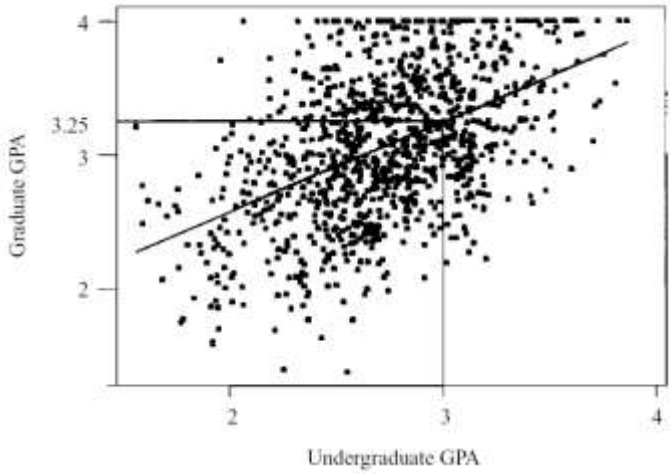
Classification examples

- Determining whether a particular credit card transaction is fraudulent
- Placing a new student into a particular track with regard to special needs
- Assessing whether a mortgage application is a good or bad credit risk
- Diagnosing whether a particular disease is present
- Determining whether a will was written by the actual deceased, or fraudulently by someone else
- Identifying whether or not certain financial or personal behavior indicates a possible terrorist threat
- …

21

Estimation

Example: Estimating the grade-point average (GPA) of a graduate student, based on that student's undergraduate GPA



regression

22

Estimation Examples

- Estimating the amount of money a randomly chosen family of four will spend for back-to-school shopping this fall.
- Estimating the number of points per game that Patrick Ewing will score.
- Estimating the grade-point average (GPA) of a graduate student, based on that student's undergraduate GPA.
- …

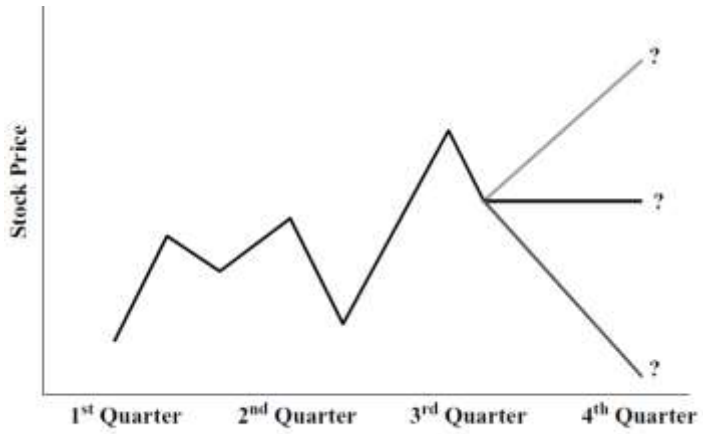
23

Difference of Estimation and Classification?

Numeric Categorical

24

Prediction



Predicting the price of a stock three months in the future.

25

Prediction

- Any of the methods and techniques used for classification and estimation may also be used for prediction.
  - simple linear regression and correlation, and multiple regression (Chapter 4)
  - *k*-nearest neighbor (Chapter 5)
  - Decision tree (Chapter 6)
  - Neural network (Chapter 7)
  - …

26

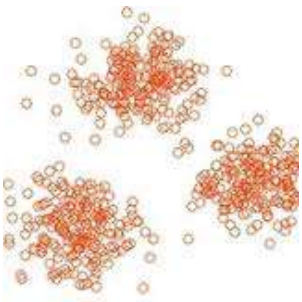
Prediction Examples

- Predicting the price of a stock three months into the future
- Predicting the percentage increase in traffic deaths next year if the speed limit is increased
- Predicting the winner of this fall's baseball World Series, based on a comparison of team statistics
- Predicting whether a particular molecule in drug discovery will lead to a profitable new drug for a pharmaceutical company
- …

27

Clustering

- In contrast to the classification, it's unsupervised
- Maximizing the similarity within the clusters, minimizing the similarity between clusters
- *k*-means clustering (Chapter 8)
- Kohonen networks (Chapter 9)



28

Clustering Example

- Grouping the clothes in a store based on the size
- Segmentation of financial behaviors into safe and suspicious categories
- As a dimension-reduction tool when the data set has hundreds of attributes
- For gene expression clustering, where very large quantities of genes may exhibit similar behavior
- …

29

Association

- A priori & GRI algorithms (Chapter 10)

market basket analysis

TID	Items
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Rules Discovered:
{Milk} --> {Coke}
{Diaper, Milk} --> {Beer}

30

Association examples

- Investigating the proportion of subscribers to a company's cell phone plan that respond positively to an offer of a service upgrade
- Examining the proportion of children whose parents read to them who are themselves good readers
- Predicting degradation in telecommunications networks
- Finding out which items in a supermarket are purchased together and which items are never purchased together
- Determining the proportion of cases in which a new drug will exhibit dangerous side effects
- …

31

Case Studies

- Case Study 1: Analyzing automobile warranty claims
- Case Study 2: Predicting abnormal stock market
- Case Study 3: Mining association rules
- Case Study 4: Predicting corporate bankruptcies
- Case Study 5: Profiling the tourism market

32

Hint!

- Data mining is easy to do badly
  - Because of the easy-to-use GUI-based tools

=> White-box approaches instead of the black-box or blind methods

33

Homework 1: Analyzing a paper

- Find a data mining paper (preferably from 2010 to 2016)
- Discuss the 6 steps of CRISP-DM on the paper

35