Data Mining

Prof. Dr. Nizamettin AYDIN

naydin@yildiz.edu.tr

http://www3.yildiz.edu.tr/~naydin

1

1

Course Details

- Course Code: BLM5224
- Course Name: Data Mining (Veri Madenciliği)
- Credit: 3
- Nature of the course: Lecture
- Course web page:
  http://www3.yildiz.edu.tr/~naydin/na_DMi.htm
- Instructors: Nizamettin AYDIN

Email: naydin@yildiz.edu.tr

2

2

Rules of the Conduct

- No eating /drinking in class
  - *except water*
- Cell phones must be kept outside of class or switched-off during class
  - *If your cell-phone rings during class or you use it in any way, you will be asked to leave and counted as unexcused absent.*
- No web surfing and/or unrelated use of computers,
  - when computers are used in class or lab.

3

3

Rules of the Conduct

- You are responsible for checking the class web page often for announcements.
  - http://www3.yildiz.edu.tr/~naydin/na_ SdA.htm
- Academic dishonesty and cheating
  - will not be tolerated
  - will be dealt with according to university rules and regulations
    - http://www.yok.gov.tr/content/view/475/
    - Presenting any work that does not belong to you is also considered academic dishonesty.

4

4

Attendance Policy

- The requirement for attendance is **70%**
  - **Hospital reports** are **not** accepted to fulfill the requirement for attendance.
  - **The students, who fail to fulfill the attendance requirement, will be excluded from the final exams and the grade of F0 will be given.**

5

5

Assesment

• Quiz	:	10%
• Midterm	:	25%
• Homework	:	20%
• Final	:	40%
• Attendance & participation	:	05%

(The requirement for attendance is 70%)

6

6

Some Recommended Books

- Introduction to Data Mining, Tan, Steinbach & Kumar
- Data Mining: The Textbook, Charu C. Aggarwal
- Data Mining- Concepts, Models, Methods, and Algorithms, Mehmed Kantardzic
- Principles of Data Mining, Max Bramer
- Data Mining Techniques, Michael Berry and Gordon Linoff
- Introduction to Algorithms for Data Mining and Machine Learning, Xin-She Yang

7

7

Introduction

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data?
- Data mining functionality
- Are all the patterns interesting?
- Classification of data mining systems
- Data Mining Task Primitives
- Integration of data mining system with a DB and DW System
- Major issues in data mining

8

8

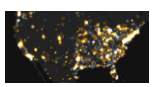
Large-scale Data is Everywhere!

- There has been enormous data growth in both commercial and scientific databases due to advances in data generation and collection technologies
- New mantra
  - Gather whatever data you can whenever and wherever possible.
- Expectations
  - Gathered data will have value either for the purpose collected or for a purpose not envisioned.

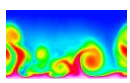

Cyber Security


E-Commerce


Traffic Patterns


Social Networking: Twitter


Sensor Networks


Computational Simulations

9

9

Why Data Mining?

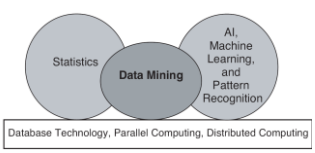
- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, …
    - Science: Remote sensing, bioinformatics, scientific simulation, …
    - Society and everyone: news, digital cameras,
- We are drowning in data but starving for knowledge!
- "Necessity is the mother of invention"—Data mining—Automated analysis of massive data sets

10

10

Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems
- Traditional techniques may be unsuitable due to data that is
  - Large-scale
  - High dimensional
  - Heterogeneous
  - Complex
  - Distributed



- A key component of the emerging field of data science and data-driven discovery

11

11

Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS
- 1970s:
  - Relational data model, relational DBMS implementation
- 1980s:
  - RDBMS, advanced data models (extended-relational, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)
- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases
- 2000s
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems

12

12

Why Data Mining? Commercial Viewpoint

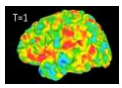
- Lots of data is being collected and warehoused
  - Web data
    - Google has Peta Bytes of web data
    - Facebook has billions of active users
  - purchases at department/ grocery stores, e-commerce
  - Amazon handles millions of visits/day
  - Bank/Credit Card transactions
- Computers have become cheaper and more powerful
- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)

13

13

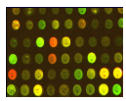
Why Data Mining? Scientific Viewpoint

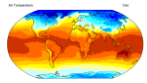
- Data collected and stored at enormous speeds
  - remote sensors on a satellite
    - NASA EOSDIS archives over petabytes of earth science data / year
  - telescopes scanning the skies
    - Sky survey data
  - High-throughput biological data
  - scientific simulations
    - terabytes of data generated in a few hours
- Data mining helps scientists
  - in automated analysis of massive datasets
  - in hypothesis formation



fMRI Data from Brain


Sky Survey Data


Gene Expression Data


Surface Temperature of Earth

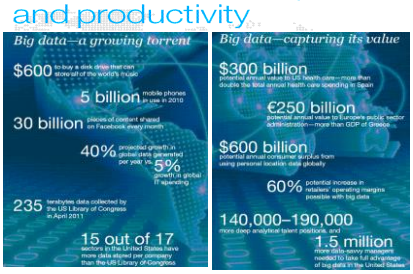
14

14

Great opportunities to improve productivity in all walks of life

McKinsey Global Institute

Big data: The next frontier for innovation, competition, and productivity

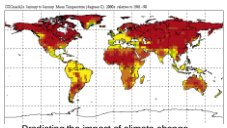


15

15

Great Opportunities to Solve Society's Major Problems


Improving health care and reducing costs


Predicting the impact of climate change


Finding alternative/ green energy sources


Reducing hunger and poverty by increasing agriculture production

16

16

What Is Data Mining?

- Data mining (knowledge discovery from data)
  - 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 databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.
- Watch out: Is everything "data mining"?
  - Simple search and query processing
  - (Deductive) expert systems

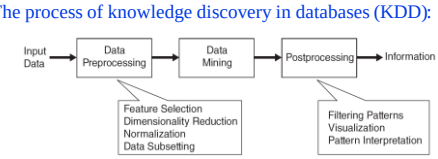
17

17

What is Data Mining?

- Many Definitions
  - Non-trivial extraction of implicit, previously unknown and potentially useful information from data
  - Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns

The process of knowledge discovery in databases (KDD):



18

18

Potential Applications

- Data analysis and decision support
  - Market analysis and management
    - Target marketing, customer relationship management (CRM), market basket analysis, cross selling, market segmentation
  - Risk analysis and management
    - Forecasting, customer retention, improved underwriting, quality control, competitive analysis
  - Fraud detection and detection of unusual patterns (outliers)
- Other Applications
  - Text mining (news group, email, documents) and Web mining
  - Stream data mining
  - Bioinformatics and bio-data analysis

19

19

Data Mining Tasks…

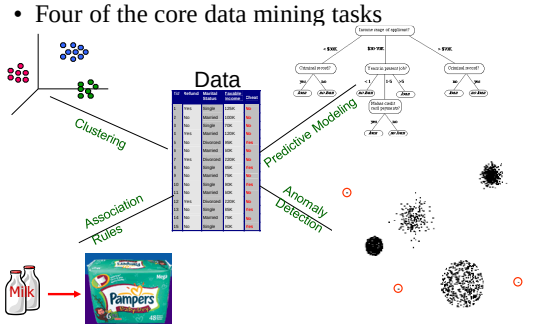
- Prediction Methods
  - Use some variables to predict unknown or future values of other variables.
    - attribute to be predicted is known as the target or dependent variable,
    - attributes used for making the prediction are known as the explanatory or independent variables.
- Description Methods
  - Find human-interpretable patterns (correlations, trends, clusters, trajectories, and anomalies) that describe the data.
    - exploratory in nature and frequently require postprocessing techniques to validate and explain the results

20

20

…Data Mining Tasks …

- Four of the core data mining tasks



21

21

Predictive Modeling

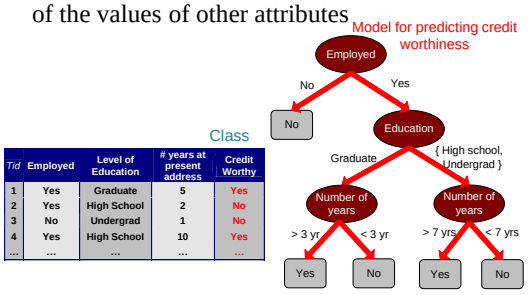
- refers to the task of building a model for the target variable as a function of the explanatory variables
- Two types of predictive modeling tasks:
  - Classification
    - used for discrete target variables
      - For example, predicting whether a web user will make a purchase at an online bookstore
  - Regression
    - used for continuous target variables
      - For example, forecasting the future price of a stock
- The goal of both tasks is to learn a model that minimizes the error between the predicted and true values of the target variable

22

22

Predictive Modeling: Classification

- Find a model for class attribute as a function of the values of other attributes



Class

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

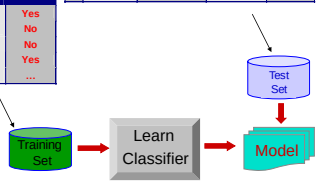
23

23

Classification Example

Tid	Employed	Level of Education	# years at present address	Credit Worthy
	categorical	categorical	quantitative	class
1	Yes	Graduate	5	Yes
2	Yes	High School	2	No
3	No	Undergrad	1	No
4	Yes	High School	10	Yes
…	…	…	…	…

Tid	Employed	Level of Education	# years at present address	Credit Worthy
1	Yes	Undergrad	7	?
2	No	Graduate	3	?
3	Yes	High School	2	?
…	…	…	…	…



24

24

Examples of Classification Task

- Classifying credit card transactions as legitimate or fraudulent
- Classifying land covers (water bodies, urban areas, forests, etc.) using satellite data
- Categorizing news stories as finance, weather, entertainment, sports, etc
- Identifying intruders in the cyberspace
- Predicting tumor cells as benign or malignant
- Classifying secondary structures of protein as alpha-helix, beta-sheet, or random coil

25

25

Classification: Application 1

- Fraud Detection
  - Goal:
    - Predict fraudulent cases in credit card transactions.
  - Approach:
    - Use credit card transactions and the information on its account-holder as attributes.
      - When does a customer buy, what does (s)he buy, how often (s)he pays on time, etc
    - Label past transactions as fraud or fair transactions.
      - This forms the class attribute.
    - Learn a model for the class of the transactions.
    - Use this model to detect fraud by observing credit card transactions on an account.

26

26

Classification: Application 2

- Churn prediction for telephone customers
  - Goal:
    - To predict whether a customer is likely to be lost to a competitor.
  - Approach:
    - Use detailed record of transactions with each of the past and present customers, to find attributes.
      - How often the customer calls, where he calls, what time-of-the day he calls most, his financial status, marital status, etc.
  - Label the customers as loyal or disloyal.
  - Find a model for loyalty.

From [Berry & Linoff] Data Mining Techniques, 1997

27

27

Classification: Application 3

- Sky Survey Cataloging
  - Goal:
    - To predict class (star or galaxy) of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory).
      - 3000 images with 23,040 x 23,040 pixels per image.
  - Approach:
    - Segment the image.
    - Measure image attributes (features) - 40 of them per object.
    - Model the class based on these features.
    - Success Story: Could find 16 new high red-shift quasars, some of the farthest objects that are difficult to find!

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996

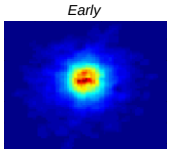
28

28

Classifying Galaxies

Courtesy: http://aps.umn.edu

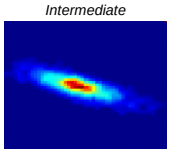
Early



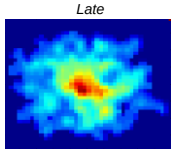
Class:
- Stages of Formation

Attributes:
- Image features,
- Characteristics of light waves received, etc.

Intermediate



Late



Data Size:
- 72 million stars, 20 million galaxies
- Object Catalog: 9 GB
- Image Database: 150 GB

29

29

Regression

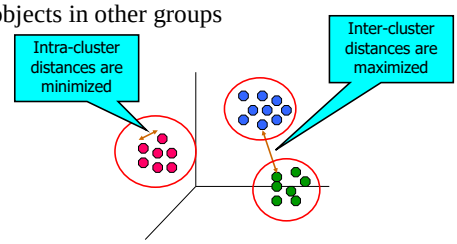
- Predict a value of a given continuous valued variable based on the values of other variables, assuming a linear or nonlinear model of dependency.
- Extensively studied in statistics, neural network fields.
- Examples:
  - Predicting sales amounts of new product based on advetising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
  - Time series prediction of stock market indices.

30

30

Clustering

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups

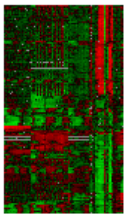


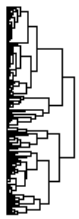
31

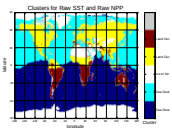
31

Applications of Cluster Analysis

- Understanding
  - Custom profiling for targeted marketing
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
- Summarization
  - Reduce the size of large data sets







Use of K-means to partition Sea Surface Temperature (SST) and Net Primary Production (NPP) into clusters that reflect the Northern and Southern Hemispheres.

32

32

Clustering: Application 1

- Market Segmentation:
  - Goal:
    - subdivide a market into distinct subsets of customers where any subset may conceivably be selected as a market target to be reached with a distinct marketing mix.
  - Approach:
    - Collect different attributes of customers based on their geographical and lifestyle related information.
    - Find clusters of similar customers.
    - Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters.

33

33

Clustering: Application 2

- Document Clustering:
  - Goal:
    - To find groups of documents that are similar to each other based on the important terms appearing in them.
  - Approach:
    - To identify frequently occurring terms in each document.
      - Form a similarity measure based on the frequencies of different terms. Use it to cluster.

Enron email dataset

34

34

Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection
  - Produce dependency rules which will predict occurrence of an item based on occurrences of other items.

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}

35

35

Association Analysis: Applications

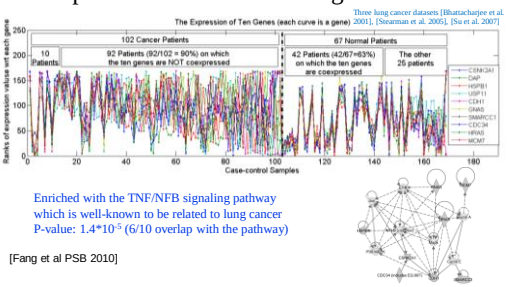
- Market-basket analysis
  - Rules are used for sales promotion, shelf management, and inventory management
- Telecommunication alarm diagnosis
  - Rules are used to find combination of alarms that occur together frequently in the same time period
- Medical Informatics
  - Rules are used to find combination of patient symptoms and test results associated with certain diseases

36

36

Association Analysis: Applications

- An Example Subspace Differential Coexpression Pattern from lung cancer dataset



Enriched with the TNF/NFB signaling pathway which is well-known to be related to lung cancer
P-value: $1.4*10^{-5}$ (6/10 overlap with the pathway)

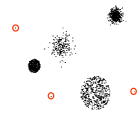
[Fang et al PSB 2010]

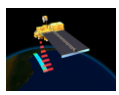
37

37

Deviation/Anomaly/Change Detection

- Detect significant deviations from normal behavior
- Applications:
  - Credit Card Fraud Detection
  - Network Intrusion Detection
  - Identify anomalous behavior from sensor networks for monitoring and surveillance.
  - Detecting changes in the global forest cover.





38

38

Motivating Challenges

- Scalability
  - Because of advances in data generation and collection, data sets with sizes of terabytes, petabytes, or even exabytes are becoming common.
    - If data mining algorithms are to handle these massive data sets, they must be scalable
- High Dimensionality
  - It is now common to encounter data sets with hundreds or thousands of attributes instead of the handful common a few decades ago
  - Data sets with temporal or spatial components also tend to have high dimensionality

39

39

Motivating Challenges

- Heterogeneous and Complex Data
  - Traditional data analysis methods often deal with data sets containing attributes of the same type, either continuous or categorical.
  - As the role of data mining in business, science, medicine, and other fields has grown, so has the need for techniques that can handle heterogeneous attributes.
  - Recent years have also seen the emergence of more complex data objects.

40

40

Motivating Challenges

- Data Ownership and Distribution
  - Sometimes, the data needed for an analysis is not stored in one location or owned by one organization.
    - Instead, the data is geographically distributed among resources belonging to multiple entities.
  - This requires the development of distributed data mining techniques.
    - The key challenges faced by distributed data mining algorithms include the following:
      - how to reduce the amount of communication needed to perform the distributed computation,
      - how to effectively consolidate the data mining results obtained from multiple sources,
      - how to address data security and privacy issues.

41

41

Motivating Challenges

- Non-traditional Analysis
  - The traditional statistical approach is based on a hypothesize-and-test paradigm.
    - an experiment is designed to gather the data, and then the data is analyzed with respect to the hypothesis
  - Current data analysis tasks often require the generation and evaluation of thousands of hypotheses, and consequently, the development of some data mining techniques has been motivated by the desire to automate the process of hypothesis generation and evaluation.

42

42