DATA MINING: **INTRODUCTION**

References:

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, ***Introduction to Data Mining***, Pearson.
- Jiawei Han, Micheline Kamber, Jian Pei, ***Data Mining Concepts and Techniques***, Third Edition, Elsevier.
- Mohammed J. Zaki, Wagner Meira Jr., ***Data Mining and Analysis: Fundamental Concepts and Algorithms,*** Cambridge University press.

Email:mazlaghani@aut.ac.ir

Files address:
fileserver\common\mazlaghani\Data Mining

Grading

- Homework + Quiz (15%+10%)
- Seminar (10%)
- Midterm + Final(25%+40%)

Introduction

- vast amounts of data
- Gather whatever data you can whenever and wherever possible.
- extracting useful challenging
- well-known applications

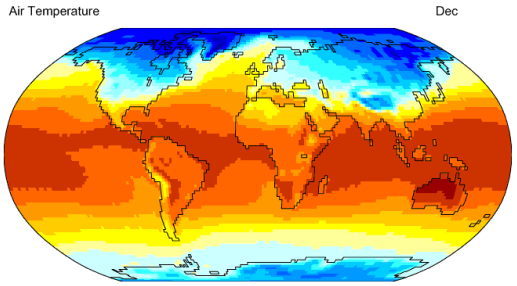
Business

- data about customer purchases
- better understand the needs of customers
- make more informed business decisions

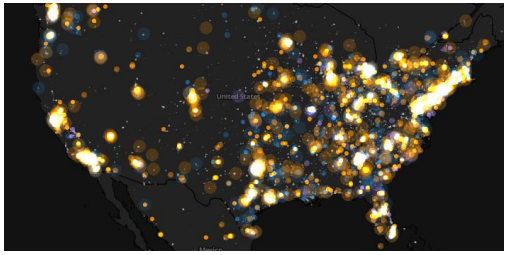
Introduction

Medicine, Science, and Engineering

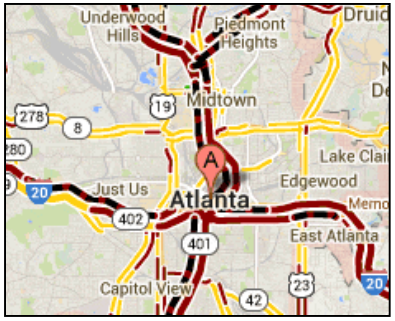
- Earth's climate system
- genomic data:microarray
- *Sensor Networks*



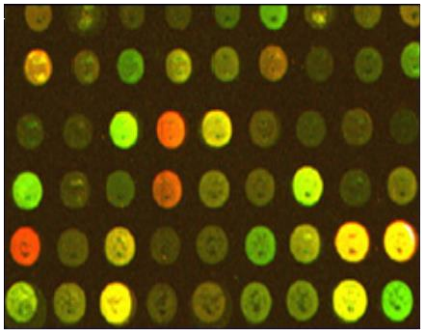
Surface Temperature of Earth



Social Networking



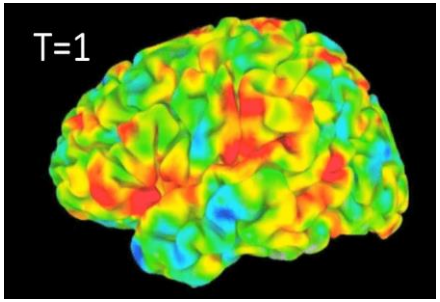
Traffic Patterns



Gene Expression Data

Introduction

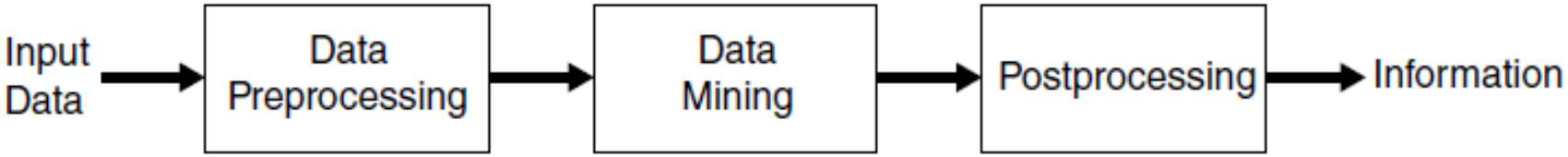
- Web data
  - Yahoo has Peta Bytes of web data
  - Facebook has billions of active users
- Bank/Credit Card transactions
- FMRI



fMRI Data from Brain

What id data mining?

- Data mining is the process of automatically discovering useful information in large data repositories
- Non-trivial extraction of implicit, previously unknown and potentially useful information from data
- Predict future



Data Mining

preprocessing : transform the raw input data into an appropriate format for subsequent analysis.

fusing data
cleaning data
selecting features

Postprocessing: only valid and useful results are incorporated

Visualization
Statistical measures

Motivating Challenges

Origins of Data Mining

- Statistics
- Machine Learning
  - ❖ Optimization
  - ❖ Information theory
  - ❖ Signal Processing

Data Mining Tasks

- Prediction Methods
  Use some variables to predict unknown or future values of other variables

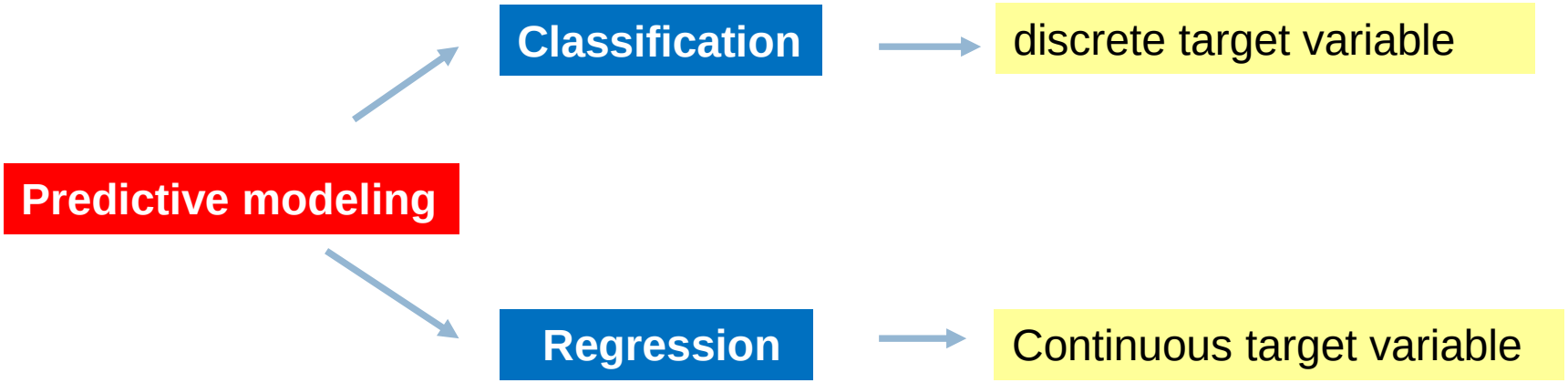
- Description Methods
  Find human-interpretable patterns that describe the data.

Core Data Mining Tasks

1. Predictive modeling

building a model for the target variable as a function of the explanatory variables

Data Mining Tasks

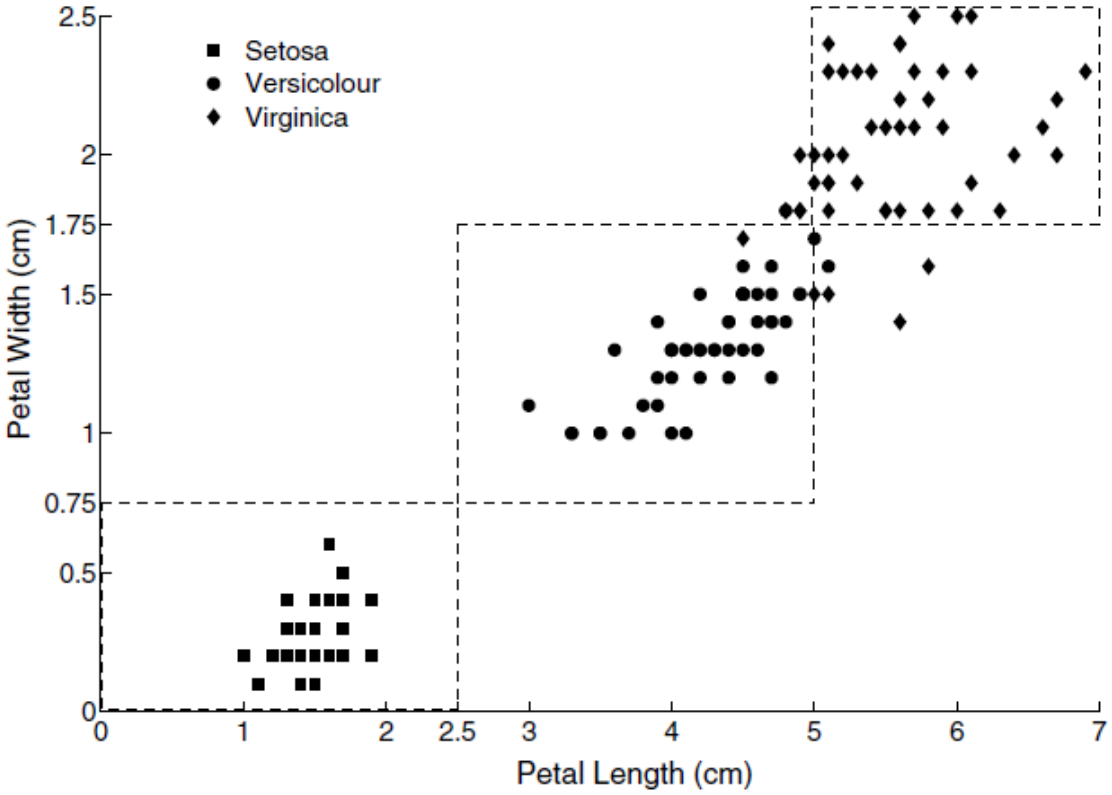


The goal of both tasks is to learn a model that minimizes the error between the predicted and true values of the target variable

Data Mining Tasks

Classification

Example :Predicting the Type of a Flower



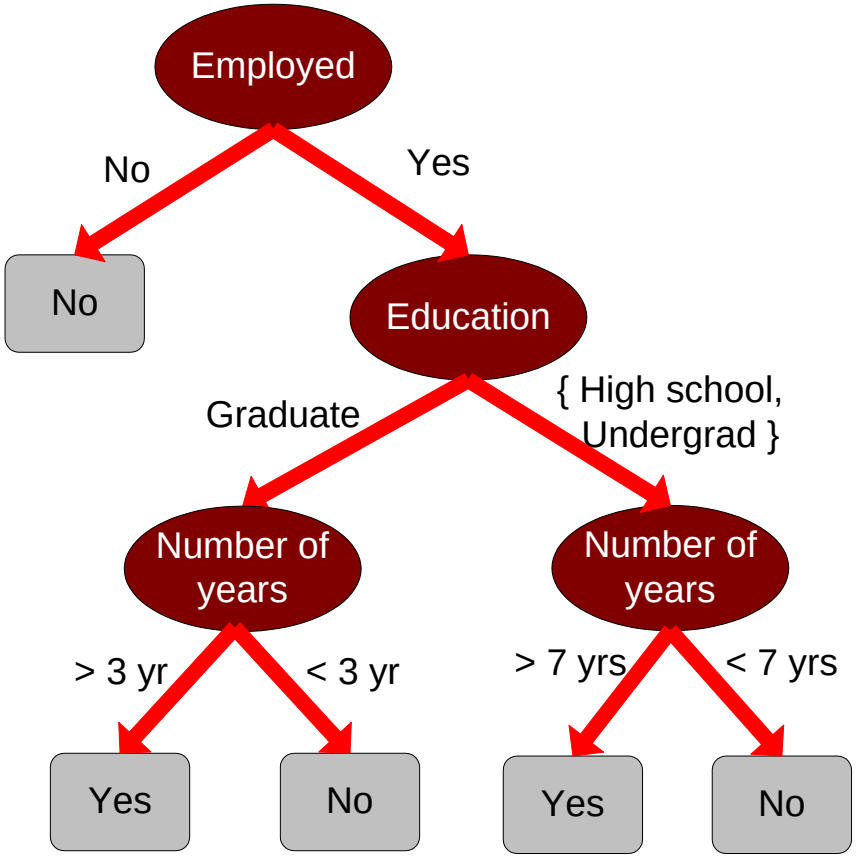
Data Mining Tasks

Example :Predicting credit worthiness

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
…	…	…	…	…

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

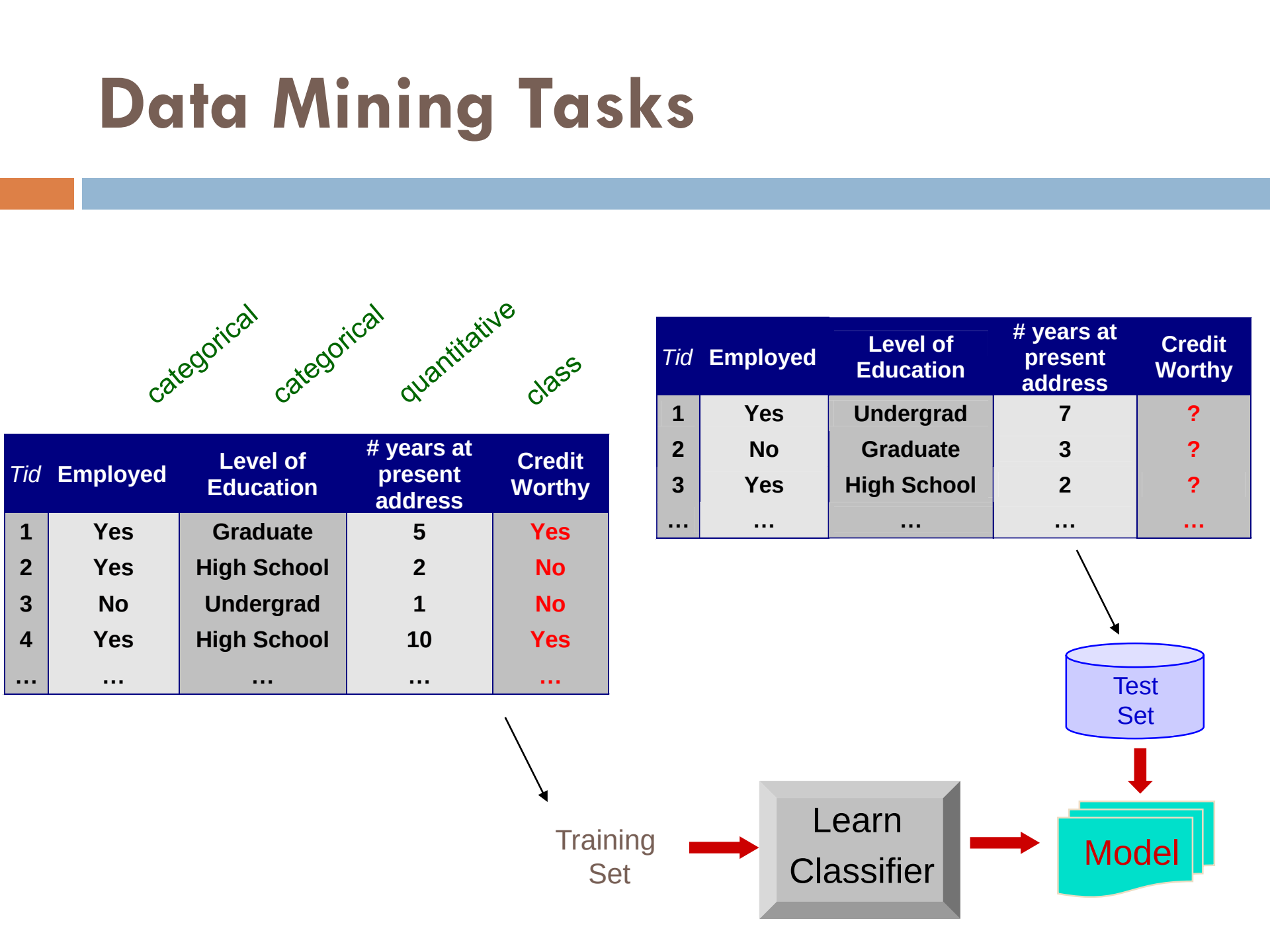


Data Mining Tasks

categorical categorical quantitative 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
…	…	…	…	…

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



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
- Predicting class of sky objects

Data Mining Tasks

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.

Examples:

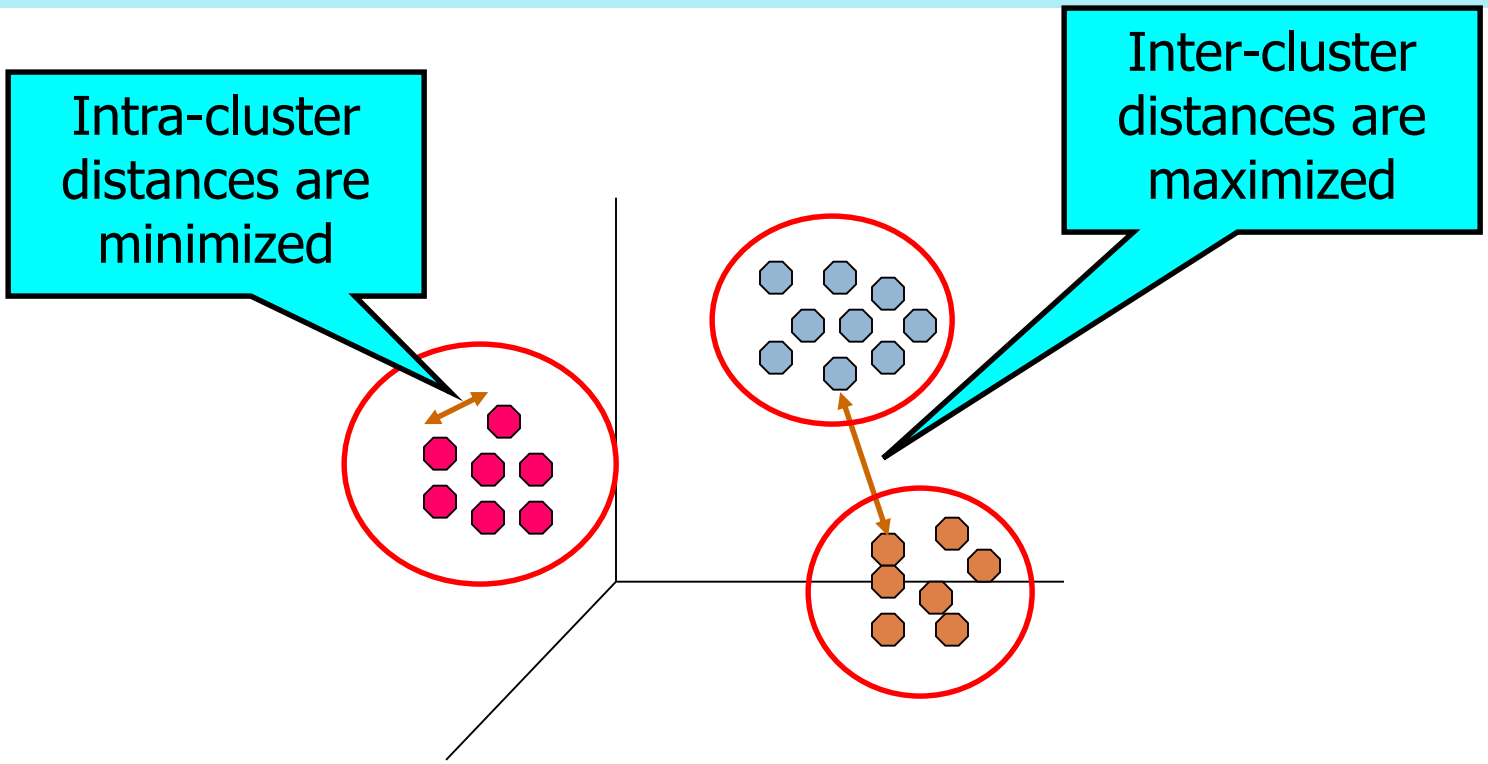
- Predicting sales amounts of new product based on advertising expenditure.
- Predicting wind velocities as a function of temperature, humidity, air pressure, etc.
- Time series prediction of stock market indices.

Clustering

Core Data Mining Tasks

2.Clustering

Goal : find groups of closely related observations so that observations that belong to the same cluster are more similar to each other than observations that belong to other clusters.



Clustering

Example (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.

Clustering

Example (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.

Clustering

Example (Document Clustering)

Article	Words
1	dollar: 1, industry: 4, country: 2, loan: 3, deal: 2, government: 2
2	machinery: 2, labor: 3, market: 4, industry: 2, work: 3, country: 1
3	job: 5, inflation: 3, rise: 2, jobless: 2, market: 3, country: 2, index: 3
4	domestic: 3, forecast: 2, gain: 1, market: 2, sale: 3, price: 2
5	patient: 4, symptom: 2, drug: 3, health: 2, clinic: 2, doctor: 2
6	pharmaceutical: 2, company: 3, drug: 2, vaccine: 1, flu: 3
7	death: 2, cancer: 4, drug: 3, public: 4, health: 3, director: 2
8	medical: 2, cost: 3, increase: 2, patient: 2, health: 3, care: 1

Association Rules

Core Data Mining Tasks

3. Association Rules

Goal: discover patterns that describe strongly associated features in the data

Examples

- ✓ Finding groups of genes that have related functionality
- ✓ Identifying Web pages that are accessed together
- ✓ Market-basket analysis
- ✓ Medical Informatics
- ✓ Finding region of brain with related functionality

Association Rules

Example: Market Basket Analysis

Transaction ID	Items
1	{Bread, Butter, Diapers, Milk}
2	{Coffee, Sugar, Cookies, Salmon}
3	{Bread, Butter, Coffee, Diapers, Milk, Eggs}
4	{Bread, Butter, Salmon, Chicken}
5	{Eggs, Bread, Butter}
6	{Salmon, Diapers, Milk}
7	{Bread, Tea, Sugar, Eggs}
8	{Coffee, Sugar, Chicken, Eggs}
9	{Bread, Diapers, Milk, Salt}
10	{Tea, Eggs, Cookies, Diapers, Milk}

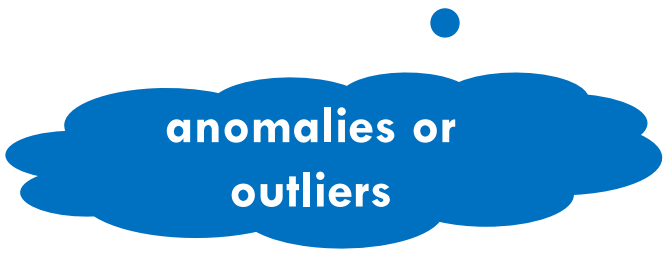
$$\{\text{Diapers}\} \longrightarrow \{\text{Milk}\}$$

Anomaly Detection

Core Data Mining Tasks

4. Anomaly detection

Goal: identifying observations whose characteristics are significantly different from the rest of the data



Examples:

- Network Intrusion Detection
- Identify anomalous behavior from sensor networks for monitoring and surveillance.
- Detecting changes in the global forest cover.