Introduction to Data Mining

Introduction Outline

Goal: Provide an overview of data mining.

- Define data mining
- Data mining vs. databases
- Basic data mining tasks
- Data mining development
- Data mining issues

Introduction

- Data is produced at a phenomenal rate
- Our ability to store has grown
- Users expect more sophisticated information
- How?

UNCOVER HIDDEN INFORMATION

DATA MINING

Data Mining

- Objective: Fit data to a model
- Potential Result: Higher-level meta information that may not be obvious when looking at raw data
- Similar terms
  – Exploratory data analysis
  – Data driven discovery
  – Deductive learning

Data Mining Algorithm

- Objective: Fit Data to a Model
  - Descriptive
  - Predictive
- Preferential Questions
  - Which technique to choose?
    - ARM/Classification/Clustering
    - Answer: Depends on what you want to do with data?
  - Search Strategy – Technique to search the data
    - Interface? Query Language?
    - Efficiency

Database Processing vs. Data Mining Processing

- Query
  - Well defined
  - SQL

- Query
  - Poorly defined
  - No precise query language

- Output
  - Precise
  - Subset of database

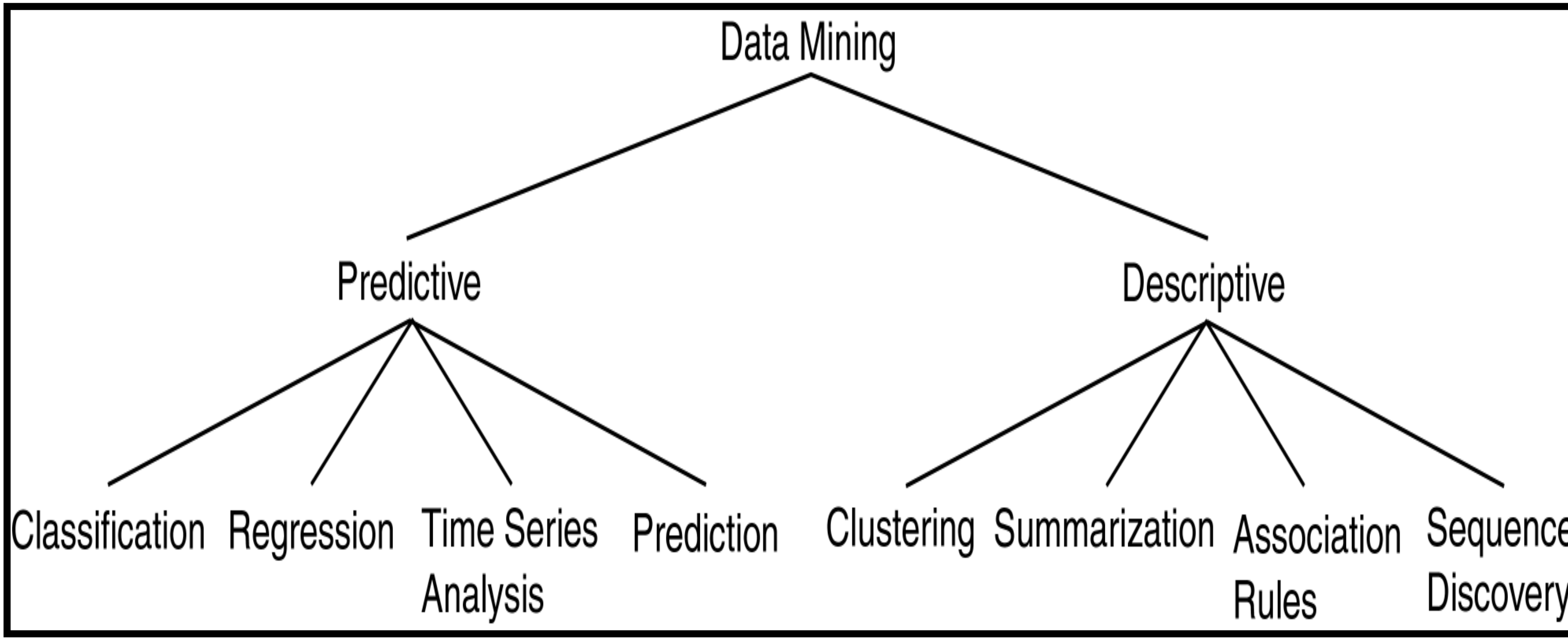
- Output
  - Fuzzy
  - Not a subset of database

Query Examples

- Database
  - Find all credit applicants with last name of Smith.
  - Identify customers who have purchased more than $10,000 in the last month.
  - Find all customers who have purchased milk

- Data Mining
  - Find all credit applicants who are poor credit risks. (classification)
  - Identify customers with similar buying habits. (Clustering)
  - Find all items which are frequently purchased with milk. (association rules)

Data Mining Models and Tasks



Basic Data Mining Tasks

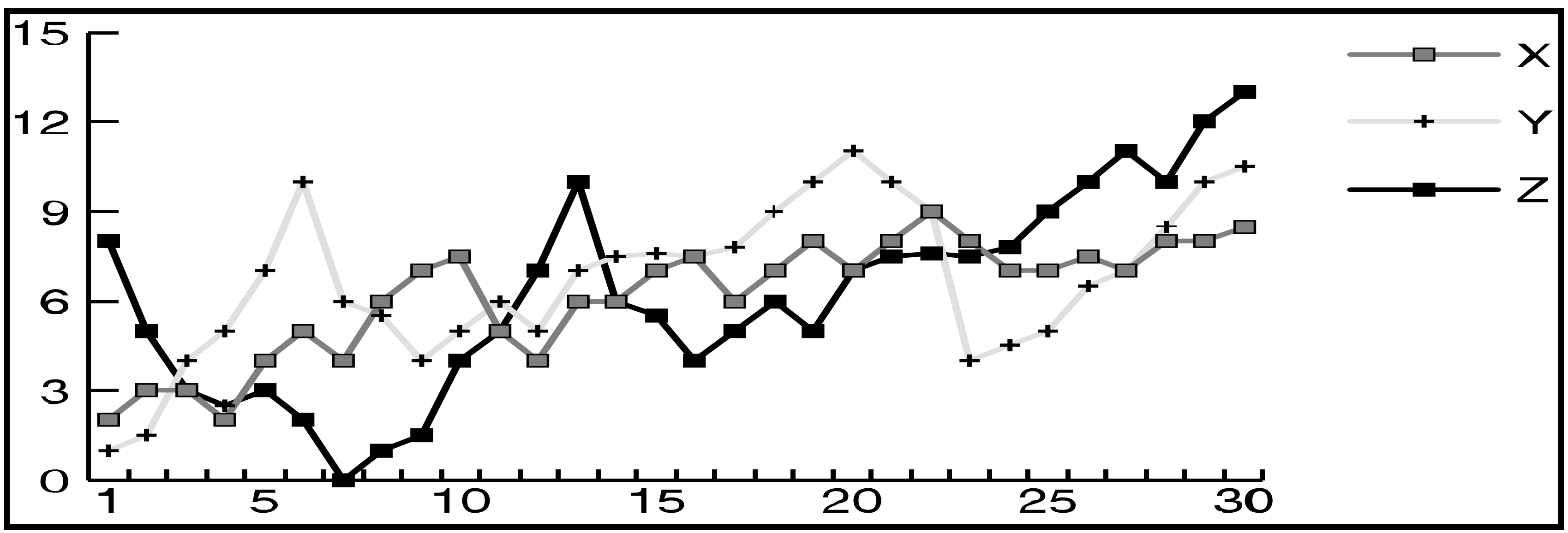
- ***Classification*** maps data into predefined groups or classes
  - Supervised learning
  - Pattern recognition
  - Prediction
- ***Regression*** is used to map a data item to a real valued prediction variable.
- ***Clustering*** groups similar data together into clusters.
  - Unsupervised learning
  - Segmentation
  - Partitioning

Basic Data Mining Tasks (cont'd)

- ***Summarization*** maps data into subsets with associated simple descriptions.
  - Characterization
  - Generalization
- ***Link Analysis*** uncovers relationships among data.
  - Affinity Analysis
  - Association Rules
  - Sequential Analysis determines sequential patterns.

Ex: Time Series Analysis

- Example: Stock Market
- Predict future values
- Determine similar patterns over time
- Classify behavior

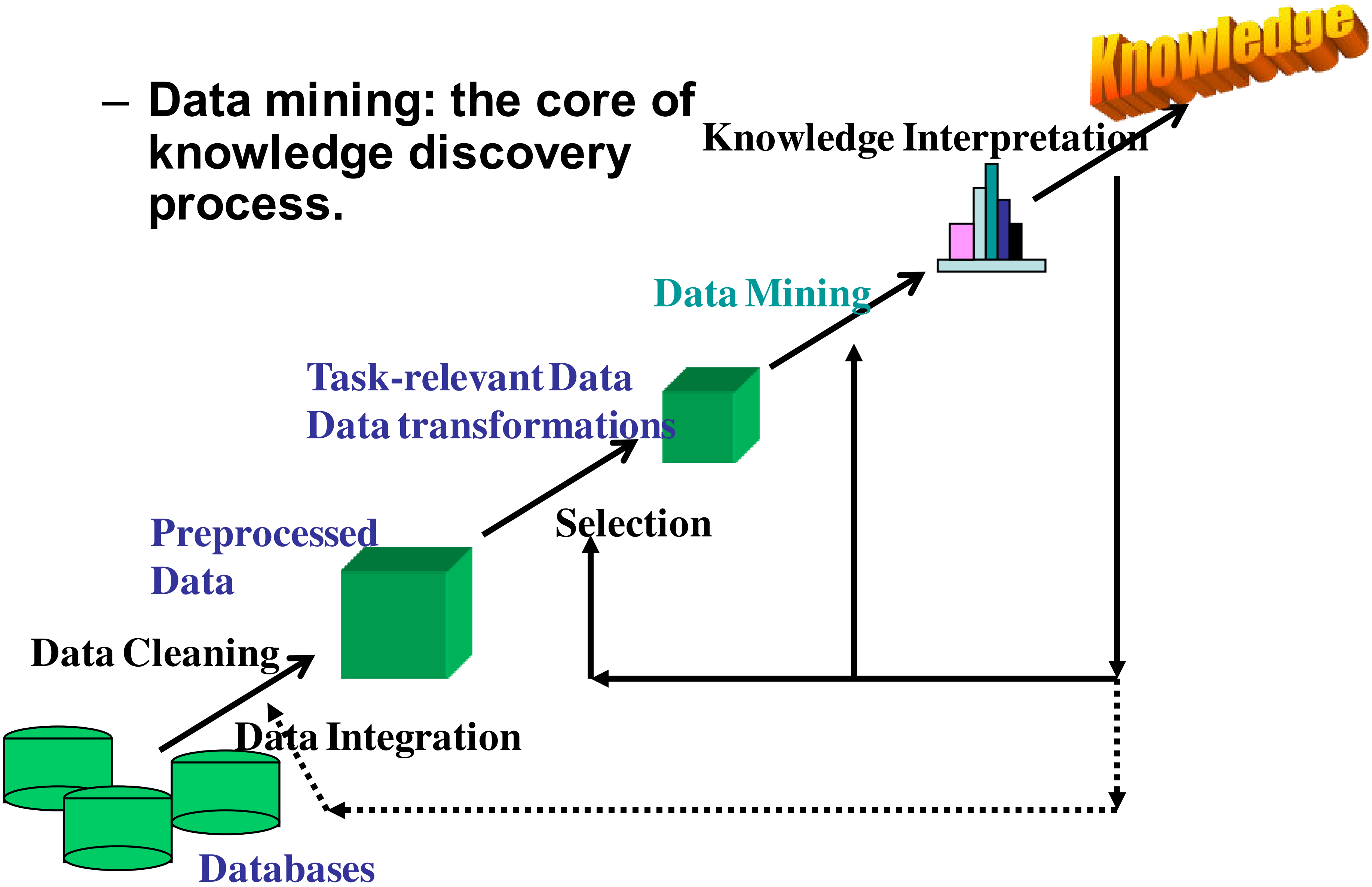


Data Mining vs. KDD

- ***Knowledge Discovery in Databases (KDD):*** process of finding useful information and patterns in data.
- ***Data Mining:*** Use of algorithms to extract the information and patterns derived by the KDD process.

Knowledge Discovery Process

– Data mining: the core of knowledge discovery process.



KDD Process Ex: Web Log

- ***Selection:***
  - Select log data (dates and locations) to use
- ***Preprocessing:***
  - Remove identifying URLs
  - Remove error logs
- ***Transformation:***
  - Sessionize (sort and group)
- ***Data Mining:***
  - Identify and count patterns
  - Construct data structure
- ***Interpretation/Evaluation:***
  - Identify and display frequently accessed sequences.
- ***Potential User Applications:***
  - Cache prediction
  - Personalization

Data Mining Development

- Relational Data Model
- SQL
- Association Rule Algorithms
- Data Warehousing
- Scalability Techniques

- Similarity Measures
- Hierarchical Clustering
- IR Systems
- Imprecise Queries
- Textual Data
- Web Search Engines

DATA MINING

- Bayes Theorem
- Regression Analysis
- EM Algorithm
- K-Means Clustering
- Time Series Analysis

- Algorithm Design Techniques
- Algorithm Analysis
- Data Structures

- Neural Networks
- Decision Tree Algorithms

[**HIGH PERFORMANCE**]

KDD Issues

- **Human Interaction**
- **Overfitting**
- **Outliers**
- **Interpretation**
- **Visualization**
- **Large Datasets**
- **High Dimensionality**

KDD Issues (cont’d)

- **Multimedia Data**
- **Missing Data**
- **Irrelevant Data**
- **Noisy Data**
- **Changing Data**
- **Integration**
- **Application**

Social Implications of DM

- Privacy
- Profiling
- Unauthorized use

Data Mining Metrics

- Usefulness
- Return on Investment (ROI)
- Accuracy
- Space/Time

Database Perspective on Data Mining

- Scalability
- Real World Data
- Updates
- Ease of Use

Outline of Today's Class

- **Statistical Basics**
  - **Point Estimation**
  - **Models Based on Summarization**
  - **Bayes Theorem**
  - **Hypothesis Testing**
  - Regression and Correlation
- **Similarity Measures**

Point Estimation

- ***Point Estimate:*** estimate a population parameter.
- May be made by calculating the parameter for a sample.
- May be used to predict value for missing data.
- Ex:
  – R contains 100 employees
  – 99 have salary information
  – Mean salary of these is $50,000
  – Use $50,000 as value of remaining employee's salary.

Is this a good idea?

Estimation Error

- ***Bias:*** Difference between expected value and actual value.

$$\boxed{Bias = E(\hat{\Theta}) - \Theta}$$

- ***Mean Squared Error (MSE):*** expected value of the squared difference between the estimate and the actual value:

$$\boxed{MSE(\hat{\Theta}) = E(\hat{\Theta} - \Theta)^2}$$

- Why square?
- Root Mean Square Error (RMSE)

Jackknife Estimate

- ***Jackknife Estimate:*** estimate of parameter is obtained by omitting one value from the set of observed values.
  - Treat the data like a population
  - Take samples from this population
  - Use these samples to estimate the parameter
- Let θ(hat) be an estimate on the entire pop.
- Let $\theta_{(j)}$(hat) be an estimator of the same form with observation j deleted
- Allows you to examine the impact of outliers!

Maximum Likelihood Estimate (MLE)

- Obtain parameter estimates that maximize the probability that the sample data occurs for the specific model.
- Joint probability for observing the sample data by multiplying the individual probabilities. Likelihood function:

$$\boxed{L(\Theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \Theta)}$$

- Maximize L.

MLE Example

- Coin toss five times: {H,H,H,H,T}
- Assuming a perfect coin with H and T equally likely, the likelihood of this sequence is:

$$\boxed{L(p \mid 1,1,1,1,0) = \prod_{i=1}^{5} 0.5 = 0.03.}$$

- However if the probability of a H is 0.8 then:

$$\boxed{L(p \mid 1,1,1,1,0) = 0.8 \times 0.8 \times 0.8 \times 0.8 \times 0.2 = 0.08.}$$

MLE Example (cont'd)

- General likelihood formula:

$$L(p \mid x_1, \ldots, x_5) = \prod_{i=1}^{5} p^{x_i} \, (1-p)^{1-x_i} = p^{\sum_{i=1}^{5} x_i} \, (1-p)^{5-\sum_{i=1}^{5} x_i}.$$

$$l(p) = logL(p) = \sum_{i=1}^{5} x_i log(p) + (5 - \sum_{i=1}^{5} x_i) \; log(1-p)$$

$$\frac{\partial l(p)}{\partial p} = \sum_{i=1}^{5} \frac{x_i}{p} - \frac{5 - \sum_{i=1}^{5} x_i}{1-p}.$$

$$p = \frac{\sum_{i=1}^{5} x_i}{5}$$

- Estimate for p is then 4/5 = 0.8

Expectation-Maximization (EM)

- Solves estimation with incomplete data.
- Obtain initial estimates for parameters.
- Iteratively use estimates for missing data and continue until convergence.

EM Example

$\{1, 5, 10, 4\}$; $n = 6$ $k = 4$; **Guess** $\hat{\mu}^0 = 3$.

$$\hat{\mu}^1 = \frac{\sum_{i=1}^{k} x_i}{n} + \frac{\sum_{i=k+1}^{n} x_i}{n} = 3.33 + \frac{3+3}{6} = 4.33$$

$$\hat{\mu}^2 = \frac{\sum_{i=1}^{k} x_i}{n} + \frac{\sum_{i=k+1}^{n} x_i}{n} = 3.33 + \frac{4.33+4.33}{6} = 4.77$$

$$\hat{\mu}^3 = \frac{\sum_{i=1}^{k} x_i}{n} + \frac{\sum_{i=k+1}^{n} x_i}{n} = 3.33 + \frac{4.77+4.77}{6} = 4.92$$

$$\hat{\mu}^4 = \frac{\sum_{i=1}^{k} x_i}{n} + \frac{\sum_{i=k+1}^{n} x_i}{n} = 3.33 + \frac{4.92+4.92}{6} = 4.97$$

EM Algorithm

Input:

$\Theta = \{\theta_1, ..., \theta_p\}$ **//Parameters to be Estimated**

$X_{obs} = \{x_1, ..., x_k\}$ **//Input Database Values Observed**

$X_{miss} = \{x_{k+1}, ..., x_n\}$ **//Input Database Values Missing**

Output:

$\hat{\Theta}$ **//Estimates for** Θ

EM Algorithm:

i := 0;

Obtain initial parameter MLE estimate,$\hat{\Theta}^i$**;**

repeat

Estimate missing data,$\hat{X}^i_{miss}$**;**

i++;

Obtain next parameter estimate,$\hat{\theta}^i$ **to maximize data;**

until estimate converges;

Bayes Theorem Example

- Credit authorizations (hypotheses): h_1=authorize purchase, h_2 = authorize after further identification, h_3=do not authorize, h_4= do not authorize but contact police
- Assign twelve data values for all combinations of credit and income:

	1	2	3	4
Excellent	x_1	x_2	x_3	x_4
Good	x_5	x_6	x_7	x_8
Bad	x_9	x_{10}	x_{11}	x_{12}

- From training data: $P(h_1) = 60\%$; $P(h_2)=20\%$; $P(h_3)=10\%$; $P(h_4)=10\%$.

Bayes Example(cont'd)

- Training Data:

ID	Income	Credit	Class	x_i
1	4	Excellent	h_1	x_4
2	3	Good	h_1	x_7
3	2	Excellent	h_1	x_2
4	3	Good	h_1	x_7
5	4	Good	h_1	x_8
6	2	Excellent	h_1	x_2
7	3	Bad	h_2	x_{11}
8	2	Bad	h_2	x_{10}
9	3	Bad	h_3	x_{11}
10	1	Bad	h_4	x_9

Bayes Example(cont'd)

- Calculate $P(x_i|h_j)$ and $P(x_i)$
- Ex: $P(x_7|h_1)=2/6$; $P(x_4|h_1)=1/6$; $P(x_2|h_1)=2/6$; $P(x_8|h_1)=1/6$; $P(x_i|h_1)=0$ for all other x_i.
- Predict the class for x_4:
  - Calculate $P(h_j|x_4)$ for all h_j.
  - Place x_4 in class with largest value.
  - Ex:
    - $P(h_1|x_4)=(P(x_4|h_1)(P(h_1))/P(x_4)$
      $=(1/6)(0.6)/0.1=1.$
    - x_4 in class h_1.

Other Statistical Measures

$$\chi^2 = \sum \frac{(O - E)^2}{E}.$$

- Chi-Squared
  - O – observed value
  - E – Expected value based on hypothesis.
- Jackknife Estimate
  - estimate of parameter is obtained by omitting one value from the set of observed values.
- Regression
  - Predict future values based on past values
  - ***Linear Regression*** assumes linear relationship exists.

$$y = c_0 + c_1 x_1 + \ldots + c_n x_n$$

    - Find values to best fit the data
- Correlation

Similarity Measures

- Determine similarity between two objects.
- Similarity characteristics:

- $\forall t_i \in D, sim(t_i, t_i) = 1$
- $\forall t_i, t_j \in D, sim(t_i, t_j) = 0$ if t_i and t_j are not alike at all.
- $\forall t_i, t_j, t_k \in D, sim(t_i, t_j) < sim(t_i, t_k)$ if t_i is more like t_k than it is like t_j.

- Alternatively, distance measure measure how unlike or dissimilar objects are.

Similarity Measures

Dice: $$sim(t_i, t_j) = \frac{2\sum_{h=1}^{k} t_{ih}t_{jh}}{\sum_{h=1}^{k} t_{ih}^2 + \sum_{h=1}^{k} t_{jh}^2}$$

Jaccard: $$sim(t_i, t_j) = \frac{\sum_{h=1}^{k} t_{ih}t_{jh}}{\sum_{h=1}^{k} t_{ih}^2 + \sum_{h=1}^{k} t_{jh}^2 - \sum_{h=1}^{k} t_{ih}t_{jh}}$$

Cosine: $$sim(t_i, t_j) = \frac{\sum_{h=1}^{k} t_{ih}t_{jh}}{\sqrt{\sum_{h=1}^{k} t_{ih}^2 \sum_{h=1}^{k} t_{jh}^2}}$$

Overlap: $$sim(t_i, t_j) = \frac{\sum_{h=1}^{k} t_{ih}t_{jh}}{min(\sum_{h=1}^{k} t_{ih}^2, \sum_{h=1}^{k} t_{jh}^2)}$$

Distance Measures

- Measure dissimilarity between objects

$$\textbf{Euclidean:}\ dis(t_i, t_j) = \sqrt{\sum_{h=1}^{k} (t_{ih} - t_{jh})^2}$$

$$\textbf{Manhattan:}\ dis(t_i, t_j) = \sum_{h=1}^{k} \mid (t_{ih} - t_{jh}) \mid$$

Information Retrieval

- ***Information Retrieval (IR):*** retrieving desired information from textual data.
- Library Science
- Digital Libraries
- Web Search Engines
- Traditionally keyword based
- Sample query:

  Find all documents about "data mining".

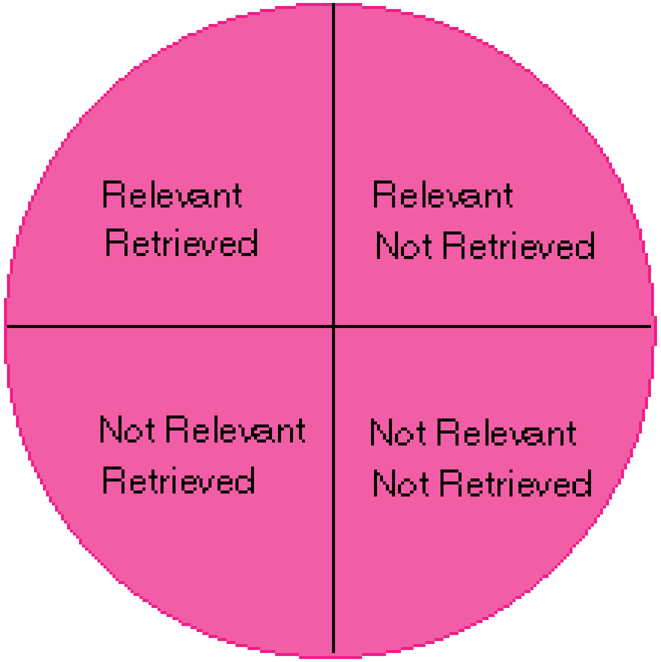
DM: Similarity measures;

Mine text/Web data.

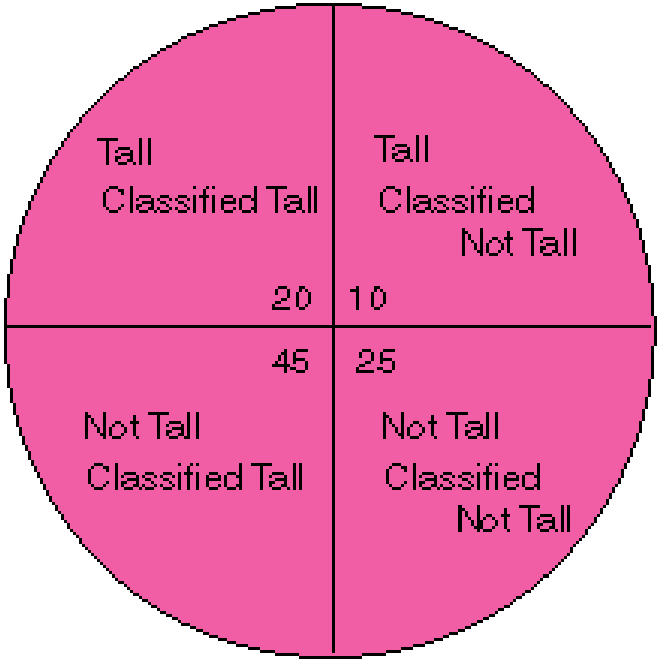
Information Retrieval (cont'd)

- ***Similarity:*** measure of how close a query is to a document.
- Documents which are "close enough" are retrieved.
- Metrics:
  - ***Precision*** = $\dfrac{|\text{Relevant and Retrieved}|}{|\text{Retrieved}|}$
  - ***Recall*** = $\dfrac{|\text{Relevant and Retrieved}|}{|\text{Relevant}|}$

IR Query Result Measures and Classification



IR



Classification