Principles of Knowledge Discovery in Data

Winter 2007

Dr. Osmar R. Zaïane

University of Alberta

Database Laboratory

Who Am I?

蔡頤安

UNIVERSITY OF ALBERTA

Osmar R. Zaïane, Ph.D.
Associate Professor
Department of Computing Science

221 Athabasca Hall
Edmonton, Alberta
Canada T6G 2E8

Telephone: Office +1 (780) 492 2860
Fax +1 (780) 492 1071
E-mail: zaiane@cs.ualberta.ca
http://www.cs.ualberta.ca/~zaiane/

PhD on *Web Mining* & *Multimedia Mining* With **Dr. Jiawei Han** at Simon Fraser University, Canada

Research Interests:
Data Mining,
Web Mining,
Multimedia Mining,
Data Visualization,
Information Retrieval.

Applications:
Analytic Tools,
Adaptive Systems,
Intelligent Systems,
Diagnostic and Categorization,
Recommender Systems

Achievements:
(in last 6 years):
2 PhD and 18 MSc,
80+ publications,
WEBKDD and MDM/KDD co-chair (2000 to 2003)
Currently: 4 PhD and 0 MSc students

Principles of Knowledge Discovery in Data

Class and Office Hours

Class:
Mondays-Wednesday-Fridays from 10:00 to 10:50

Office Hours:
Mondays from 13:00 to 14:00

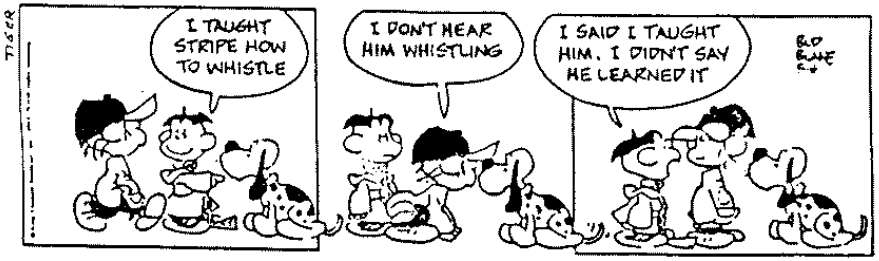
But I prefer Class:
Tuesdays and Thursdays from 9:30 to 10:50 or even better
Once a week from 9:00 to 11:50 (with one break)

CMPT Schedule

	Monday	Tuesday	Wednesday	Thursday	Friday
8:00-9:00					
9:00-10:00	611	631	611	631	611
10:00-11:00	695		695		695
11:00-12:00	606	675	606	675	606
12:00-13:00	530 617	674	530 617	674	530 617
13:00-14:00	511 610		511 610		511 610
14:00-15:00	652	551	652	551	652
15:00-16:00					
16:00-17:00		690		690	
17:00-18:00					

Course Requirements

- Understand the basic concepts of database systems
- Understand the basic concepts of artificial intelligence and machine learning
- Be able to develop applications in C/C^{++} or Java



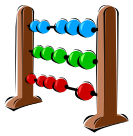
Course Objectives

To provide an introduction to knowledge discovery in databases and complex data repositories, and to present basic concepts relevant to real data mining applications, as well as reveal important research issues germane to the knowledge discovery domain and advanced mining applications.

Students will understand the fundamental concepts underlying knowledge discovery in databases and gain hands-on experience with implementation of some data mining algorithms applied to real world cases.

Evaluation and Grading

There is no final exam for this course, but there are assignments, presentations, a midterm and a project.
I will be evaluating all these activities out of 100% and give a final grade based on the evaluation of the activities.
The midterm has two parts: a <u>take-home exam</u> + <u>oral exam</u>.



- Assignments 20%
- Midterm 25%
- Project 39%
  - Quality of presentation + quality of report + quality of demos
  - Preliminary project demo (week 11) and final project demo (week 14) have the same weight (could be week 15)
- Class presentations 16%
  - Quality of presentation + quality of slides + peer evaluation

- A+ will be given only for outstanding achievement.

More About Evaluation

Re-examination.

None, except as per regulation.

Collaboration.

Collaborate on assignments and projects, etc; do not merely copy.

Plagiarism.

Work submitted by a student that is the work of another student or any other person is considered plagiarism. Read **Sections 26.1.4** and **26.1.5** of the University of Alberta calendar. Cases of plagiarism are immediately referred to the Dean of Science, who determines what course of action is appropriate.

About Plagiarism

Plagiarism, cheating, misrepresentation of facts and participation in such offences are viewed as serious academic offences by the University and by the Campus Law Review Committee (CLRC) of General Faculties Council.
Sanctions for such offences range from a reprimand to suspension or expulsion from the University.

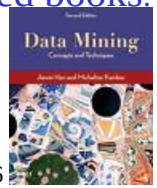
Notes and Textbook

Course home page:
http://www.cs.ualberta.ca/~zaiane/courses/cmput695/
We will also use the **Sakai** system for on-line delivery.

We will also have a mailing list and newsgroup for the course.

No Textbook but recommended books:

Data Mining: Concepts and Techniques
Jiawei Han and Micheline Kamber
Morgan Kaufmann Publisher

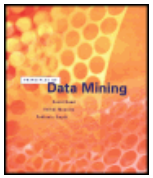


http://www-faculty.cs.uiuc.edu/~hanj/bk2/

2006
ISBN 1-55860-901-6
800 pages

2001
ISBN 1-55860-489-8
550 pages

Other Books



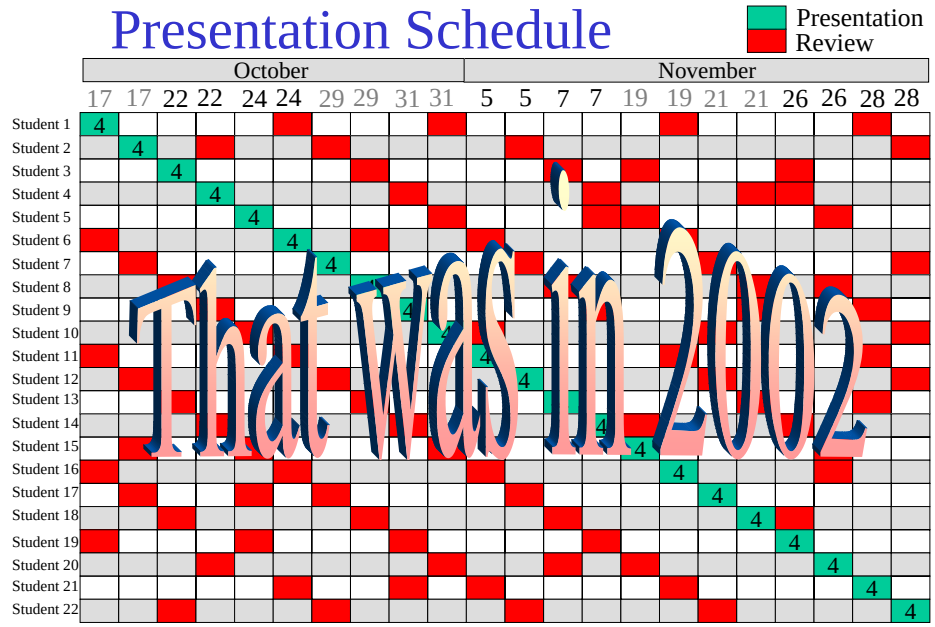
- Principles of Data Mining
  - David Hand, Heikki Mannila, Padhraic Smyth,
    MIT Press, 2001, ISBN 0-262-08290-X, 546 pages

- Data Mining: Introductory and Advanced Topics
  - Margaret H. Dunham,
    Prentice Hall, 2003, ISBN 0-13-088892-3, 315 pages

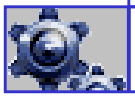
- Dealing with the data flood: Mining data, text and multimedia
  - Edited by Jeroen Meij,
    SST Publications, 2002, ISBN 90-804496-6-0, 896 pages

- Introduction to Data Mining
  - Pang-Ning Tan, Michael Steinbach, Vipin Kumar
    Addison Wesley, ISBN: 0-321-32136-7, 769 pages

Presentation Schedule



Projects

	Choice	Deliverables
	Implement data mining project	Project proposal + 10' proposal presentation + project pre-demo + final demo + project report

Examples and details of data mining projects will be posted on the course web site.

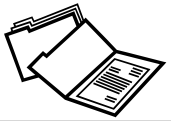
Assignments

1- Competition in one algorithm implementation
2- Devising Exercises and solving them
3- Review of a paper

More About Projects

Students should write a project proposal (1 or 2 pages).

- project topic;
- implementation choices;
- approach;
- schedule.



All projects are demonstrated at the end of the semester.
April 9-11-13 to the whole class.
Preliminary project demos are private demos given to the instructor on **week March 19**.

Implementations: C/C++ or Java,
OS: Linux, Window XP/2000 , or other systems.

Course Schedule

(**Tentative**, subject to changes)

There are 14 weeks from January 8th to April 13th.

Week 1:	January 8	: Introduction to Data Mining
Week 2:	January 15	: Association Rules
Week 3:	January 22	: Association Rules (advanced topics)
Week 4:	January 29	: Sequential Pattern Analysis
Week 5:	February 5	: Classification (Neural Networks)
Week 6:	February 12	: Classification (Decision Trees and others)
Week 7:	February 19	: Winter Reading Week
Week 8:	February 26	: Data Clustering
Week 9:	March 5	: Data Clustering in subspaces
Week 10:	March 12	: Outlier Detection
Week 11:	March 19	: Contrast sets
Week 12:	March 26	: Web Mining
Week 13:	April 2	: Web Mining (Good Friday April 6)
Week 14:	April 9	: Project Demos (Easter Monday April 9)

Due dates
-Midterm
week 8
-Assignment 1
week 6
-Assignment 2
variable dates

Course Content

- Introduction to Data Mining
- Association analysis
- Sequential Pattern Analysis
- Classification and prediction
- Contrast Sets
- Data Clustering
- Outlier Detection
- Web Mining
- Other topics if time permits (spatial data, biomedical data, etc.)

Let's do some Data Mining!

For those of you who watch what you eat...
Here's the final word on nutrition and health. It's a relief to know the truth after all those conflicting medical studies.

- **The Japanese eat very little fat and suffer fewer heart attacks than the British or Americans.**
- **The Mexicans eat a lot of fat and suffer fewer heart attacks than the British or Americans.**
- **The Japanese drink very little red wine and suffer fewer heart attacks than the British or Americans**
- **The Italians drink excessive amounts of red wine and suffer fewer heart attacks than the British or Americans.**
- **The Germans drink a lot of beers and eat lots of sausages and fats and suffer fewer heart attacks than the British or Americans.**

CONCLUSION:
Eat and drink what you like. Speaking English is apparently what kills you.

Quick Overview of some Data Mining Operations

Association Rules
Clustering
Classification

What Is Association Mining?

- **Association rule mining searches for relationships between items in a dataset**:
  - Finding association, correlation, or causal structures among sets of items or objects in transaction databases, relational databases, and other information repositories.
  - Rule form: "**Body → Head [support, confidence]**".
- **Examples:**
  - buys(x, "bread") → buys(x, "milk") [0.6%, 65%]
  - major(x, "CS") ^ takes(x, "DB") → grade(x, "A") [1%, 75%]

Basic Concepts

A transaction is a set of items: $T=\{i_a, i_b,\ldots i_t\}$

$T \subset I$, where I is the set of all possible items $\{i_1, i_2,\ldots i_n\}$

D, the task relevant data, is a set of transactions.

An association rule is of the form:
$P \rightarrow Q$, where $P \subset I$, $Q \subset I$, and $P \cap Q = \varnothing$

$P \rightarrow Q$ holds in D with <u>support</u> s
and
$P \rightarrow Q$ has a <u>confidence</u> c in the transaction set D.

Support($P \rightarrow Q$) = Probability($P \cup Q$)
Confidence($P \rightarrow Q$)=Probability(Q/P)

Association Rule Mining

Frequent Itemset Mining

FIM (1) abc

Bound by a **support** threshold

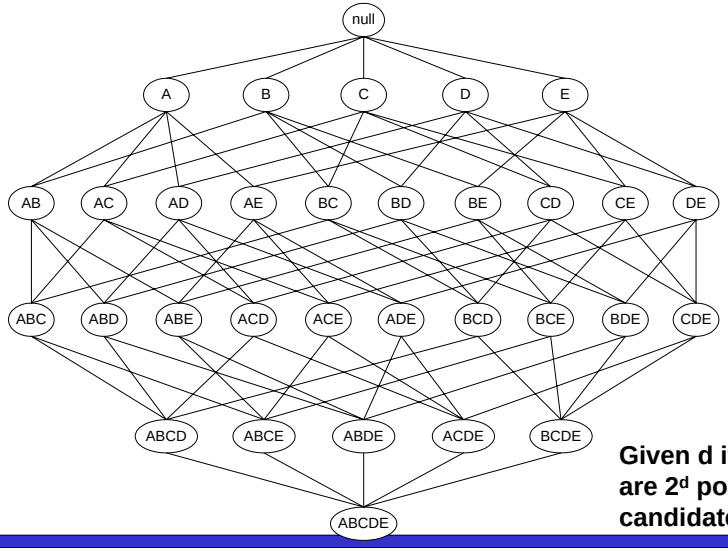
Association Rules Generation

(2) ab→c
b→ac

Bound by a **confidence** threshold

- Frequent itemset generation is still computationally expensive

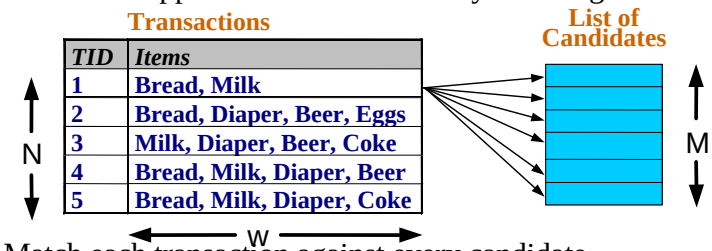
Frequent Itemset Generation



Given d items, there are 2^d possible candidate itemsets

Frequent Itemset Generation

- Brute-force approach (Basic approach):
  - Each itemset in the lattice is a candidate frequent itemset
  - Count the support of each candidate by scanning the database

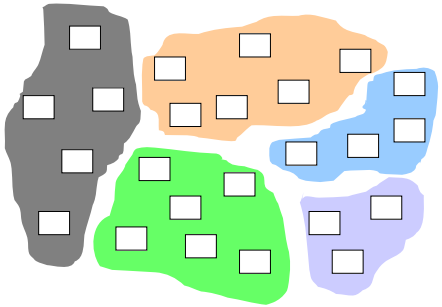


TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

  - Match each transaction against every candidate
  - Complexity ~ O(NMw) => **Expensive since M = 2^d!!!**

Obviously not the right way to do it.

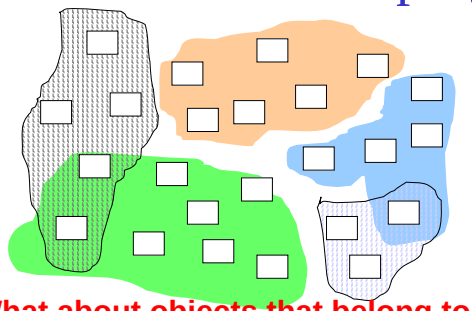
Grouping



Grouping
Clustering
Partitioning

- **We need a notion of similarity or closeness (what features?)**
- **Should we know apriori how many clusters exist?**
- **How do we characterize members of groups?**
- **How do we label groups?**

Grouping

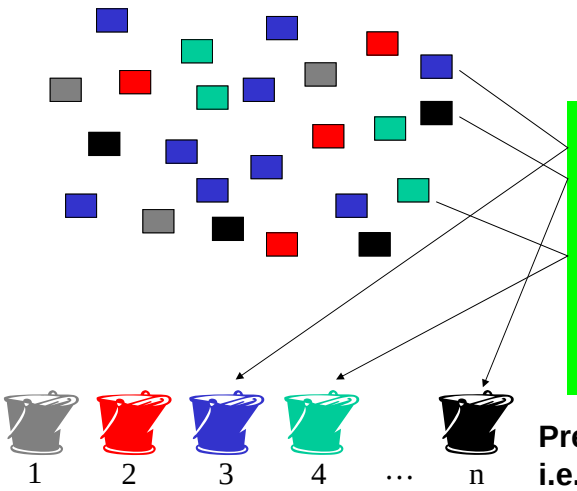


Grouping
Clustering
Partitioning

What about objects that belong to different groups?

- **We need a notion of similarity or closeness (what features?)**
- **Should we know apriori how many clusters exist?**
- **How do we characterize members of groups?**
- **How do we label groups?**

Classification



Classification
Categorization

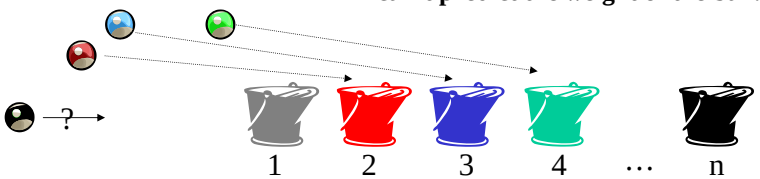
Predefined buckets
i.e. known labels

What is Classification?

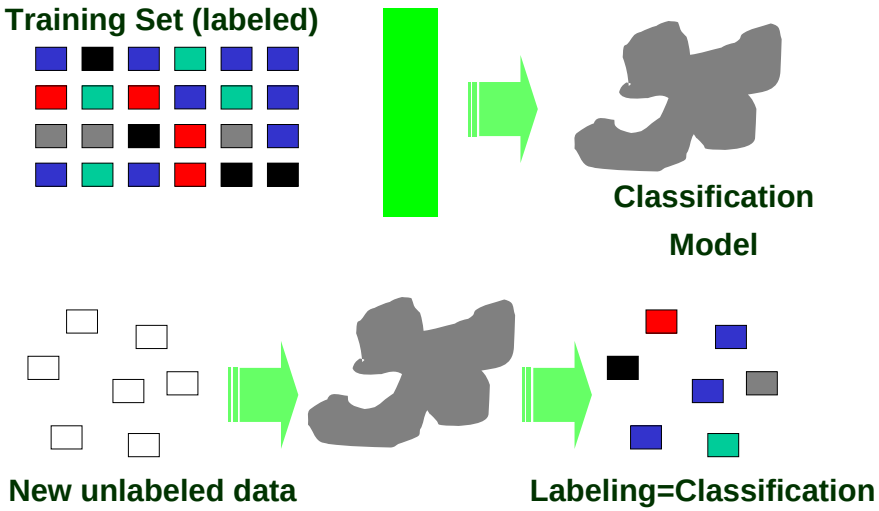
The goal of data classification is to organize and categorize data in distinct classes.

- A model is first created based on the data distribution.
- The model is then used to classify new data.
- Given the model, a class can be predicted for new data.

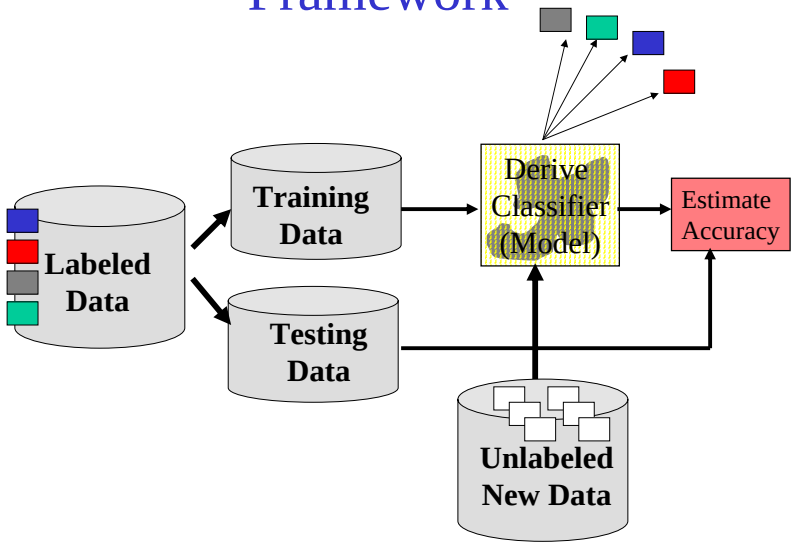
With classification, I can predict in which bucket to put the ball, but I can't predict the weight of the ball.



Classification = Learning a Model



Framework



Classification Methods

- Decision Tree Induction
- Neural Networks
- Bayesian Classification
- K-Nearest Neighbour
- Support Vector Machines
- Associative Classifiers
- Case-Based Reasoning
- Genetic Algorithms
- Rough Set Theory
- Fuzzy Sets
- Etc.