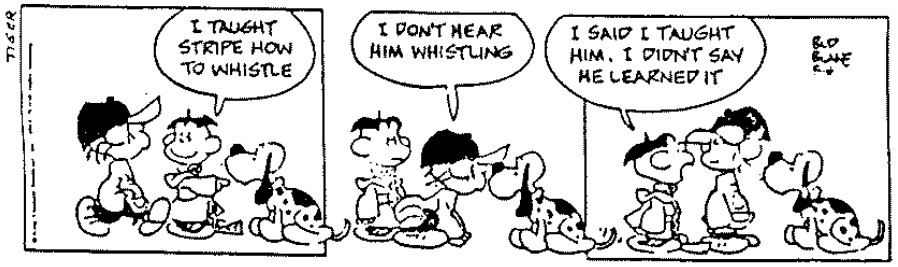
Principles of Knowledge Discovery in Data

Fall 2004

Dr. Osmar R. Zaïane

蔡頤安

University of Alberta



Class and Office Hours

Class:
Tuesdays and Thursdays from 9:30 to 10:50

Office Hours:
Wednesdays from 9:30 to 11:00

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

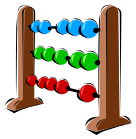
Course Objectives

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 take-home exam + oral exam.



- Assignments (4) 20%
- Midterm 25%
- Project 39%
  - Quality of presentation + quality of report + quality of demos
  - Preliminary project demo (week 12) and final project demo (week 16) have the same weight
- 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.

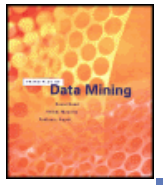
Notes and Textbook

Course home page:
http://www.cs.ualberta.ca/~zaiane/courses/cmput695/

We will also have a mailing list for the course (probably also a newsgroup).

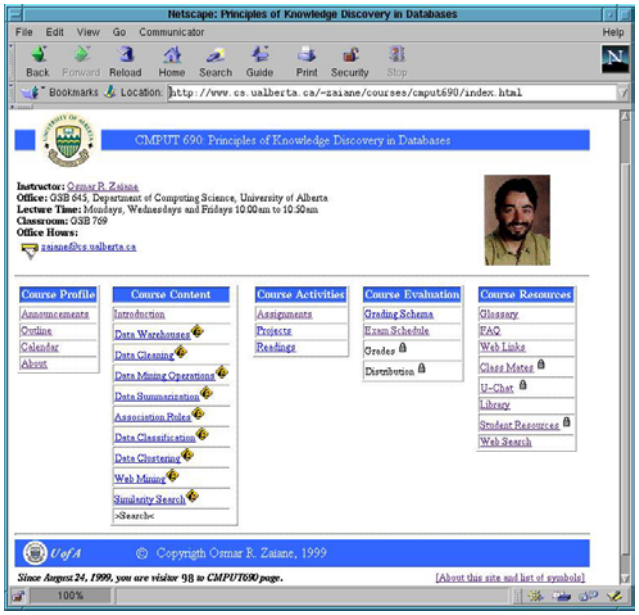
Textbook:
Data Mining: Concepts and Techniques
Jiawei Han and Micheline Kamber
Morgan Kaufmann Publisher, 2001
ISBN 1-55860-489-8

Other Books

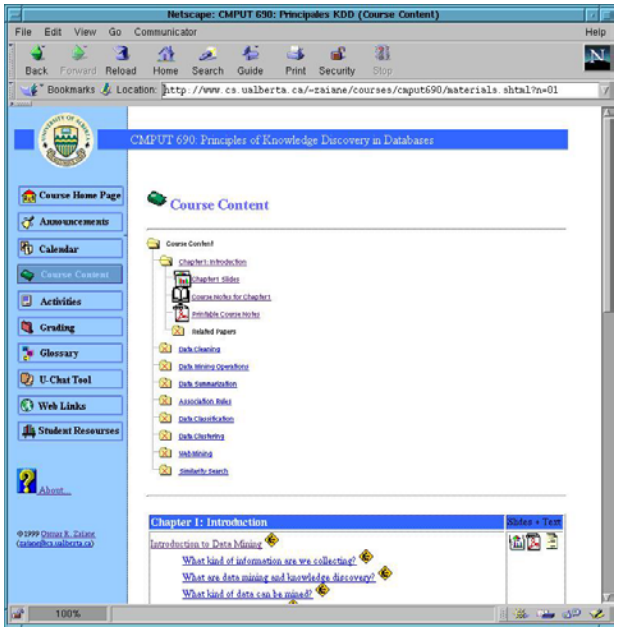


- Principles of Data Mining
  - David Hand, Heikki Mannila, Padhraic Smyth, MIT Press, 2001, ISBN 0-262-08290-X 546 pages

- 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



Course Web Page



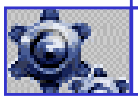
Course Content, Slides, etc.

On-line Resources

- Course notes
- Course slides
- Web links
- Glossary
- Student submitted resources
- U-Chat
- Newsgroup
- Frequently asked questions

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.

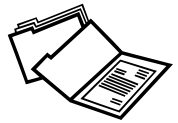
Assignments

1- Competition in one algorithm implementation
2- evaluation of off the shelf data mining tools
3- Use of educational DM tool to evaluate algorithms
4- 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.
December 2 and 7 to the whole class.
Preliminary project demos are private demos given to the instructor on **week November 22**.

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

Course Schedule (Tentative, subject to changes)

There are 14 weeks from Sept. 8th to Dec. 8th.
First class starts September 9th and classes end December 7th.

	Tuesday	Thursday
Week 1:		Sept. 9: Introduction
Week 2:	Sept. 14: Intro DM	Sept. 16: DM operations
Week 3:	Sept. 21: Assoc. Rules	Sept. 23: Assoc. Rules
Week 4:	Sept. 28: Data Prep.	Sept. 30: Data Warehouse
Week 5:	Oct. 5: Char Rules	Oct. 7: Classification
Week 6:	Oct. 12: Clustering	Oct. 14: Clustering
Week 7:	Oct. 19: Outliers	Oct. 21: Spatial & MM
Week 8:	Oct. 26: Web Mining	Oct. 31: Web Mining
Week 9:	Nov. 2: PPDM	Nov. 4: Advanced Topics
Week 10:	Nov. 9: Papers 1&2	Nov. 11: No class
Week 11:	Nov. 16: Papers 1&2	Nov. 18: Papers 3&4
Week 12:	Nov. 23: Papers 5&6	Nov. 25: Papers 7&8
Week 13:	Nov. 30 Papers 9&10	Dec. 2: Project Presentat.
Week 14:	Dec. 7: Final Demos	

Away (out of town)
To be confirmed
November 2nd
November 4th
Nov. 1-4: ICDM

Due dates
-Midterm **week 8**
-Project proposals **week 5**
-Project preliminary demo **week 12**
- Project reports **week 13**
- Project final demo **week 14**

Course Content

- Introduction to Data Mining
- Data warehousing and OLAP
- Data cleaning
- Data mining operations
- Data summarization
- Association analysis
- Classification and prediction
- Clustering
- Web Mining
- Multimedia and Spatial Mining
- *Other topics if time permits*

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 $\{i_1, i_2,\ldots i_n\}$

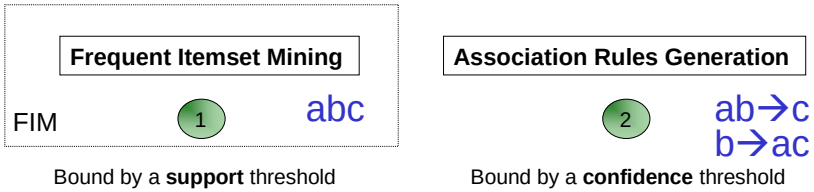
D, the task relevant data, is a set of transactions.

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 support s
and
$P \rightarrow Q$ has a confidence c in the transaction set D.

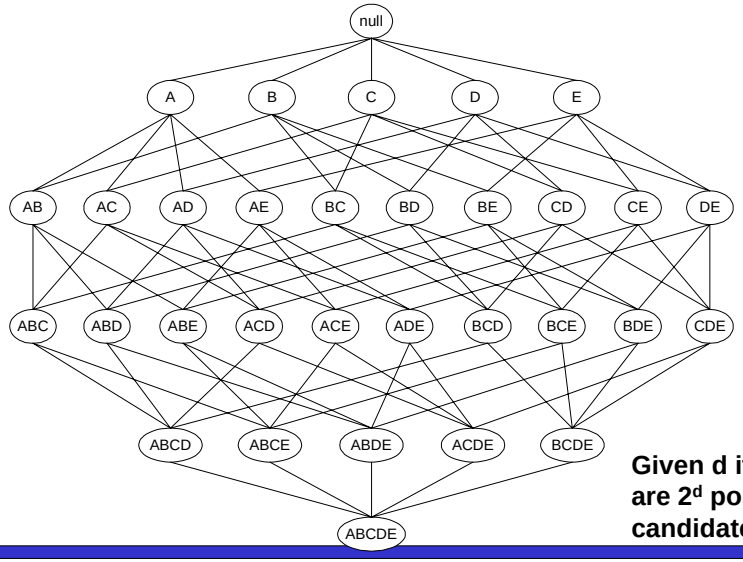
$Support(P \rightarrow Q) = Probability(P \cup Q)$
$Confidence(P \rightarrow Q) = Probability(Q/P)$

Association Rule Mining



- Frequent itemset 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

Transactions

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

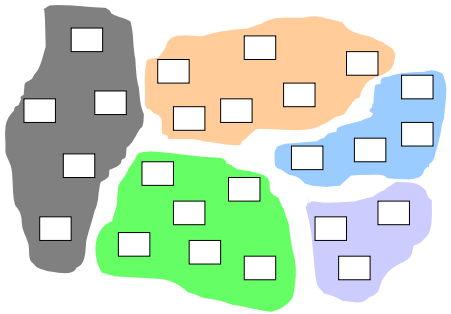
N (rows) × w (items)

List of Candidates — M

  - 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?**

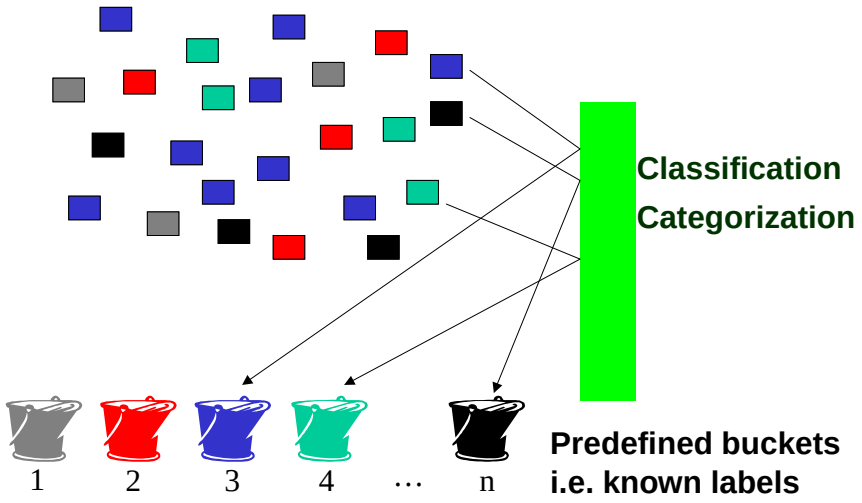
Grouping

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

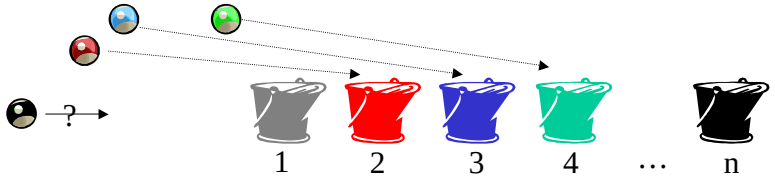


What is Classification?

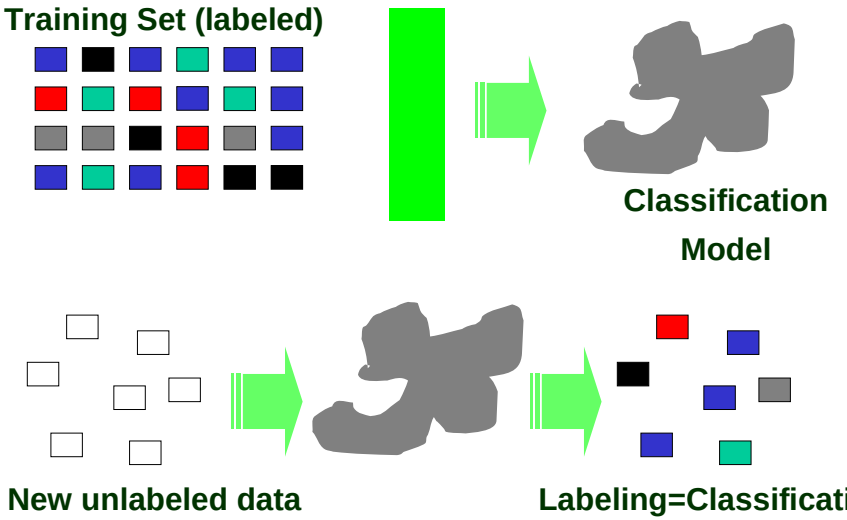
The goal of data classification is to organize and categorize data in distinct classes.

- 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.

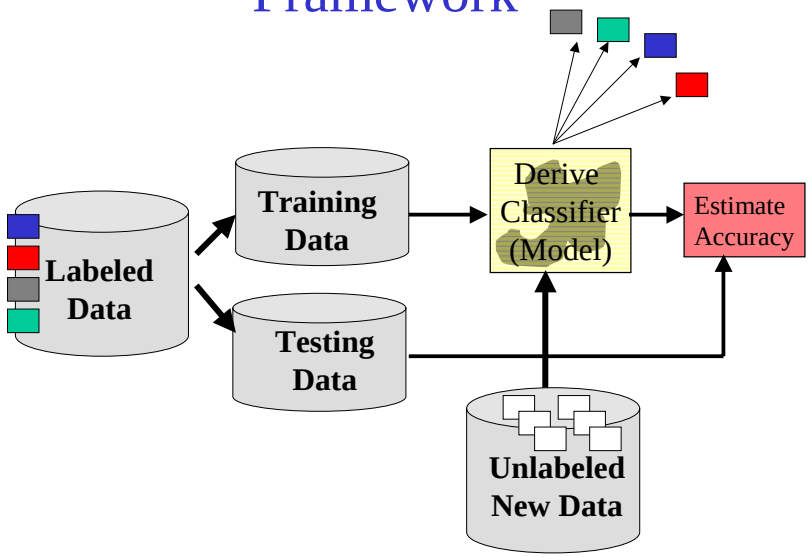
With classification, I can predict in which bucket to put the ball, but I can't predict the weight of the ball.



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.