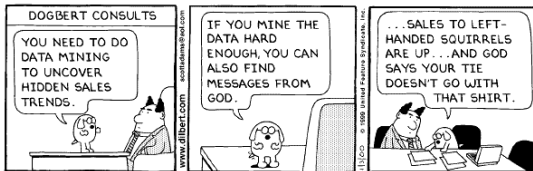


C(I)S 330: Applied Database Systems

A Break: A Mini-Introduction to Data Mining

(Some slides courtesy of Rich Caruana)

What Is Data Mining?



Copyright © 2000 United Feature Syndicate, Inc.
Redistribution in whole or in part prohibited

Definition

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.

Example pattern (Census Bureau Data):
If (relationship = husband), then (gender = male). 99.6%

Definition (Cont.)

Data mining is the exploration and analysis of large quantities of data in order to discover valid, novel, potentially useful, and ultimately understandable patterns in data.

Valid: The patterns hold in general.

Novel: We did not know the pattern beforehand.

Useful: We can devise actions from the patterns.

Understandable: We can interpret and comprehend the patterns.

Why Use Data Mining Today?

Human analysis skills are inadequate:

- Volume and dimensionality of the data
- High data growth rate

Availability of:

- Data
- Storage
- Computational power
- Off-the-shelf software
- Expertise

An Abundance of Data

- Supermarket scanners, POS data
- Preferred customer cards
- Credit card transactions
- Direct mail response
- Call center records
- ATM machines
- Demographic data
- Sensor networks
- Cameras
- Web server logs
- Customer web site trails

Evolution of Database Technology

- 1960s: IMS, network model
- 1970s: The relational data model, first relational DBMS implementations
- 1980s: Maturing RDBMS, application-specific DBMS, (spatial data, scientific data, image data, etc.), OODBMS
- 1990s: Mature, high-performance RDBMS technology, parallel DBMS, terabyte data warehouses, object-relational DBMS, middleware and web technology
- 2000s: High availability, zero-administration, seamless integration into business processes
- 2010: Sensor database systems, databases on embedded systems, P2P database systems, large-scale pub/sub systems, ???

Computational Power

- Moore's Law:
In 1965, Intel Corporation cofounder Gordon Moore predicted that the density of transistors in an integrated circuit would double every year. (Later changed to reflect 18 months progress.)
- Experts on ants estimate that there are 10^{16} to 10^{17} ants on earth. In the year 1997, we produced one transistor per ant.



Much Commercial Support

- Many data mining tools
 - <http://www.kdnuggets.com/software>
- Database systems with data mining support
- Visualization tools
- Data mining process support
- Consultants

Why Use Data Mining Today?

Competitive pressure!

"The secret of success is to know something that nobody else knows."

Aristotle Onassis

- Competition on service, not only on price (Banks, phone companies, hotel chains, rental car companies)
- Personalization, CRM
- The real-time enterprise
- "Systemic listening"
- Security, homeland defense

The Knowledge Discovery Process

Steps:

1. Identify business problem
2. Data mining
3. Action
4. Evaluation and measurement
5. Deployment and integration into businesses processes

Data Mining Step in Detail

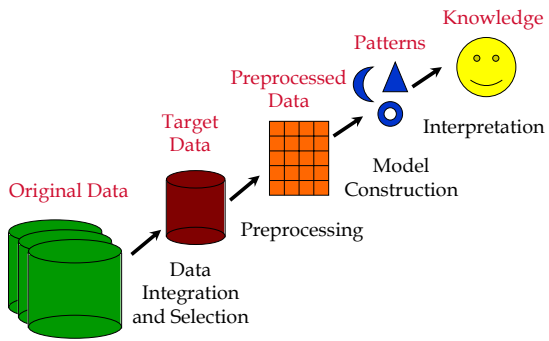
2.1 Data preprocessing

- Data selection: Identify target datasets and relevant fields
- Data cleaning
 - Remove noise and outliers
 - Data transformation
 - Create common units
 - Generate new fields

2.2 Data mining model construction

2.3 Model evaluation

Preprocessing and Mining



Example Application: Sports

IBM Advanced Scout analyzes NBA game statistics

- Shots blocked
 - Assists
 - Fouls
- Google: "IBM Advanced Scout"



Advanced Scout

- Example pattern: An analysis of the data from a game played between the New York Knicks and the Charlotte Hornets revealed that "*When Glenn Rice played the shooting guard position, he shot 5/6 (83%) on jump shots.*"
- Pattern is interesting:
The average shooting percentage for the Charlotte Hornets during that game was 54%.



Example Application: Sky Survey

- Input data: 3 TB of image data with 2 billion sky objects, took more than six years to complete
- Goal: Generate a catalog with all objects and their type
- Method: Use decision trees as data mining model
- Results:
 - 94% accuracy in predicting sky object classes
 - Increased number of faint objects classified by 300%
 - Helped team of astronomers to discover 16 new high red-shift quasars in one order of magnitude less observation time

Gold Nuggets?

- Investment firm mailing list: Discovered that old people do not respond to IRA mailings
- Bank clustered their customers. One cluster: Older customers, no mortgage, less likely to have a credit card
- "Bank of 1911"
- Customer churn example

What is a Data Mining Model?

A data mining model is a description of a specific aspect of a dataset. It produces output values for an assigned set of input values.

Examples:

- Linear regression model
- Classification model
- Clustering

Data Mining Models (Contd.)

A data mining model can be described at two levels:

- Functional level:
 - Describes model in terms of its intended usage.
Examples: Classification, clustering
- Representational level:
 - Specific representation of a model.
Example: Log-linear model, classification tree, nearest neighbor method.
- Black box models versus transparent models

Data Mining: Types of Data

- Relational data and transactional data
- Spatial and temporal data, spatio temporal observations
- Time series data
- Text
- Images, video
- Mixtures of data
- Sequence data

- Features from processing other data sources

Types of Variables

- *Numerical*: Domain is ordered and can be represented on the real line (e.g., age, income)
- *Nominal* or *categorical*: Domain is a finite set without any natural ordering (e.g., occupation, marital status, race)
- *Ordinal*: Domain is ordered, but absolute differences between values is unknown (e.g., preference scale, severity of an injury)

Data Mining Techniques

- Supervised learning
 - Classification and regression
- Unsupervised learning
 - Clustering
- Dependency modeling
 - Associations, summarization, causality
- Outlier and deviation detection
- Trend analysis and change detection

Supervised Learning

- $F(x)$: true function (usually not known)
- D : training sample drawn from $F(x)$

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0	1
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0	0
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,1,0,0	1
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0	0
40,M,205,0,115,90,37,18,0	0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0	1
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,1,1	0

Supervised Learning

- $F(x)$: true function (usually not known)
- D : training sample $(x,F(x))$

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	0
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0	1

- $G(x)$: model learned from D
- Goal: $E[(F(x)-G(x))^2]$ is small (near zero) for future samples

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0	?
--	---

Supervised Learning

Well-defined goal:

Learn $G(x)$ that is a good approximation to $F(x)$ from training sample D

Well-defined error metrics:

Accuracy, RMSE, ROC, ...

Supervised Learning

Training dataset:

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0	1
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0	0
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,1,0,0	1
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0	0
40,M,205,0,115,90,37,18,0	0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0	1
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,1	0

Test dataset:

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0

?

Un-Supervised Learning

Training dataset:

57,M,195,0,125,95,39,25,0,1,0,0,0,1,0,0,0,0,0,0,1,1,0,0,0,0,0,0	0
78,M,160,1,130,100,37,40,1,0,0,0,1,0,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0	1
69,F,180,0,115,85,40,22,0,0,0,0,0,1,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0	0
18,M,165,0,110,80,41,30,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0	1
54,F,135,0,115,95,39,35,1,1,0,0,0,1,0,0,0,1,0,0,0,0,1,0,0,0,1,0,0,0	1
84,F,210,1,135,105,39,24,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0	0
89,F,135,0,120,95,36,28,0,0,0,0,0,0,0,0,0,0,0,1,1,0,0,0,0,0,1,0,0	1
49,M,195,0,115,85,39,32,0,0,0,1,1,0,0,0,0,0,0,1,0,0,0,0,1,0,0,0	0
40,M,205,0,115,90,37,18,0	0
74,M,250,1,130,100,38,26,1,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0	1
77,F,140,0,125,100,40,30,1,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,1,1	0

Test dataset:

71,M,160,1,130,105,38,20,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0

?

Classification Example

- Example training database

- Two predictor attributes: Age and Car-type (**S**port, **M**inivan and **T**ruck)
- Age is ordered, Car-type is categorical attribute
- Class label indicates whether person bought product
- Dependent attribute is *categorical*

Age	Car	Class
20	M	Yes
30	M	Yes
25	T	No
30	S	Yes
40	S	Yes
20	T	No
30	M	Yes
25	M	Yes
40	M	Yes
20	S	No

Regression Example

- Example training database

- Two predictor attributes: Age and Car-type (**S**port, **M**inivan and **T**ruck)
- Spent indicates how much person spent during a recent visit to the web site
- Dependent attribute is *numerical*

Age	Car	Spent
20	M	\$200
30	M	\$150
25	T	\$300
30	S	\$220
40	S	\$400
20	T	\$80
30	M	\$100
25	M	\$125
40	M	\$500
20	S	\$420

Types of Variables (Review)

- *Numerical*: Domain is ordered and can be represented on the real line (e.g., age, income)
- *Nominal* or *categorical*: Domain is a finite set without any natural ordering (e.g., occupation, marital status, race)
- *Ordinal*: Domain is ordered, but absolute differences between values is unknown (e.g., preference scale, severity of an injury)

Definitions

- Random variables X_1, \dots, X_k (*predictor variables*) and Y (*dependent variable*)
- X_i has domain $\text{dom}(X_i)$, Y has domain $\text{dom}(Y)$
- P is a probability distribution on $\text{dom}(X_1) \times \dots \times \text{dom}(X_k) \times \text{dom}(Y)$
Training database D is a random sample from P
- A *predictor* d is a function
 $d: \text{dom}(X_1) \times \dots \times \text{dom}(X_k) \rightarrow \text{dom}(Y)$

Classification Problem

- If Y is categorical, the problem is a *classification problem*, and we use C instead of Y .
 $|\text{dom}(C)| = J$.
- C is called the *class label*, d is called a *classifier*.
- Take r be record randomly drawn from P .
Define the *misclassification rate* of d :
 $RT(d, P) = P(d(r.X_1, \dots, r.X_k) \neq r.C)$
- **Problem definition:** Given dataset D that is a random sample from probability distribution P , find classifier d such that $RT(d, P)$ is minimized.

Regression Problem

- If Y is numerical, the problem is a *regression problem*.
- Y is called the dependent variable, d is called a *regression function*.
- Take r be record randomly drawn from P .
Define mean squared error rate of d :
 $RT(d, P) = E(r.Y - d(r.X_1, \dots, r.X_k))^2$
- **Problem definition:** Given dataset D that is a random sample from probability distribution P , find regression function d such that $RT(d, P)$ is minimized.

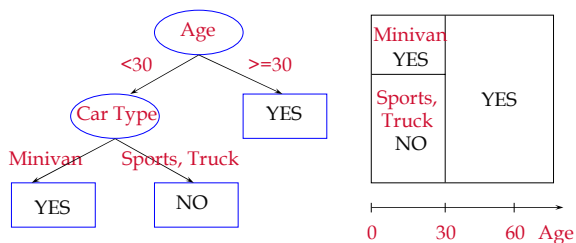
Goals and Requirements

- Goals:
 - To produce an accurate classifier/regression function
 - To understand the structure of the problem
- Requirements on the model:
 - High accuracy
 - Understandable by humans, interpretable
 - Fast construction for very large training databases

Different Types of Classifiers

- Linear discriminant analysis (LDA)
- Quadratic discriminant analysis (QDA)
- Density estimation methods
- Nearest neighbor methods
- Logistic regression
- Neural networks
- Fuzzy set theory
- Decision Trees

What are Decision Trees?



Decision Trees

- A *decision tree* T encodes d (a classifier or regression function) in form of a tree.
- A node t in T without children is called a *leaf node*. Otherwise t is called an *internal node*.

Internal Nodes

- Each internal node has an associated *splitting predicate*. Most common are binary predicates.

Example predicates:

- Age ≤ 20
- Profession in {student, teacher}
- $5000 \cdot \text{Age} + 3 \cdot \text{Salary} - 10000 > 0$

Internal Nodes: Splitting Predicates

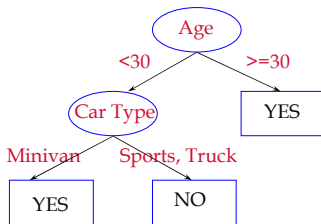
- Binary Univariate splits:
 - Numerical or ordered X : $X \leq c$, c in $\text{dom}(X)$
 - Categorical X : X in A , A subset $\text{dom}(X)$
- Binary Multivariate splits:
 - Linear combination split on numerical variables:
 $\sum a_i X_i \leq c$
- k -ary ($k > 2$) splits analogous

Leaf Nodes

Consider leaf node t

- Classification problem: Node t is labeled with one class label c in $\text{dom}(C)$
- Regression problem: Two choices
 - Piecewise constant model:
t is labeled with a constant y in $\text{dom}(Y)$.
 - Piecewise linear model:
t is labeled with a linear model
$$Y = y_t + \sum a_i X_i$$

Example



Encoded classifier:

If (age < 30 and
carType = Minivan)
Then YES

If (age < 30 and
(carType = Sports or
carType = Truck))
Then NO

If (age >= 30)
Then NO

Evaluation of Misclassification Error

Problem:

- In order to quantify the quality of a classifier d , we need to know its misclassification rate $RT(d, P)$.
- But unless we know P , $RT(d, P)$ is unknown.
- Thus we need to estimate $RT(d, P)$ as good as possible.

Resubstitution Estimate

The *Resubstitution estimate* $R(d,D)$ estimates $R_T(d,P)$ of a classifier d using D :

- Let D be the training database with N records.
- $R(d,D) = 1/N \sum I(d(r.X) \neq r.C)$
- Intuition: $R(d,D)$ is the proportion of training records that is misclassified by d
- Problem with resubstitution estimate:
Overly optimistic; classifiers that overfit the training dataset will have very low resubstitution error.

Test Sample Estimate

- Divide D into D_1 and D_2
- Use D_1 to construct the classifier d
- Then use resubstitution estimate $R(d,D_2)$ to calculate the estimated misclassification error of d
- Unbiased and efficient, but removes D_2 from training dataset D

V-fold Cross Validation

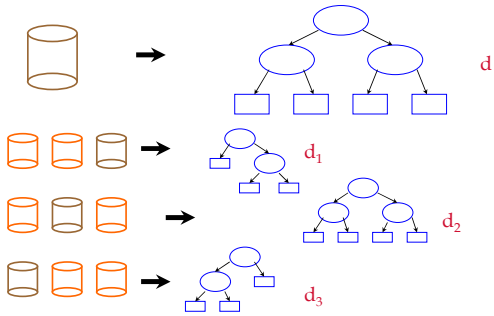
Procedure:

- Construct classifier d from D
- Partition D into V datasets D_1, \dots, D_V
- Construct classifier d_i using $D \setminus D_i$
- Calculate the estimated misclassification error $R(d_i, D_i)$ of d_i using test sample D_i

Final misclassification estimate:

- Weighted combination of individual misclassification errors:
 $R(d,D) = 1/V \sum R(d_i, D_i)$

Cross-Validation: Example



Cross-Validation

- Misclassification estimate obtained through cross-validation is usually nearly unbiased
- Costly computation (we need to compute d , and d_1, \dots, d_v); computation of d_i is nearly as expensive as computation of d
- Preferred method to estimate quality of learning algorithms in the machine learning literature

Decision Tree Construction

- Top-down tree construction schema:
 - Examine training database and find best splitting predicate for the root node
 - Partition training database
 - Recurse on each child node

Top-Down Tree Construction

BuildTree(Node t , Training database D , Split Selection Method S)

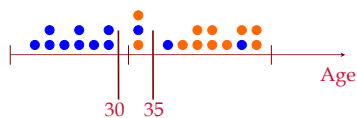
- (1) Apply S to D to find splitting criterion
- (2) **if** (t is not a leaf node)
- (3) Create children nodes of t
- (4) Partition D into children partitions
- (5) Recurse on each partition
- (6) **endif**

Decision Tree Construction

- Three algorithmic components:
 - Split selection (CART, C4.5, QUEST, CHAID, CRUISE, ...)
 - Pruning (direct stopping rule, test dataset pruning, cost complexity pruning, statistical tests, bootstrapping)
 - Data access (CLOUDS, SLIQ, SPRINT, RainForest, BOAT, UnPivot operator)

Split Selection Method

- Numerical or ordered attributes: Find a split point that separates the (two) classes



(Yes: ● No: ●)



Split Selection Method (Contd.)

- Categorical attributes: How to group?

Sport:  Truck:  Minivan: 

(Sport, Truck) - (Minivan)  

(Sport) - (Truck, Minivan)  

(Sport, Minivan) - (Truck)  

Pruning Method

- For a tree T , the misclassification rate $R(T,P)$ and the mean-squared error rate $R(T,P)$ depend on P , but not on D .
- The goal is to do well on records randomly drawn from P , not to do well on the records in D
- If the tree is too large, it overfits D and does not model P . The pruning method selects the tree of the right size.

Data Access Method

- Recent development: Very large training databases, both in-memory and on secondary storage
- Goal: Fast, efficient, and scalable decision tree construction, using the complete training database.

Decision Trees: Summary

- Many application of decision trees
- There are many algorithms available for:
 - Split selection
 - Pruning
 - Handling Missing Values
 - Data Access
- Decision tree construction still active research area (after 20+ years!)
- Challenges: Performance, scalability, evolving datasets, new applications

Market Basket Analysis

- Consider shopping cart filled with several items
- Market basket analysis tries to answer the following questions:
 - Who makes purchases?
 - What do customers buy together?
 - In what order do customers purchase items?

Market Basket Analysis

Given:

- A database of customer transactions
- Each transaction is a set of items
- Example:
Transaction with TID 111 contains items {Pen, Ink, Milk, Juice}

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

Market Basket Analysis (Contd.)

- Co-occurrences
 - 80% of all customers purchase items X, Y and Z together.
- Association rules
 - 60% of all customers who purchase X and Y also buy Z.
- Sequential patterns
 - 60% of customers who first buy X also purchase Y within three weeks.

Confidence and Support

We prune the set of all possible association rules using two interestingness measures:

- **Confidence** of a rule:
 - $X \Rightarrow Y$ has confidence c if $P(Y|X) = c$
- **Support** of a rule:
 - $X \Rightarrow Y$ has support s if $P(XY) = s$

We can also define

- **Support** of an itemset (a cocurrence) XY :
 - XY has support s if $P(XY) = s$

Example

Examples:

- $\{\text{Pen}\} \Rightarrow \{\text{Milk}\}$
Support: 75%
Confidence: 75%
- $\{\text{Ink}\} \Rightarrow \{\text{Pen}\}$
Support: 100%
Confidence: 100%

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

Exercise

- Find all itemsets with support $\geq 75\%$?

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

Exercise

- Can you find all association rules with support $\geq 50\%$?

TID	CID	Date	Item	Qty
111	201	5/1/99	Pen	2
111	201	5/1/99	Ink	1
111	201	5/1/99	Milk	3
111	201	5/1/99	Juice	6
112	105	6/3/99	Pen	1
112	105	6/3/99	Ink	1
112	105	6/3/99	Milk	1
113	106	6/5/99	Pen	1
113	106	6/5/99	Milk	1
114	201	7/1/99	Pen	2
114	201	7/1/99	Ink	2
114	201	7/1/99	Juice	4

Extensions

- Imposing constraints
 - Only find rules involving the dairy department
 - Only find rules involving expensive products
 - Only find "expensive" rules
 - Only find rules with "whiskey" on the right hand side
 - Only find rules with "milk" on the left hand side
 - Hierarchies on the items
 - Calendars (every Sunday, every 1st of the month)

Market Basket Analysis: Applications

- Sample Applications
 - Direct marketing
 - Fraud detection for medical insurance
 - Floor/shelf planning
 - Web site layout
 - Cross selling

Beyond Support and Confidence

Example: 5000 students

- 3000 students play basketball
- 3750 students eat cereal
- 2000 students both play basketball and eat cereal

	Basketball	No basketball	Sum
Cereal	2000	1750	3750
No cereal	1000	250	1250
Sum	3000	2000	5000

Misleading Association Rules

	Basketball	No basketball	Sum
Cereal	2000	1750	3750
No cereal	1000	250	1250
Sum	3000	2000	5000

- Basketball => Cereal (support: 40%, confidence: 66.7%) is misleading because 75% of students eat cereal
- Basketball => No cereal (support: 20%, confidence: 33.3%) is more interesting, although with lower support and confidence

Interest

Interest of rule $A \Rightarrow B$: $P(AB)/(P(A)*P(B))$

- Symmetric (uses both $P(A)$ and $P(B)$)
- Note that confidence is not symmetric (confidence of rule $A \Rightarrow B$: $P(AB)/P(A)$)

Interest values:

- Interest = 1: A and B are independent ($P(AB)=P(B)*P(A)$)
- Interest > 1: A and B are positively correlated
- Interest < 1: A and B are negatively correlated

Interest: Example

	Basketball	No basketball	Sum
Cereal	2000	1750	3750
No cereal	1000	250	1250
Sum	3000	2000	5000

Itemset	Support	Interest
Cereal, basketball	40%	1.125
Cereal, no basketball	35%	0.857
No cereal, basketball	20%	0.750
No cereal, no basketball	5%	2.000
