University of Colorado
Boulder

CSCI 4502/5502
Data Mining

Fall 2019
Lecture 07 (Sep 17)

Announcements

- Homework 2
  - due at 9:30am, Thursday, Sep 19
- Office hours
  - Tu 11-12pm, Fr 1-2pm (Instructor)
  - Tu 4-5pm (TA), W 11-12pm (GSS)
- Course project
  - team, project idea, data sets
  - project proposal: Week 7

Review (1)

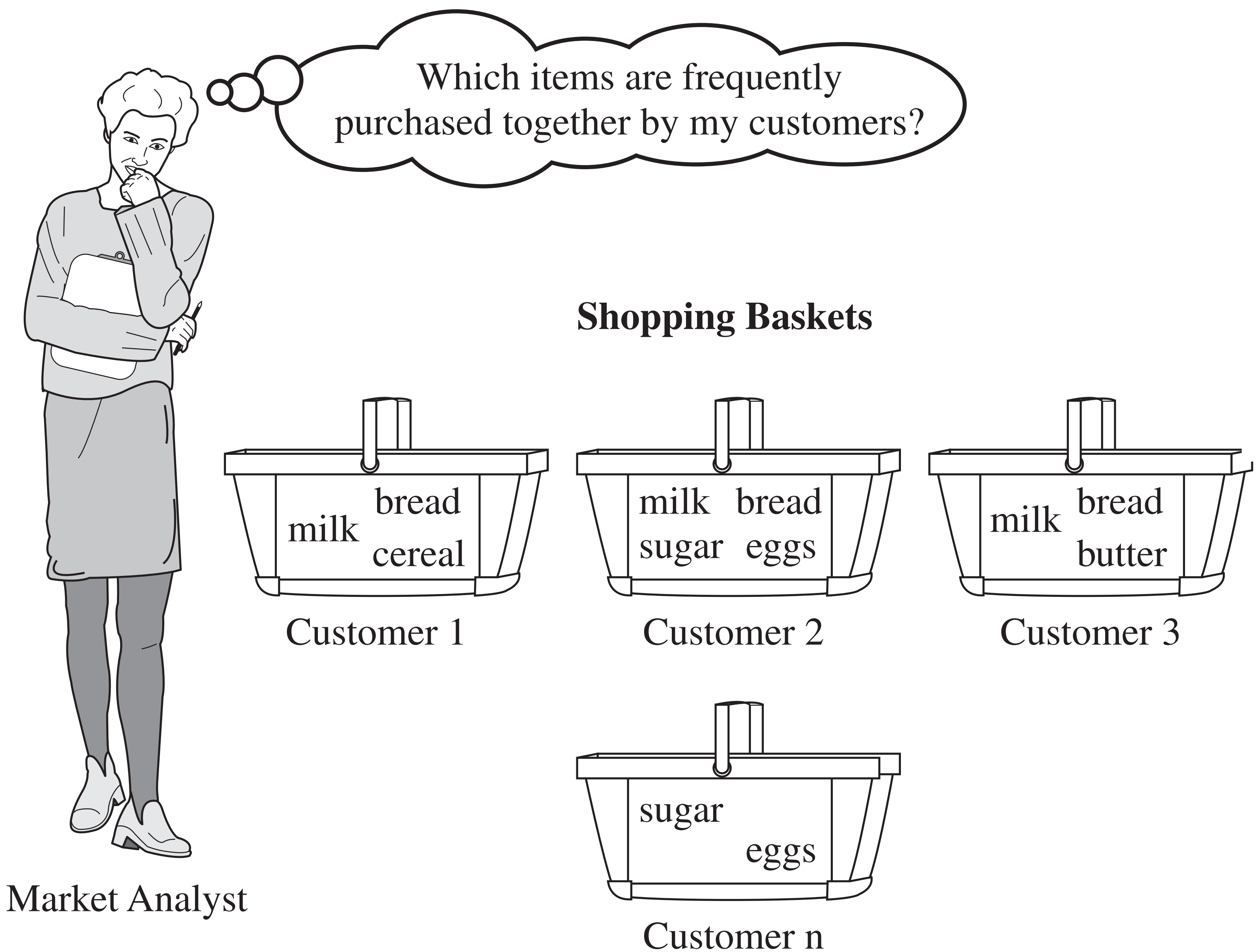
- Chap 1: Introduction to data mining
  - why? data? knowledge? methods? app?
- Chap 2: Getting to know your data
  - central tendency, dispersion
  - attribute types, similarity/dissimilarity
- Chap 3 : Data preprocessing
  - cleaning, integration, reduction, transformation & discretization

Review (2)

- Chap 4 & 5: Data Warehouse, Data Cube
  - what is data warehouse?
  - OLTP vs. OLAP
  - what is data cube?
  - data cube operations
  - data cube computation

Chapter 6:
Mining Frequent Patterns, Associations & Correlations

Market Basket Analysis



http://www.information-drivers.com/images/beer_and_baby.gif

Frequent Pattern Analysis

- Frequent patterns in a data set
  - a set of items
  - subsequences
  - substructures
- Other examples?
  - Web log
  - Road traffic

Basic Concepts

- Frequent itemset
  - $X = \{x_1, x_2, ..., x_k\}$
- Association rule $X \Rightarrow Y$
  - support: probability that a transaction contains $X \cup Y$
  - confidence: conditional probability that a transaction containing X also contains Y
  - minimum support, minimum confidence

Example

- Let min_sup = 50%, min_conf = 50%
- Frequent patterns
  - A , B , D , E , AD
- Association rules
  - A ⇒ D (%, %)
  - D ⇒ A (%, %)

Tid	Items
1	A, B, D
2	A, C, D
3	A, D, E
4	B, E, F
5	B, C, D, E, F

Example

- Let min_sup = 50%, min_conf = 50%
- Frequent patterns
  - A 3, B 3, D 4, E 3, AD 3
- Association rules
  - A $\Rightarrow$ D (60 %, 100 %)
  - D $\Rightarrow$ A (60 %, 75 %)

Tid	Items
1	A, B, D
2	A, C, D
3	A, D, E
4	B, E, F
5	B, C, D, E, F

Mining Association Rules

- Two-step process
  - find all frequent itemsets (w/ min_sup)
  - generate strong association rules from the frequent itemsets (min_sup, min_conf)
- A long pattern contains a combinatorial number of subpatterns (e.g., 100 items)

$$\binom{100}{1} + \binom{100}{2} + \cdots + \binom{100}{100} = 2^{100} - 1 \approx 1.27 \times 10^{30}$$

Closed & Max Patterns

- Solution: mine closed patterns & max-patterns
- Closed pattern X
  - no super-pattern $Y \supset X$ w/ the same support
- Max-pattern X
  - no super-pattern $Y \supset X$
- Closed pattern is a lossless compression of frequent patterns
  - reducing the number of patterns and rules

Example

- $\{<a_1, \ldots, a_{100}>, <a_1, \ldots, a_{50}>\}$, min_sup = 0.5
- Frequent pattern?
  - all item combinations
- Closed pattern?
  - $<a_1, \ldots, a_{100}>$: 1
  - $<a_1, \ldots, a_{50}>$: 2
- Max-pattern?
  - $<a_1, \ldots, a_{100}>$: 1

Apriori Algorithm (1)

- Apriori property
  - subset of a freq. itemset is also frequent
  - e.g., {beer, diaper, nuts}, {beer, diaper}
- Apriori pruning
  - if X is infrequent,
  - then superset of X is pruned

Apriori Algorithm (2)

- Procedure
  - 1. scan DB to get freq. 1-itemset
  - 2. generate candidate (k+1)-itemsets from freq. k-itemsets
  - 3. test candidate (k+1)-itemsets against DB
  - 4. stop when no freq. or candidate itemsets can be generated

Apriori Algorithm: Example

Tid	Items
1	A, C, D
2	B, C, E
3	A, B, C, E
4	B, E

min_sup = 0.5

Itemset	sup

Itemset	sup

Itemset	sup

Apriori Algorithm: Example

Tid	Items
1	A, C, D
2	B, C, E
3	A, B, C, E
4	B, E

min_sup =0.5

Itemset	sup
{A}	0.5
{B}	0.75
{C}	0.75
{D}	0.25
{E}	0.75

Itemset	sup
{A, B}	0.25
{A, C}	0.5
{A, E}	0.25
{B, C}	0.5
{B, E}	0.75
{C, E}	0.5

Itemset	sup
{B, C, E}	0.5

Important Details

- Self-joining of k-itemsets to generate (k+1)-itemsets
  - two k-itemsets are joined if their first (k-1) items are the same
- Pruning: remove if subset not frequent
- Example: L3 = {abc, abd, acd, ace, bcd}
  - abc and abd => abcd
  - acd and ace => acde
  - acde pruned because ade is not in L3

Interestingness Measure

- Association rule
  - A ⇒ B [support, confidence]
- A strong association rule
  - play basketball ⇒ eat cereal [40%, 66.7%]
- The rule is misleading
  - overall, 75% of students eat cereal
  - play basketball ⇒ not eat cereal [20%, 33.3%]

Correlation Rules

- Correlation rule
  - A ⇒ B [support, confidence, correlation]
- Measure of dependent/correlated events

$$lift(A,B) = \frac{P(A \cup B)}{P(A)P(B)}$$

  - lift = 1? independent
  - lift < 1? negatively dependent
  - lift > 1? positively dependent

$$lift(A,B) = \frac{P(A \cup B)}{P(A)P(B)}$$

	basketball	not basketball	sum (row)
cereal	2000	1750	3750
not cereal	1000	250	1250
sum (col)	3000	2000	5000

$$lift(B,C) = \frac{2000/5000}{(3000/5000) \times (3750/5000)} = 0.89$$

$$lift(B,\overline{C}) = \frac{1000/5000}{(3000/5000) \times (1250/5000)} = 1.33$$