CARLSON SCHOOL
OF MANAGEMENT
UNIVERSITY OF MINNESOTA

IDSC 4444 (004)
Association Rules

Zihong Huang
Information & Decision Sciences
Carlson School of Management
huan0707@umn.edu

Agenda

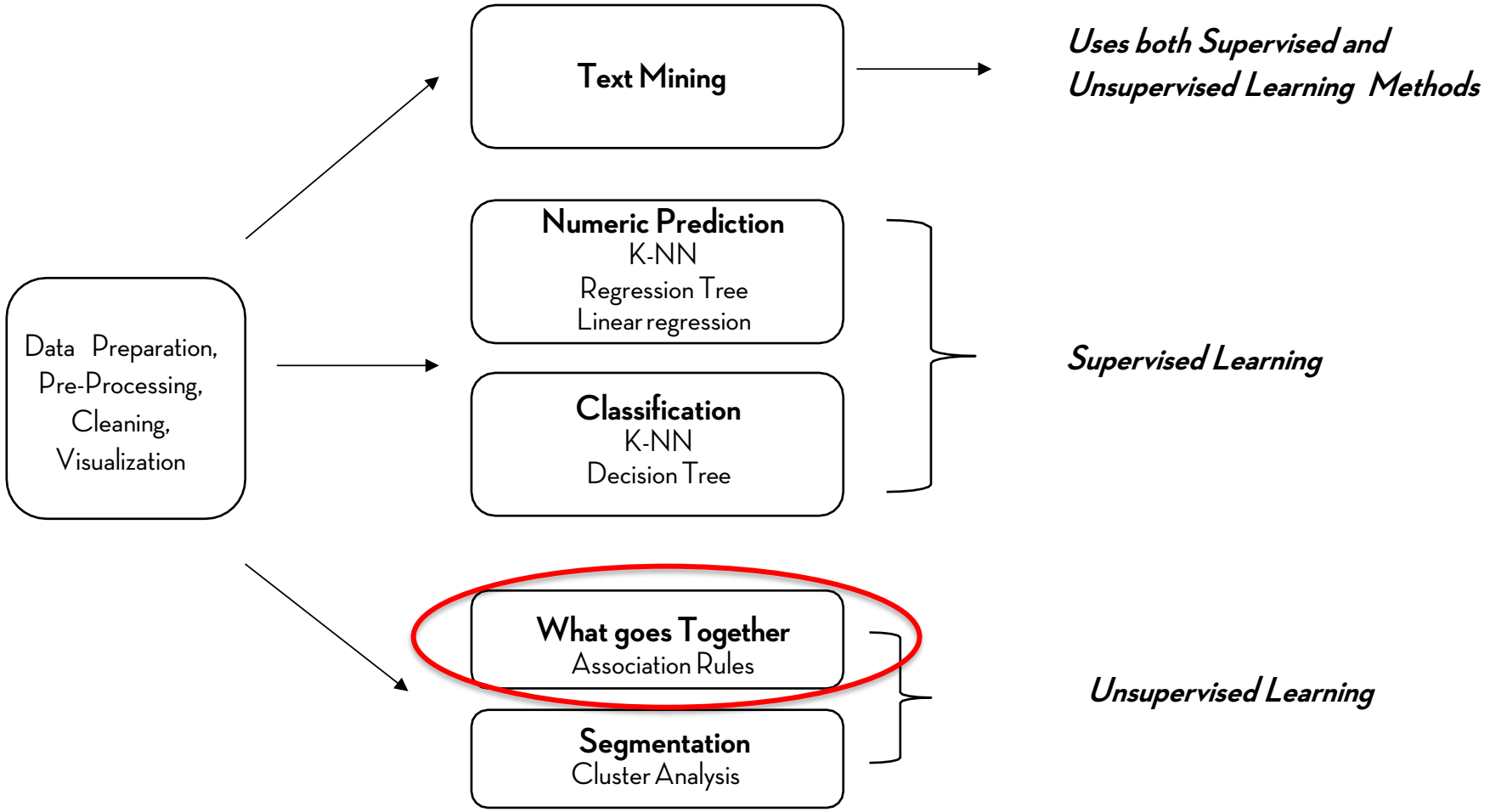
- Supervised vs. Unsupervised
- Basic Definitions of Association Rules
- Measurement of Association Rules
  - Support
  - Confidence
  - Lift
- How to find association rules?
  - Apriori Algorithm

Tentative Schedule

- Check it out on Canvas

Week	Date	Topic	HW/Quiz Posted	HW/Quiz Due By 11:59 pm
1	10/26/2021 (Tu)	Lec - Course Introduction & Visualization	HW1	
1	10/28/2021 (Th)	Lab - Working in R - Tutorial		
2	11/02/2021 (Tu)	Lec - Descriptive Analysis 1: Association Rules	Quiz 1 (Association Rule)	
2	11/04/2021 (Th)	Lab - Association Rules	HW 2	HW 1
2	11/05/2021 (Fri)			Quiz 1 (Association Rule)
3	11/9/2021 (Tu)	Lec - Descriptive Analysis 2: Cluster Analysis	Quiz 2 (Cluster Analysis)	
3	11/11/2021 (Th)	Lab - Cluster Analysis	HW 3	HW 2

An Overview


Text Mining
Uses both Supervised and Unsupervised Learning Methods
Numeric Prediction
K-NN
Regression Tree
Linear regression
Classification
K-NN
Decision Tree
Supervised Learning
Data Preparation, Pre-Processing, Cleaning, Visualization
What goes Together
Association Rules
Segmentation
Cluster Analysis
Unsupervised Learning

Two Types of Learning

❑ Unsupervised Learning

- Unlabeled Data: there is no outcome variable to predict or classify
- Data is mined for patterns in the hopes of discovering useful patterns
- Methods we will cover: Association Rules, Cluster Analysis

❑ Supervised Learning

- Labeled Data: known outcomes like purchase decisions, price of goods, etc.
- Model can be tested against known outcomes for performance.
- Methods we will cover: Classification, Numeric Prediction

Supervised vs. Unsupervised Problems

- Will this customer purchase service?
- What services are commonly purchased together by the customers?
- Which service package ($S1$, $S2$, or none) will a customer likely to purchase?
- How much money will this customer spend on the service?
- Are there groups of similar customers within the data?

Supervised vs. Unsupervised Problems

Question	Unsupervised	Supervised	Technique
Will this customer purchase service?		√	Classification (binary target variable)
What services are commonly purchased together by the customers?	√		Association analysis (No target variable)
Which service package ($S1$, $S2$, or none) will a customer likely to purchase?		√	Classification (three valued target variable)
How much money will this customer spend on the service?		√	Regression (numeric target variable)
Are there groups of similar customers within the data?	√		Clustering (No target variable)

What is Association Rules Mining?

- Discovering interesting relationships among items/events/variables
- Also known as **Market Basket Analysis** and **Affinity Analysis**
  - Popular in Marketing, used to find out which products tend to be purchased together
  - Also applied in many domains, e.g., healthcare, bio-informatics,...
- It is a type of **exploratory** data analytics

An Example

Software

The parable of the beer and diapers

Never let the facts get in the way of a good story

By Mark Whitehorn 15 Aug 2006 at 13:20 5 SHARE ▼

https://www.theregister.com/2006/08/15/beer_diapers/

Basic Definitions

- Let **U** be the universal set of **Items** in a given domain
  - E.g., **U = {Milk, Eggs, Bread, Coke, Beer,...}** or all items in a grocery store
- An **Itemset, X,** is any subset of **U**
  - E.g., **X={Bread, Milk}** is a 2-items itemset
- A **Transaction** is an instance of consumption by one consumer
  - Multiple transactions comprise a dataset

Each row is a transaction →

#	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

Basic Definitions

- An **Association Rule** describes relationship between two itemsets: **X→Y**
  - This association rule reads "**if X then Y**"
  - X and Y are two **non-overlapping** itemsets (they don't share any item in common)
  - X is called **antecedent** (or left-hand-side, or body)
  - Y is called **consequent** (or right-hand-side, or head)
  - In the shopping context, this means: "customers who buy X are also likely to buy Y"
    - ✓ Consider an association rule {Bread} →{Milk}
    - ✓ It means **customers who buy Bread are likely to buy Milk too**

Support

- Consider the following association rule: X→Y
  - X = {Milk, Diapers}, Y = {Coke}
- **Support**: a measure of how frequently X and Y occur together
  - **Support Count (σ)**: raw count of transactions containing both X and Y
    - ✓ σ(X→Y) = Count(X and Y) = σ({Milk, Diapers, Coke}) = 2
  - **Support Percentage (S)**: **Fraction** of transactions containing both X and Y
    - ✓ $S(X \rightarrow Y) = S(\{Milk, Diapers, Coke\}) = \frac{\sigma(X \rightarrow Y)}{\#Transactions} = 2/5 = 0.4$
    - ✓ **Note:** Support metrics are **NOT directional**, i.e., S(X→Y) is equivalent to S(Y→X)
  - **Support of individual itemsets:**
    - ✓ σ(X) = σ({Milk, Diapers}) = 3, S(X) = σ({Milk, Diapers}) /#Transactions = 3/5 = 0.6
    - ✓ σ(Y) = σ({Coke}) = 2, S(Y) = σ ({Coke}) /#Transactions = 2/5 = 0.4

Basket	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

Confidence

- **"Among all transactions containing X, how many also have Y?"**
  - X→Y: X = {Milk, Diapers}, Y = {Coke}
- **Confidence**: a measure of how often Y appears with transactions that contain X
  - $\text{Conf}(X \rightarrow Y) = \frac{S(X \rightarrow Y)}{S(X)} = \frac{\sigma(X \rightarrow Y)}{\sigma(X)} = \frac{\sigma(\{\text{Milk, Diapers, Coke}\})}{\sigma(\{\text{Milk, Diapers}\})} = \frac{2}{3} = 0.67$
  - Conceptually related to conditional probability Pr(Y|X)
  - When a customer buys milk and diapers, 67% of the time also buys coke
- **Important**: this measure is **directional**
  - i.e., Conf (X→Y) is **not necessarily equivalent** to Conf (Y→X)
  - $\text{Conf}(Y \rightarrow X) = \frac{\sigma(Y \rightarrow X)}{\sigma(Y)} = \frac{\sigma(\{\text{Milk, Diapers, Coke}\})}{\sigma(\{\text{Coke}\})} = \frac{2}{2} = 1$
  - When a customer buys coke, 100% of the time buys milk and diapers

Basket	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

Lift

- **Support and Confidence**: both are measures of how strong a rule is.
- Consider the following situation: in a supermarket, **90%** of all customers buy milk, and **95%** of all customers buy toilet paper.
  - By pure chance, 85% (0.9 * 0.95) of all customers buy milk and toilet paper
  - Real association between them or just coincidence?
- **Lift**: a measure of how much more likely X and Y co-occur than pure chance
  - $\text{Lift}(X \rightarrow Y) = \frac{S(X \rightarrow Y)}{S(X) * S(Y)} = \frac{\text{Conf}(X \rightarrow Y)}{S(Y)}$
  - Here, we must use the support percentage in calculation
  - $S(X) * S(Y)$ is the probability of seeing X co-occurring with Y by pure chance, i.e., X and Y are independent
- Note: Lift has no direction

Lift

- Consider the following association rule: X→Y
  - X = {Milk, Diapers}, Y = {Beer}
- S({Milk, Diapers, Beer}) = 2/5 = 0.4
- S({Milk, Diapers}) = 3/5 = 0.6
- S({Beer}) = 3/5 = 0.6
- Lift (X→Y) = $\frac{0.4}{0.6*0.6} = 1.11$
  - Customers who buy Milk and Diapers are 1.11x more likely to buy Beer than other customers

Basket	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

Lift

When Lift > 1 means that customers who buy X are more likely to buy Y than other customers

When Lift = 1 means that customers who buy X are as likely to buy Y than any other customers

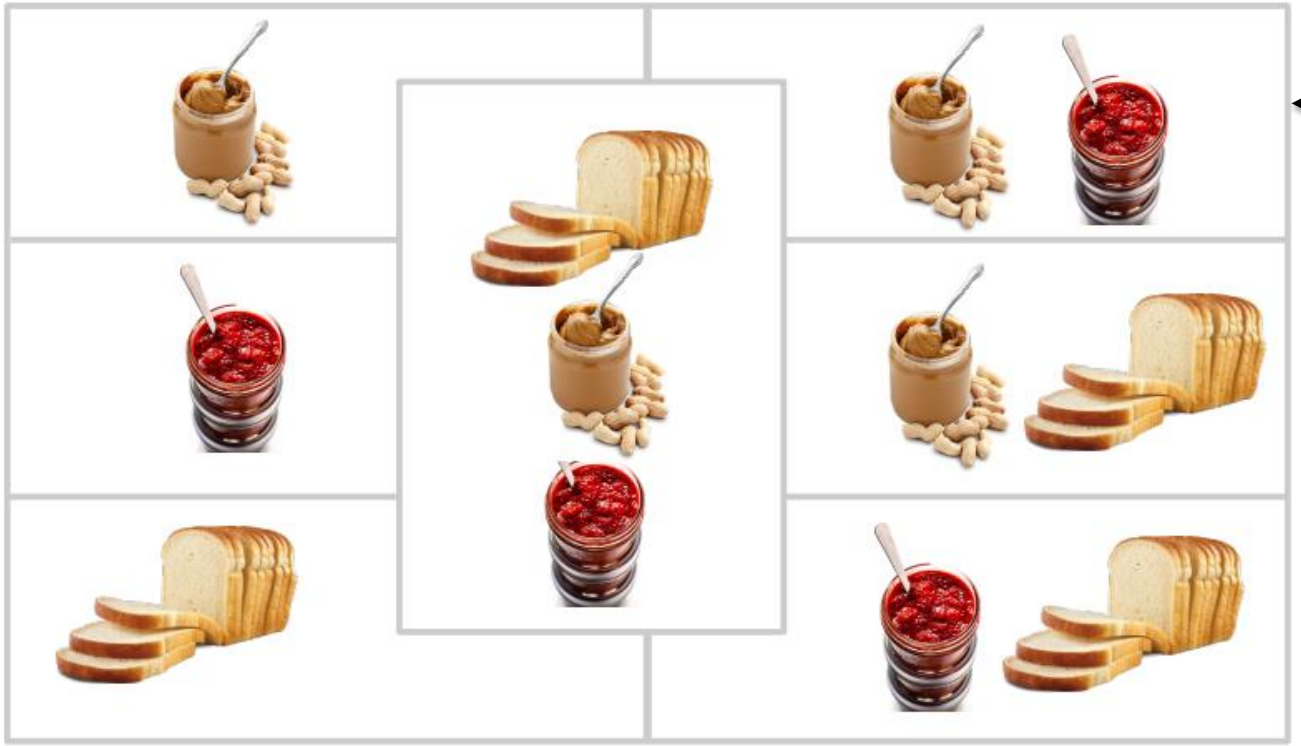
When Lift < 1 means that customers who buy X are less likely to buy Y than other customers

How to Find Association Rules

- We specify the minimum support **(minsupp)** and minimum confidence **(minconf)**
  - Find all association rules: support > = minsupp and confidence >= minconf
  - minsupp and minconf are picked based on domain knowledge or business goals.
- **Step 1:** Find all itemsets with **support >= minsupp**
  - These are called **frequent itemsets**
- **Step 2:** Based on each frequent itemset, generate all possible association rules, then keep the ones with **confidence >= minconf**
  - Note: for a frequent itemset {coffee, bagel}, we need to consider two rules: {coffee} →{bagel} and {bagel} →{coffee}

Practical Concerns

3 Items, how many possible itemsets?



$2^3 - 1 = 7$ Itemsets

- How about N? $2^N - 1$ potential itemsets (exponential)
  - E.g., 50 items, check 1000 itemsets per second, would take ~35000 years!

Apriori Algorithm

- **Apriori Algorithm** (Agrawal et al., 1993): A smart way to reduce burden
  - **Key idea:** if an itemset X is NOT frequent, then any larger itemsets containing X cannot be frequent
- **Steps:**
  - First check all 1-item itemsets, only keep the frequent ones (support >= minsupp)
  - Then check 2-items itemsets made from frequent 1-itemsets in previous step, only keep the frequent ones
  - Keep going recursively until you have checked frequent itemsets of all sizes
  - Among all frequent itemsets, generate all possible association rules
  - Find association rules that satisfy confidence >= minconf

Apriori Algorithm: Example

- Assume we want a support => 75% and confidence => 80%

Transaction	Items
T_1	K, A, D, B
T_2	D, A, C, E, B
T_3	C, A, B, E
T_4	B, A, D

- To make things easier, we can do a tabular representation of the data:

Transaction	A	B	C	D	E	K
T1	1	1	0	1	0	1
T2	1	1	1	1	1	0
T3	1	1	1	0	1	0
T4	1	1	0	1	0	0

What to do with Association Rules

- You find an association rule {beer} →{diaper} and conclude it is strong enough, Now what?
- **Possible Marketing Actions**
  - Put diapers next to beer in your store
  - Or, put diapers away from beer in your store (Why?)
  - Bundle beer and diapers in a "new parent coping kit"
  - Lower the price of diapers, raise it on beer

- **Remember:** Association rules are exploratory in nature
  - They provide some initial directions to work on.
  - Setting specific business strategies requires domain expertise and careful analysis and testing

Exercise 1

Using the dataset to the right, compute the support percentage and the confidence of the association rules listed below.

	Dataset
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

Association Rule	Support (s)	Confidence (c)
{Milk, Diapers} → {Beer}		
{Milk, Beer} → {Diapers}		
{Diapers} → {Milk, Beer}		
{Beer} → {Milk, Diapers}		
{Diapers, Beer} → {Milk}		
{Milk} →{Diapers, Beer}		

Exercise 2

Basket	Items
1	Bread, Milk
2	Bread, Diapers, Beer, Eggs
3	Milk, Diapers, Beer, Coke, Bread
4	Bread, Milk, Diapers, Beer
5	Bread, Milk, Diapers, Coke

- Consider the dataset above and calculate the **confidence** and the **lift** of the following association rules. What would you conclude about these association rules?
- {Diapers} → {Eggs}
- {Bread} → {Milk}

Before Next Class

- Do the 2 exercises in this slides, Quiz related
  - The solution will be posted on Canvas
- Check the additional materials on Canvas in case that you need them
  - Relevant textbook chapters and Apriori Algorithm related materials
  - Additional links
- Download two *.zip files for Lab on Thursday
- HW1 is due on Thursday
- Quiz 1 is posted and due on Friday

Questions?