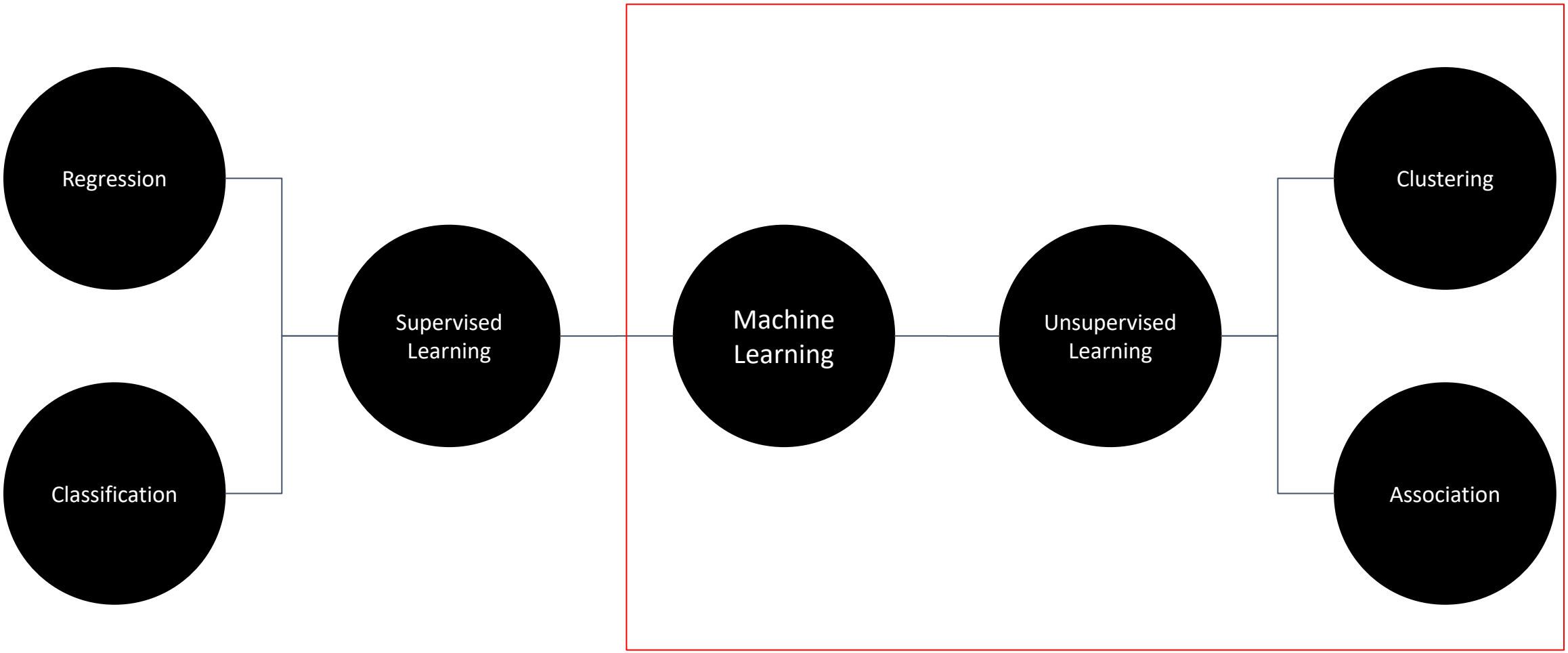
Data Prediction Model and Machine Learning

Online course #10

Market Basket Analysis


Regression
Classification
Supervised Learning
Machine Learning
Unsupervised Learning
Clustering
Association

Have you ever bought something unplanned in grocery stores?

i.e.) Have you ever bought gum or chocolate bars at the checkout line at the grocery store?

This impulsive purchase is not a coincidence.

This was possible because they used sophisticated data analysis techniques to find patterns that induce purchasing behavior.

- Barcode scanner records
- Inventory management
- Online finance records

Big data → Purchasing pattern → Recommendation

Market Basket Analysis

Understanding of the Association Rule

MBA

- Basic unit: Item
- One or more items constitute itemset
  - {bread, peanut butter, jelly)

LHS		RHS
{peanut butter, jelly}	→	{bread}

- If peanut butter and jelly are bought together, there is a possibility that bread will also be purchased.
- Peanut butter and jelly hint at bread

Understanding of the Association Rule

Not only for MBA but for…

- Discovery of interesting and frequent DNA patterns and protein sequences in cancer data
- Discover purchasing patterns arising from the use of fraudulent credit cards and insurance
- Implementation of a recommendation algorithm by discovering the pattern that people who watched drama “AAA” and “BBB” watch “CCC” together

Apriori

$$2^{k} \qquad 2^{100} = 1.27 \times 10^{30}$$

Normal PC can do the job, but still not plausible..

Let’s use the fact that some combinations are really non-sense like {motor oil, lipstick} {iphone, hammer}

Many studies have been conducted to find heuristic algorithms to reduce the number of item sets to be searched

Apriori: This is the most widely used method for efficiently finding rules in large databases

Apriori

Apriori: A priori

Use simple prior (pre-known or a priori) beliefs about the properties of frequent item sets

{motor oil, lipstick} combination is only considerable when..

1. {motor oil} is frequently chosen in the market basket
2. {lipstick} is frequently chosen in the market basket

Apriori. How it works?

transaction ID	items purchased
1	{flowers, get well card, soda}
2	{plush toy bear, flowers, balloons, candy bar}
3	{get well card, candy bar, flowers}
4	{plush toy bear, balloons, soda}
5	{flowers, get well card, soda}

Apriori. How it works?

Visitors to friends who are sick..

transaction ID	items purchased
1	{flowers, get well card, soda}
2	{plush toy bear, flowers, balloons, candy bar}
3	{get well card, candy bar, flowers}
4	{plush toy bear, balloons, soda}
5	{flowers, get well card, soda}

Apriori. How it works?

People who visit mothers tend to buy bears and balloons.

transaction ID	items purchased
1	{flowers, get well card, soda}
2	{plush toy bear, flowers, balloons, candy bar}
3	{get well card, candy bar, flowers}
4	{plush toy bear, balloons, soda}
5	{flowers, get well card, soda}

Apriori. How it works?

Support: How often it occurs in the data

$$support(X) = \frac{count(X)}{N}$$

Confidence: A measure of how much Y appears together when X appears

$$confidence(X \rightarrow Y) = \frac{support(X, Y)}{support(X)}$$

$$\frac{support(X, Y)}{support(X)} = \frac{count(X, Y)}{count(X)}$$

Apriori. How it works?

Support: How often it occurs in the data

transaction ID	items purchased
1	{flowers, get well card, soda}
2	{plush toy bear, flowers, balloons, candy bar}
3	{get well card, candy bar, flowers}
4	{plush toy bear, balloons, soda}
5	{flowers, get well card, soda}

Let's measure support of {get well card, flower}

$$support(X) = \frac{count(X)}{N}$$

$$support(\{gwc, flw\}) = \frac{3}{5} = 0.6$$

Apriori. How it works?

Support: How often it occurs in the data

transaction ID	items purchased
1	{flowers, get well card, soda}
2	{plush toy bear, flowers, balloons, candy bar}
3	{get well card, candy bar, flowers}
4	{plush toy bear, balloons, soda}
5	{flowers, get well card, soda}

Let's measure support of {candy bar}

$$support(X) = \frac{count(X)}{N}$$

$$support(\{can.bar\}) = \frac{2}{5} = 0.4$$

Apriori. How it works?

Confidence: A measure of how much Y appears together when X appears

transaction ID	items purchased
1	{flowers, get well card, soda}
2	{plush toy bear, flowers, balloons, candy bar}
3	{get well card, candy bar, flowers}
4	{plush toy bear, balloons, soda}
5	{flowers, get well card, soda}

Let's measure **confidence** of
1. {flower} → {get well card}
2. {get well card} → {flower}

$$confidence(X \rightarrow Y) = \frac{support(X,Y)}{support(X)}$$

$$\frac{support(X,Y)}{support(X)} = \frac{count(X,Y)}{count(X)}$$

$$confidence(\{\text{flower}\} \rightarrow \{\text{get well card}\}) = \frac{count(\{\text{get well card, flower}\})}{count(\{\text{flower}\})} = \frac{3}{4} = 0.75$$

Apriori. How it works?

Confidence: A measure of how much Y appears together when X appears

transaction ID	items purchased
1	{flowers, get well card, soda}
2	{plush toy bear, flowers, balloons, candy bar}
3	{get well card, candy bar, flowers}
4	{plush toy bear, balloons, soda}
5	{flowers, get well card, soda}

Let's measure **confidence** of
1. {flower} → {get well card}
2. {get well card} → {flower}

$$confidence(X \rightarrow Y) = \frac{support(X,Y)}{support(X)}$$

$$\frac{support(X,Y)}{support(X)} = \frac{count(X,Y)}{count(X)}$$

Strong rules
(High support,
High confidence)

$$confidence(\{\text{get well card}\} \rightarrow \{\text{flower}\}) = \frac{count(\{\text{get well card, flower}\})}{count(\{\text{get well card}\})} = \frac{3}{3} = 1.0$$

Apriori. How it works?

Lift: A measure of how much an item bought with another has risen against coincidence

transaction ID	items purchased
1	{flowers, get well card, soda}
2	{plush toy bear, flowers, balloons, candy bar}
3	{get well card, candy bar, flowers}
4	{plush toy bear, balloons, soda}
5	{flowers, get well card, soda}

$$lift(X \rightarrow Y) = \frac{confidence(X \rightarrow Y)}{support(Y)}$$

Constructing a rule set using the Apriori principle

Principle: Frequent All subsets of a set of items should be frequent. In other words, if {A, B} is frequent, both {A} and {B} must be frequent.

1. Identify all item sets that meet the minimum support threshold
2. Create a rule from this set of items with a set of items that meet the minimum confidence threshold.

iteration	must evaluate	frequent itemsets	infrequent itemsets
1	{A}, {B}, {C}, {D}	{A}, {B}, {C}	{D}
2	{A, B}, {A, C}, {B, C} ~~{A, D}, {B, D}, {C, D}~~	{A, B}, {B, C}	{A, C}
3	~~{A, B, C}~~		
4	~~{A, B, C, D}~~		

Practice in r

Association rule syntax

using the `apriori()` function in the `arules` package

Finding association rules:

```
myrules <- apriori(data = mydata, parameter =
  list(support = 0.1, confidence = 0.8, minlen = 1))
```

- `data` is a sparse item matrix holding transactional data
- `support` specifies the minimum required rule support
- `confidence` specifies the minimum required rule confidence
- `minlen` specifies the minimum required rule items

The function will return a rules object storing all rules that meet the minimum criteria.

Examining association rules:

```
inspect(myrules)
```

- `myrules` is a set of association rules from the `apriori()` function

This will output the association rules to the screen. Vector operators can be used on myrules to choose a specific rule or rules to view.

Example:

```
groceryrules <- apriori(groceries, parameter =
 list(support = 0.01, confidence = 0.25, minlen = 2))
inspect(groceryrules[1:3])
```