What is an Association Rule?

ASSOCIATION RULES

Session 3.1 Lecture Video Slides

Background

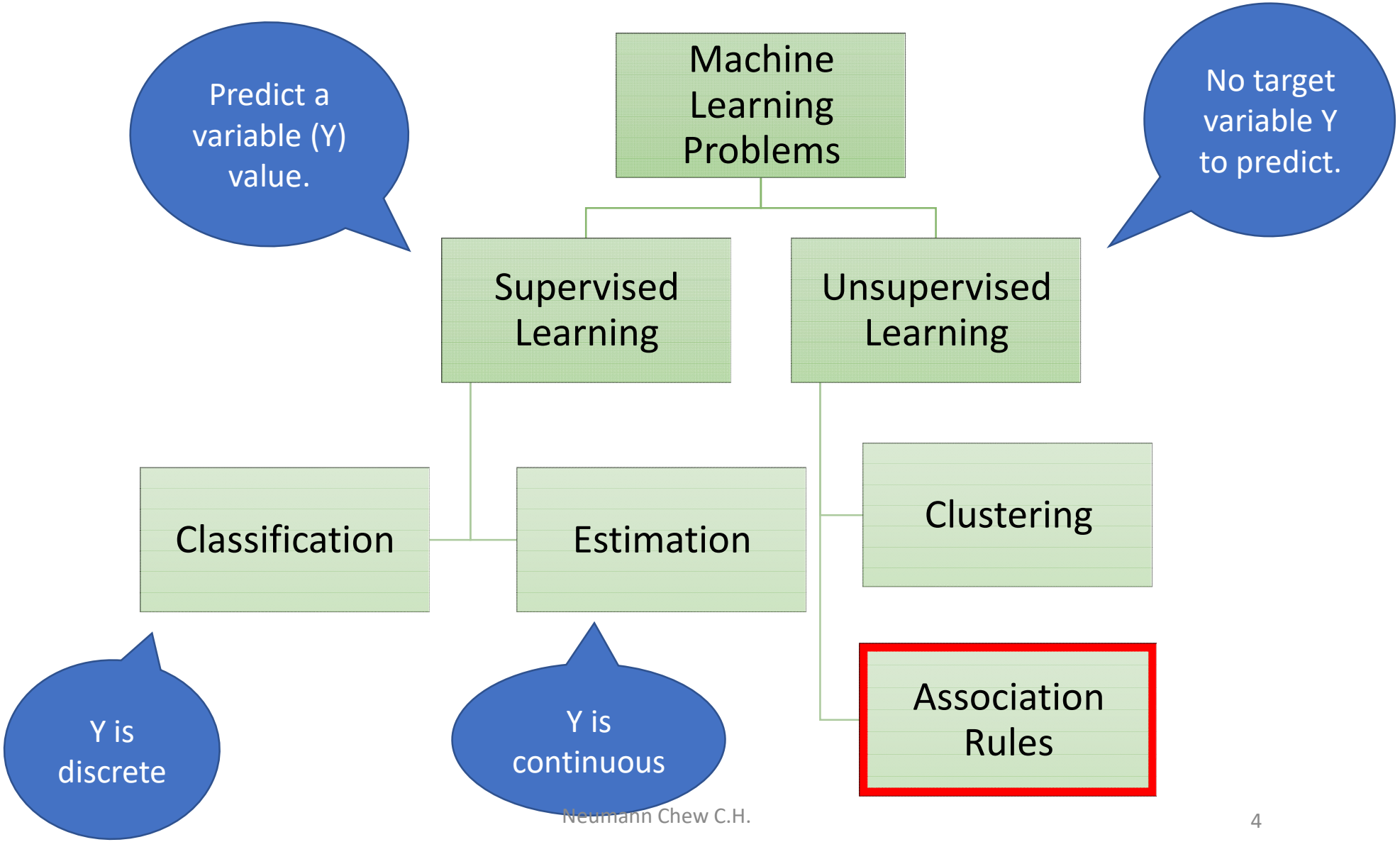
Originally started to analyze retail transaction (i.e. market basket analysis) for the purpose of:

- Which combinations of items are popular?
- Recommend products to Customer.
- Re-design store shelfs to optimize retail sales.

Modern Applications (Recommendation Sys)

- Amazon.com (e-commerce)
- Tao Bao
- Netflix
- Spotify
- In most Recommendation System

Classification of Problems



Data Format

- Which combinations of items are popular?
- Requires Transactional Data (i.e. Receipt-like)
- A record of all the purchases made in one transaction.
  - Examples: Receipt from NTUC supermarket, Invoice for mobile phone contract, Invoice of software purchase, Bill from Amazon.com, etc…
- Two Popular Formats to store the data:
  - Wide Data Format
  - Long Data Format

Example of Wide Data Format

Example database with 5 transactions and 5 items

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

Source: Wikipedia.

- Each row is a transaction record.
- Each column represents a unique item.
- Each row then records a sequence of 0 or 1.
- Each row shows the items purchased in that single transaction.
- If many possible items, dataset is sparse.
  - i.e. many zeros.

Example of Long Data Format

	A	B
1	ID	Item
2	1	Milk
3	4	Milk
4	1	Bread
5	4	Bread
6	5	Bread
7	2	Butter
8	4	Butter
9	3	Beer
10	3	Diapers

- Same dataset as previous slide.
- Each row contains only the transaction ID and one item.
- Analysis Implications…

Fundamental Concepts

1. Itemset
   - A set of items in a transaction.
   - An itemset of size 3 means there are 3 items in the set.
   - In general, it can be of any size, unless specified otherwise.
2. Association Rule
   - Notation: X → Y
   - X associated to Y.
   - X is the “antecedent” itemset, Y is the “consequent” itemset.
     - There might be more than one item in X or Y.

3 Key Concepts

- Many textbooks and websites (e.g. Wikipedia), defines and explains the 3 key concepts (Support, Confidence, Lift) in terms of Set.
- The 3 key concepts are easier to understand and appreciate if we express them in terms of Probability instead of Sets.

3 Key Concepts in an Association Rules X → Y

Define the following in terms of Probability:

- Antecedent Support: Supp (X) ≡ P(contains X)
- Rule Support: Supp (X and Y) ≡ P(contains X and Y)
- Rule Confidence: Conf (X → Y) ≡ P(contains Y|contains X)
- Rule Lift: Lift (X → Y) ≡ $\frac{P(contains\ Y|contains\ X)}{P(contains\ Y)}$

Using the keyword "contains" avoids the awkward definition of P(Y | X) = P(X intersect Y)/P(X).
In probability, X and Y are events but in Association Rules, X and Y are itemsets that typically has no items in common and thus no intersection. X and Y in itemsets actually mean items in X together with items in Y, and hence is a union of two itemsets and not the intersection.

Source: Chew C.H. (2023) AI, Analytics and Data Science Vol. 2.

Simple Numerical Example

- Given the rule R1:
  - {milk, bread} → Butter
- Calculate:
  - Supp ({milk})
  - Supp ({milk, bread})
  - Supp ({milk, bread, butter})
  - Confidence of the rule R1.
  - Lift of the rule R1.

Example database with 5 transactions and 5 items

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

Simple Numerical Example (Answers)

Assoc Rule: {milk, bread} → Butter

- Supp ({milk}) = 2/5
- Supp ({milk, bread}) = 2/5
- Supp ({milk, bread, butter}) = 1/5 = 0.2
- Confidence of the rule = ½ = 0.5
- Lift of the rule = ½ / 2/5 = 1.25

Example database with 5 transactions and 5 items

transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

Modern Applications (Non-Recommendation Sys)

- Definition of item can be any event (not just product for sales).
- Workplace Accidents and Hazards.
- Medical Diagnosis.
- Name: Association Rules (not just market basket analysis).

What is an Association Rule?

- An association rule associates a set of items to another set of items.
- Notation: X → Y
- 3 Key Concepts to measure usefulness of an association rule:
  1. Support
     - Applicability
  2. Confidence
     - Predictive Strength
  3. Lift
     - Context

The Key Question

Given a database with a huge number of transactions, which combinations of items are popular?