Association rules (apriori)
or
Market Basket Analysis

Association rules

- Unsupervised Machine Learning technique
- Association rules (AR) are methods to find some interesting relationship among variables in a large dataset[1]
- Used in domains where there is a need to understand relationships between items
  - E.g: Market-Basket Analysis is a term used in the retail domain that studies the purchase behaviour of a customer
    - ❑ Bread + Egg → will also purchase Milk
    - ❑ Cereal + Milk → will also purchase Juice

    etc…
- Enables retail stores to strategize business needs matching the consumer needs
- Association Rules tell which products are closely related to each other
- Applicable to any industry / domain

Important terms

Transaction	Items purchased
1	Bread, Juice
2	Juice, Eggs, Milk
3	Milk, Butter, Eggs, Chips
4	Eggs, Milk, Biscuit, Noodles
5	Noodles, Sauce
6	Biscuit, Milk

Itemset
(collection of different items)

Frequent Pattern Analysis
A set of 2-items or more repeating frequently

Milk-Eggs (3 times)

Association Rules
A set of frequent items
e.g. {Milk} ---→ {Eggs} mean that the sale of Milk has a strong relation with the sale of Eggs
Interpretation: Customers who buy Milk are highly likely to purchase Eggs

- Computationally expensive on a large dataset
- Information only indicative; not exact

Support **Confidence** **Lift**

Support

- Frequency of the rule in a transaction
- Fraction of transactions that contain an itemset (both X and Y items)

- **Formula**
  Support (I) = # transactions containing I / # transactions

Transaction	Items purchased
1	Bread, Juice
2	Juice, Eggs, Milk
3	Milk, Butter, Eggs, Chips
4	Eggs, Milk, Biscuit, Noodles
5	Noodles, Sauce
6	Biscuit, Milk

- **S(Milk)** 4/6 = 0.66
- **S(Bread)** 1/6 = 0.16
- **S(Egg)** 3/6 = 0.5

- Support can also be for multiple items
- **S(Biscuit+Milk)** = 2/6

Confidence

- If "X" then "Y" (conditional probability)
- How often item **Y** purchased in a transaction when item **X** is also purchased

- **Formula**

  Confidence (X->Y) = # transactions containing X and Y/ # transactions with X

Transaction	Items purchased
1	Bread, Juice
2	Juice, Eggs, Milk
3	Milk, Butter, Eggs, Chips
4	Eggs, Milk, Biscuit, Noodles
5	Noodles, Sauce
6	Biscuit, Milk

• **C(Milk -> Egg)**	3/4 = 0.75
• **C(Bread -> Juice)**	1/1 = 1
• **C(Noodles -> Biscuit)**	1/2 = 0.5

Lift

- Ratio of Confidence and Support
- Measures how many times items X and Y occur together than coincidentally happening - if they are statistically independent
- Greater the Lift ratio, more significant the association between items

- **Formula**

  Lift (X->Y) = Confidence(X -> Y) / Support(X)

Transaction	Items purchased
1	Bread, Juice
2	Juice, Eggs, Milk
3	Milk, Butter, Eggs, Chips
4	Eggs, Milk, Biscuit, Noodles
5	Noodles, Sauce
6	Biscuit, Milk

• **L(Milk -> Egg)**	0.75 / 0.66 = 1.136

• **L(Bread -> Juice)**	1/ 0.16 = 6.25

Sample Association rules

lhs	rhs	support	confidence	lift
Sliced cheese	Whole milk	0.01077783	0.43983402	1.7213560
Whole milk	Sliced cheese	0.01077783	0.04218066	1.7213560
oil	Whole milk	0.01128622	0.40217391	1.5739675
Whole milk	Oil	0.01128622	0.04417031	1.5739675
onions	Other vegetables	0.01423488	0.459[illegible]1639	2.3722681

Example source: Edureka

- Every rule has an LHS and RHS
- Represented as LHS => RHS
- Means items on RHS were frequently purchased with items on LHS

- A large number of rules are generated
- For each rule, Support and Confidence are calculated which may be a wastage of time and resource
- Rules may be redundant
- Select only the strong rules; discard the others
- Strong rules are derived from *frequent item sets* – decisions are made from these strong rules

Interpreting the Association Rules

Filter

	rules	support	confidence	lift	count
1	{Saving.accounts=moderate} => {Checking.account=moderate}	0.286	1.0000000	3.496503	286
2	{Checking.account=moderate} => {Saving.accounts=moderate}	0.286	1.0000000	3.496503	286
3	{Checking.account=rich} => {Saving.accounts=little}	0.603	1.0000000	1.658375	603
4	{Saving.accounts=little} => {Checking.account=rich}	0.603	1.0000000	1.658375	603
5	{Housing=rent,Checking.account=rich} => {Saving.accounts=little}	0.106	1.0000000	1.658375	106
6	{Housing=rent,Saving.accounts=little} => {Checking.account=rich}	0.106	1.0000000	1.658375	106
7	{Job=1,Checking.account=rich} => {Saving.accounts=little}	0.128	1.0000000	1.658375	128
8	{Job=1,Saving.accounts=little} => {Checking.account=rich}	0.128	1.0000000	1.658375	128
9	{Job=1,Checking.account=rich} => {Housing=own}	0.103	0.8046875	1.128594	103
10	{Job=1,Saving.accounts=little} => {Housing=own}	0.103	0.8046875	1.128594	103
11	{Job=1,CreditScore=Bad} => {Housing=own}	0.126	0.8025478	1.125593	126
12	{Job=1,Housing=own} => {CreditScore=Bad}	0.126	0.8181818	1.208540	126
13	{Sex=male,Job=1} => {Housing=own}	0.112	0.8235294	1.155020	112
14	{Job=2,Saving.accounts=moderate} => {Checking.account=moderate}	0.186	1.0000000	3.496503	186
15	{Job=2,Checking.account=moderate} => {Saving.accounts=moderate}	0.186	1.0000000	3.496503	186
16	{Saving.accounts=moderate,MaritalStatus=Divorced} => {Checking.account=moderate}	0.172	1.0000000	3.496503	172
17	{Checking.account=moderate,MaritalStatus=Divorced} => {Saving.accounts=moderate}	0.172	1.0000000	3.496503	172
18	{Saving.accounts=moderate,CreditScore=Bad} => {Checking.account=moderate}	0.194	1.0000000	3.496503	194
19	{Checking.account=moderate,CreditScore=Bad} => {Saving.accounts=moderate}	0.194	1.0000000	3.496503	194
20	{Sex=male,Saving.accounts=moderate} => {Checking.account=moderate}	0.205	1.0000000	3.496503	205
21	{Sex=male,Checking.account=moderate} => {Saving.accounts=moderate}	0.205	1.0000000	3.496503	205
22	{Housing=own,Saving.accounts=moderate} => {Checking.account=moderate}	0.201	1.0000000	3.496503	201

Rule 3

60.3% have ***savings account*** = "little" and ***checking account*** = "rich"

Confidence

Rule 10:

If ***job***=1 and ***savings*** = "little", then there is a 80% chance that the ***Housing*** = "own" will occur