association analysis

↳ given a set of transactions, find rules that will predict the occurance of an item based on the occurences of other items in the transaction

* implication means cooccurence, not causality

Market-Basket transactions

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

support count (σ) = frequency of occurence of an itemset → $\sigma(\{milk, bread, diaper\}) = 2$

association rule = $X \to Y$ (item sets) ⇒ {milk, diaper} → {beer}

support (s) = fraction of transactions that contain both X and Y ⇒ $s = \frac{\sigma(\{milk, bread, diaper\})}{|T|} = \frac{2}{5}$ ($X+Y$; (All))

↳ ≥ minsup threshold → frequent itemset

confidence (c) = measure of how often Y appear in transactions that contain X ⇒ $c = \frac{\sigma(\{milk, bread, diaper\})}{\sigma(\{milk, bread\})} = \frac{2}{3}$ ($X+Y$; X)

↳ ≥ minconf treshold

how to find rules? first find frequent itemset (≥ minsup), then generate rules from these that have high confidence

frequent itemset generation strategies

- reducing number of candidates = for d items, 2^d candidates

↳ apriori principle = if an itemset is frequent, then all of its subsets must also be frequent

↳ anti-monotone property = support of an itemset never exceeds the support of its subsets

- if AB is not frequent, ABC, ABCD, ABCDE → all its supersets are also not frequent (prune)

minsup = 3

if every subset is considered: $\binom{6}{1} + \binom{6}{2} + \binom{6}{3} = \boxed{41}$

Items (1-itemsets)

Item	Count
Bread	4
~~Coke~~	~~2~~
Milk	4
Beer	3
Diaper	4
~~Eggs~~	~~1~~

6 $\xrightarrow{4}$

2-itemsets

Itemset	Count
{Bread,Milk}	3
~~{Beer, Bread}~~	~~2~~
{Bread,Diaper}	3
~~{Beer,Milk}~~	~~2~~
{Diaper,Milk}	3
{Beer,Diaper}	3

$\binom{4}{2} = 6$

3-itemsets

Itemset	Count
{ Beer, Diaper, Milk}	2
{ Beer,Bread, Diaper}	2
{Bread, Diaper, Milk}	2
{Beer, Bread, Milk}	1

$\binom{4}{3} = 4$

with support-based pruning, $6 + 6 + 4 = \boxed{16}$ subsets are considered

after this → use $F_{k-1} \times F_{k-1}$ method

↳ {Bread, Diaper, Milk} only $6 + 6 + 1 = \boxed{13}$

↳ candidate generation: $F_{k-1} \times F_{k-1}$ method

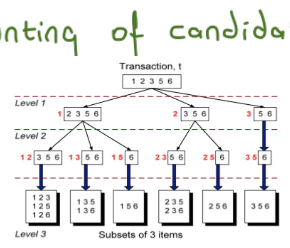
↳ Merge (ABC, ABD) = ABCD merge if first k items are same

↳ do not merge ACD, ABD → must have same prefix

↳ alternate $F_{k-1} \times F_{k-1}$ method

↳ Merge (ABC, BCD) = ABCD merge if last k items of first one are same with the first k items of second one

support counting of candidate items = instead of matching each transaction against every candidate, match with hash buckets



for transactions with length 3

→ {1,3,5} supported by transaction {1,2,3,5,6}?

rule generation = from frequent itemsets, find the ones that have confidence ≥ minconf

↳ for {A,B,C,D} ⇒ A → BCD B → ACD ... $\binom{4}{1} + \binom{4}{2} + \binom{4}{3} = 14$ possible rules $2^4 - 2$ $\boxed{(2^k - 2)}$

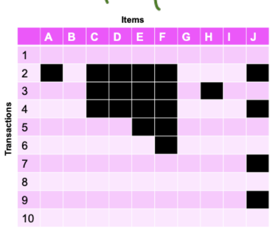
* confidence does not have anti-monotone property (A → BCD and AB → D are independent) for different itemsets but have for same itemset

$c(ABC \to D) \geq c(AB \to CD) \geq c(A \to BCD)$ ($\frac{ABCD}{ABC}$, $\frac{ABCD}{AB}$, $\frac{ABCD}{A}$) payda büyüyor

↳ if this below the minimum confidence, no need to check rest, eliminate, prune

complexity of apriori = increases with decrease of support threshold

maximal frequent itemset = if none of itemset's immediate supersets (one level up) is frequent



support threshold =	5	4	3
frequent itemsets =	F	E, F, J, EF	C, CD, D...
maximal itemsets =	F	EF, J	CDEF, J

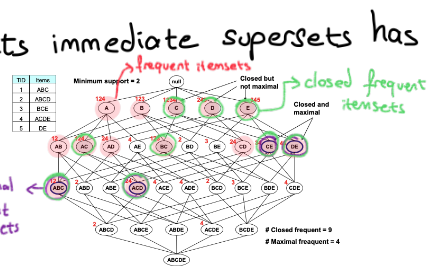
closed itemset = if none of an itemsets immediate supersets has the same support as X

TID	Items
1	{A,B}
2	{B,C,D}
3	{A,B,C,D}
4	{A,B,D}
5	{A,B,C,D}

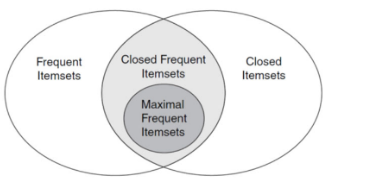
Itemset	Support
{A}	4
{B}	5
{C}	3
{D}	4
{A,B}	4
{A,C}	2
{A,D}	3
{B,C}	3
{B,D}	4
{C,D}	3

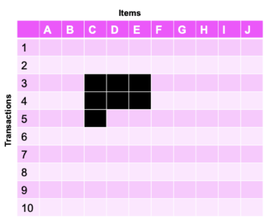
not closed / closed / closed

Itemset	Support
{A,B,C}	2
{A,B,D}	3
{A,C,D}	2
{B,C,D}	2
{A,B,C,D}	2



frequent itemsets / closed frequent itemsets / maximal frequent itemsets





Itemsets	Support (counts)	Closed itemsets
{C}	3	✓
{D}	2	✗
{E}	2	✗
{C,D}	2	✗
{C,E}	2	✗
{D,E}	2	✗ ✓
{C,D,E}	2	✓

interestigness measure = $X \to Y$

Contingency table

	Y	$\overline{Y}$	
X	f_{11}	f_{10}	f_{1+}
$\overline{X}$	f_{01}	f_{00}	f_{0+}
	f_{+1}	f_{+0}	N

↳ supports

tea → coffee

	Coffee	$\overline{Coffee}$	
Tea	150	50	200
$\overline{Tea}$	650	150	800
	800	200	1000

confidence = P(coffee | tea) = 150/200 = 0.75

P(coffee) = 800/100 = 0.8

means people who drink tea are more likely to drink coffee than not to drink

means knowing that a person drinks tea reduces the probability that he/she drinks coffee

contradiction

shouldn't reduce

* confidence (X → Y) > support(Y) : otherwise rule will be misleading

must be high

$P(X,Y) > P(X) \times P(Y)$ → X & Y positively correlated

$P(X,Y) < P(X) \times P(Y)$ → X & Y negatively correlated

$P(X,Y) = P(X) \times P(Y)$ → X & Y are independent ⇒ confidence (X → Y) = support(Y) ⇒ $P(Y|X) = P(Y)$

lift = $\frac{P(Y|X)}{P(Y)}$ → 1 if they are independent → used to measure the importance of a rule

they are same if they are independent

interest = $\frac{P(X,Y)}{P(X) \cdot P(Y)}$ → 1 if they are independent

invarient measures to inverse = cosine, jaccard, confidence

non-varient measures to inverse = correlation, interest / lift, odds ratio → they dont change

simpson's paradox = observed relationship in data may be influenced by the presence of hidden variables

↳ recovery rate in hospitals, hidden variabe = young or old patients

cross support and H-confidence ⇒

conf (caviar → milk) → very high

conf (milk → caviar) → very low

when caviar is not bought frequently, but when someone buys milk, mostly buys cavier → hard to catch to rule because support of caviar is too low, not frequent itemset

$0 \leq hconf(X) \leq r(X) \leq 1$

h-confidence = min conf of any association rule formed from itemset X → $s(X) / \max\{s(x_i)\}$

cross-support = $\min\{s(x_i)\} / \max\{s(x_j)\}$

hyperclique = items in the itemset are strongly correlated → not necessarily frequent itemsets