Introduction to Data Science

GIRI NARASIMHAN, SCIS, FIU

APRIORI Algorithm

Frequent Itemsets (MinSup 20%)

Almonds, Beer, Cheese, Dogfood, Eggs, Fruit, Gree

TID	A	B	C	D	E	F	G
1	1	1		1			1
2			1	1	1		
3		1	1			1	
4		1				1	
5			1		1		1
6						1	
7	1		1	1			
8						1	
9			1		1		
10		1					1
11			1		1		1
12	1						
13			1			1	
14	1		1	1		1	
15							
16				1			
17	1		1			1	
18	1	1	1	1			
19	1	1	1	1			1
20					1		

APRIORI Algorithm

Frequent Itemsets

(MinSup 20%)

Almonds, Beer, Cheese, Dogfood, Eggs, Fruit, Greens

TID	A	B	C	D	E	F	G
1	1	1		1			1
2			1	1	1		
3		1	1			1	
4		1				1	
5			1		1		1
6						1	
7	1		1	1			
8						1	
9			1		1		
10		1					1
11			1		1		1
12	1						
13			1			1	
14	1		1	1		1	
15							
16				1			
17	1		1			1	
18	1	1	1	1			
19	1	1	1	1			1
20					1		

> **Central Observation:** *If a set X of items is frequent, then so is every subset of X.*

Monotonicity Property

Implication of Monotonicity

- Example: If {beer, cheese} is not frequent, then no need to consider set {beer, cheese, dogfood}
- We don't have to consider any sets with an infrequent subset
- We consider subsets in the order of increasing size and all of whose subsets are frequent
- Once a subset is eliminated, all its supersets are removed from consideration

APRIORI Algorithm

Frequent Itemsets

(MinSup 20%)

Almonds, Beer, Cheese, Dogfood, Eggs, Fruit, Greens

TID	A	B	C	D	E	F	G
1	1	1		1			1
2			1	1	1		
3		1	1			1	
4		1				1	
5			1		1		1
6						1	
7	1		1	1			
8						1	
9			1		1		
10		1					1
11			1		1		1
12	1						
13			1			1	
14	1		1	1		1	
15							
16				1			
17	1		1			1	
18	1	1	1	1			
19	1	1	1	1			1
20					1		

Iteration 1

- **Find all frequent subsets of size 1**
- **It turns out that all subsets of size 1 are frequent with support 20%**
- **L_1 = {{A}, {B}, {C}, {D}, {E}, {F}, {G}}**

Iteration 2

- Generate all possible subsets of size 2 from L_1
- C_2 = {{a, b}, {a, c}, {a, d}, {a, e}, {a, f}, {a, g}, {b, c}, {b, d}, {b, e}, {b, f}, {b, g}, {c, d}, {c, e}, {c, f}, {c, g}, {d, e}, {d, f}, {d, g}, {e, f}, {e, g}, {f, g}}
- Identify those with minimum support of 20%
- L_2 = {{a, c}, {a, d}, {c, d}, {c, e}, {c, f}}

Iteration 3

- Generate all possible subsets of size 3 from L_2
- Remember that L_2 = {{a, c}, {a, d}, {c, d}, {c, e}, {c, f}}
- C_3 = {{a, c, d}, {c, d, e}, {c, d, f}, {c, e, f}}
  - We can prune {c, d, e} since {d, e} is not in L2
  - We can prune {c, d, f} since {d, f} is not in L2
  - We can prune {c, e, f} since {e, f} is not in L2
- Identify remaining subsets with minimum support of 20%
- L_3 = {{a, c, d}} = {{almonds, cheese, dogfood}}

Iteration 4

- Generate all possible subsets of size 4 from L_3
- Remember that L_3 = {{a, c, d}}
- C_4 = { }
- L_4 = { }
- STOP!

- Only one frequent itemset: {Almonds, Cheese, Dogfood}

APRIORI Algorithm

Frequent Itemsets (MinSup 20%)

Almonds, Beer, Cheese, Dogfood, Eggs, Fruit, Greens

TID	A	B	C	D	E	F	G
1	1	1		1			1
2			1	1	1		
3		1	1			1	
4		1				1	
5			1		1		1
6						1	
7	1		1	1			
8						1	
9			1		1		
10		1					1
11			1		1		1
12	1						
13			1			1	
14	1		1	1		1	
15							
16				1			
17	1		1			1	
18	1	1	1	1			
19	1	1	1	1			1
20					1		

Association Rule

- A useful rule is one that says, if you buy almonds and dogfood, then you are likely to buy cheese as well.
  - “Diaper-Beer” rule
- Confidence: Confidence of an Association Rule is the percentage of applicable rules where the rule is true
- Rule: if X then Y
- Confidence: #(X and Y) / #(X)
- Confidence of “If almonds and dogfood, then cheese” is ?
  - 80% or 0.8

APRIORI Algorithm

Frequent Itemsets (MinSup 20%)

Almonds, Beer, Cheese, Dogfood, Eggs, Fruit, Greens

TID	A	B	C	D	E	F	G
1	1	1		1			1
2			1	1	1		
3		1	1			1	
4		1				1	
5			1		1		1
6						1	
7	1		1	1			
8						1	
9			1		1		
10		1					1
11			1		1		1
12	1						
13			1			1	
14	1		1	1		1	
15							
16				1			
17	1		1			1	
18	1	1	1	1			
19	1	1	1	1			1
20					1		

Other Applications

- Word Frequency Count
- Patterns in biomolecular sequences
- …

Patterns

Loc	Protein Name	Helix 2									Turn				Helix 3								
		-1	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
14	Cro	F	G	Q	E	K	T	A	K	D	L	G	V	Y	Q	S	A	I	N	K	A	I	H
16	434 Cro	M	T	Q	T	E	L	A	T	K	A	G	V	K	Q	Q	S	I	Q	L	I	E	A
11	P22 Cro	G	T	Q	R	A	V	A	K	A	L	G	I	S	D	A	A	V	S	Q	W	K	E
31	Rep	L	S	Q	E	S	V	A	D	K	M	G	M	G	Q	S	G	V	G	A	L	F	N
16	434 Rep	L	N	Q	A	E	L	A	Q	K	V	G	T	T	Q	Q	S	I	E	Q	L	E	N
19	P22 Rep	I	R	Q	A	A	L	G	K	M	V	G	V	S	N	V	A	I	S	Q	W	E	R
24	CII	L	G	T	E	K	T	A	E	A	V	G	V	D	K	S	Q	I	S	R	W	K	R
4	LacR	V	T	L	Y	D	V	A	E	Y	A	G	V	S	Y	Q	T	V	S	R	V	V	N
167	CAP	I	T	R	Q	E	I	G	Q	I	V	G	C	S	R	E	T	V	G	R	I	L	K
66	TrpR	M	S	Q	R	E	L	K	N	E	L	G	A	G	I	A	T	I	T	R	G	S	N
22	BlaA Pv	L	N	F	T	K	A	A	L	E	L	Y	V	T	Q	G	A	V	S	Q	Q	V	R
23	TrpI Ps	N	S	V	S	Q	A	A	E	Q	L	H	V	T	H	G	A	V	S	R	Q	L	K

- **Q1 G9 N20**
- **A5 G9 V10 I15**

Candidates generation

L_3

G1, V2, S3
G1, V2, T6
G1, V2, I7
G1, V2, E8
G1, S3, T6
G1, T6, I7
V2, T6, I7
V2, T6, E8

C_4

G1, V2, S3, T6
G1, V2, S3, I7
G1, V2, S3, E8
G1, V2, T6, I7
G1, V2, T6, E8
G1, V2, I7, E8
V2, T6, I7, E8

L_4

G1, V2, S3, T6
G1, V2, S3, I7
G1, V2, S3, E8
G1, V2, T6, E8
V2, T6, I7, E8