Section 2
Basic Metrics

1. Support

Definition and Formula

- Indication of how frequently an itemset appear in a dataset

$$Support\ (X) = \frac{freq(X)}{N}$$

The number of transactions that contain X

The total number of transactions

- Very low support → Not enough data for mining

Example: Support({bread, milk})

$$Support(\{bread, milk\})$$

$$= \frac{freq(\{bread, milk\})}{N}$$

$$= \frac{3}{6}$$

$$= 0.5$$

T1	cheese, ham
T2	bread, milk
T3	bread, milk, ham
T4	bread, cheese, ham
T5	milk
T6	bread, milk

Example: Support(cheese)

$$Support(\{cheese\})$$

$$= \frac{freq(\{cheese\})}{N}$$

$$= \frac{2}{6}$$

$$= 0.33$$

T1	cheese, ham
T2	bread, milk
T3	bread, milk, ham
T4	bread, cheese, ham
T5	milk
T6	bread, milk

Weakness of Support

- Among the itemsets with two items, {water, bread} has the highest support.

- This information is NOT useful.

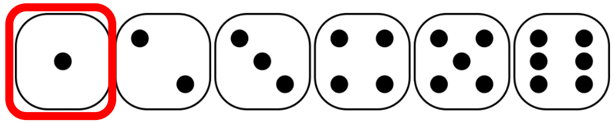
  Because most transactions contain water and bread

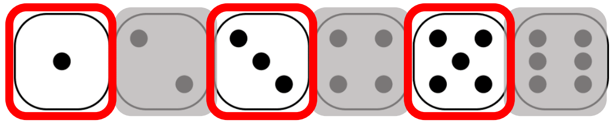
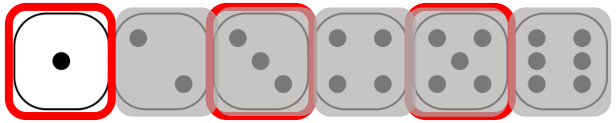
- By using conditional probability, we can deal with this problem.

T1	water, bread
T2	water, bread, cookie
T3	water, yogurt
T4	water, bread, ham
T5	water, bread, ham, butter
T6	water, bread, jam

2. Basic Math: Conditional Probability

Roll a Dice

- Probability of getting a *1*:  $= \frac{1}{6} = 0.17$

- Probability of getting an odd number, and it is a *1*.
  - Getting an odd number: 
  - And it is a *1*:  $= \frac{1}{3} = 0.33$

Conditional Probability

Definition: Probability of an event A occurring given that another event B has already occurred.

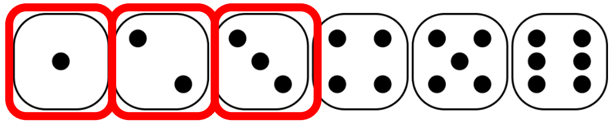
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$P(B)$: Probability of an event B occuring

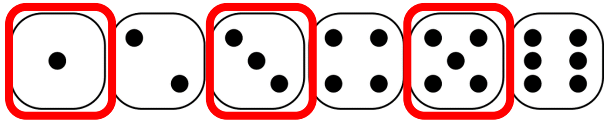
$P(A \cap B)$: Probability of both event A and B occuring

Another Example: Roll a Dice

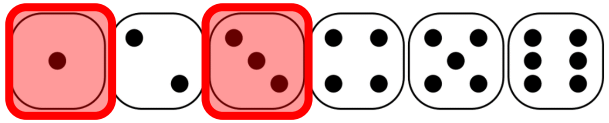
$$P(A){:}\, 3\ \mathit{or\ less}\ \ = 0.5$$



$$P(B){:}\, \mathit{Odd\ number}\ \ = 0.5$$

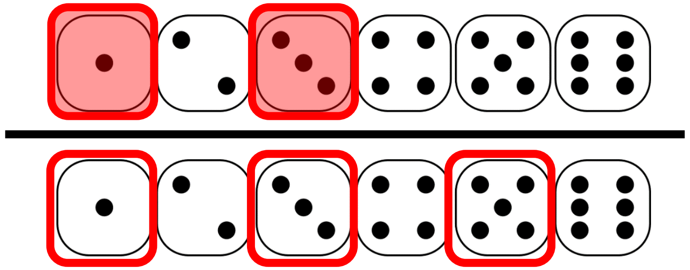


$$P(A \cap B)\ \ = 0.33$$



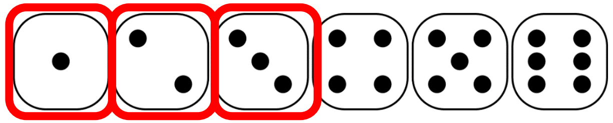
$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

$$= \frac{0.33}{0.5}\ \ = 0.67$$

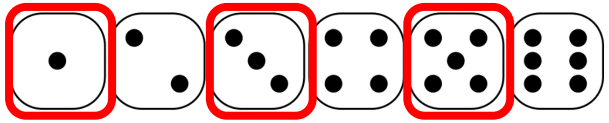


Another Example: Roll a Dice —by frequency

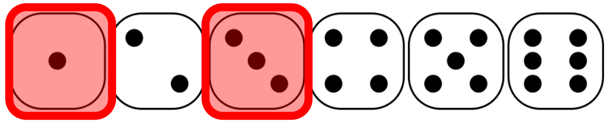
$$freq(A) = 3$$



$$freq(B) = 3$$

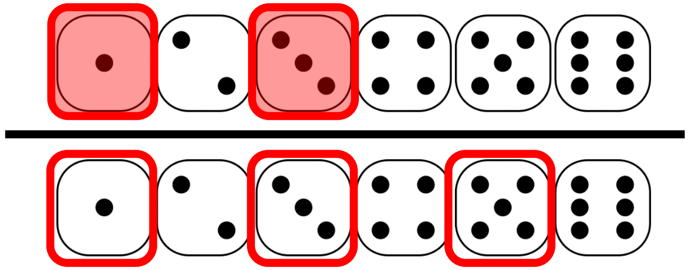


$$freq(A, B) = 2$$



$$P(A \mid B) = \frac{freq(A, B)}{freq(B)}$$

$$= \frac{2}{3} = 0.67$$



3. Confidence

Confidence

Definition: The probability of itemset Y (consequent) appearing, given that itemset X (antecedent) has already appeared.

$$P(Y|\, X)$$

$$Confidence(X \to Y) = \frac{freq(X, Y)}{freq(X)}$$

The number of transactions that contains both X and Y

The number of transactions that contains X

Example: Confidence ({water} → {ham})

$$Confidence(\{water\} \rightarrow \{ham\})$$

$$= \frac{freq(water, ham)}{freq(water)} = \frac{2}{6} = 0.33$$

T1	water, bread
T2	water, bread, cookie
T3	water, yogurt
T4	water, bread, ham
T5	water, bread, ham, butter
T6	water, bread, jam

Water is purchased in all transations.

→ Conditioning with water is meaningless.

Example: Confidence ({bread} → {ham})

$$Confidence(\{bread\} \rightarrow \{ham\})$$

$$= \frac{freq(bread, ham)}{freq(bread)}$$

$$= \frac{2}{4}$$

$$= 0.5$$

T1	water, bread
T2	water, bread, cookie
T3	water, yogurt
T4	water, bread, ham
T5	water, bread, ham, butter
T6	water, bread, jam

Weakness of Confidence

$$Confidence(X \rightarrow Y) = \frac{freq(X,Y)}{freq(X)}$$

When the frequency of Y is very high, the confidence will be high irrespective of actual association.

Example: Weakness of Confidence

$$Confidence(detergent \to water)$$

$$= \frac{freq(detergent, water)}{freq(detergent)}$$

$$= \frac{2}{3} = 0.67$$

Is the co-occurrence high ?

T1	bread, butter, water
T2	aluminum foil, towel
T3	milk, beef, water
T4	detergent, chicken, water
T5	bread, ham, butter, water
T6	bread, mik, water
T7	detergent, water
T8	bread, bacon, egg, water
T9	detergent, cookie
T10	potato chips, coffee

Example: Weakness of Confidence (Continued)

$$Confidence(water \rightarrow detergent)$$

$$= \frac{freq(water, detergent)}{freq(water)}$$

$$= \frac{2}{7} = 0.29$$

Just swapping lowered the value of confidence!

T1	bread, butter, water
T2	aluminum foil, towel
T3	milk, beef, water
T4	detergent, chicken, water
T5	bread, ham, butter, water
T6	bread, mik, water
T7	detergent, water
T8	bread, bacon, egg, water
T9	detergent, cookie
T10	potato chips, coffee

4. Lift

Lift

$$Lift(X \rightarrow Y) = \underline{\frac{freq(X,Y)}{freq(X)}} \cdot \underline{\frac{N}{freq(Y)}}$$

$$= \underline{Confidence(X \rightarrow Y)} \cdot \underline{\frac{1}{Support(Y)}}$$

$$= \frac{Confidence(X \rightarrow Y)}{Support(Y)}$$

Meaning of Lift

$$Lift(X \rightarrow Y) = \frac{Confidence(X \rightarrow Y)}{Support(Y)} = \frac{P(Y|X)}{P(Y)}$$

$Lift(X \rightarrow Y) > 1$ ➡ $P(Y|X) > P(Y)$

⬇

The occurrence of X increased the probability of occurrence of Y

Example: Roll a Dice

X_1: $Odd\ number\ (Antecedent\ 1)$

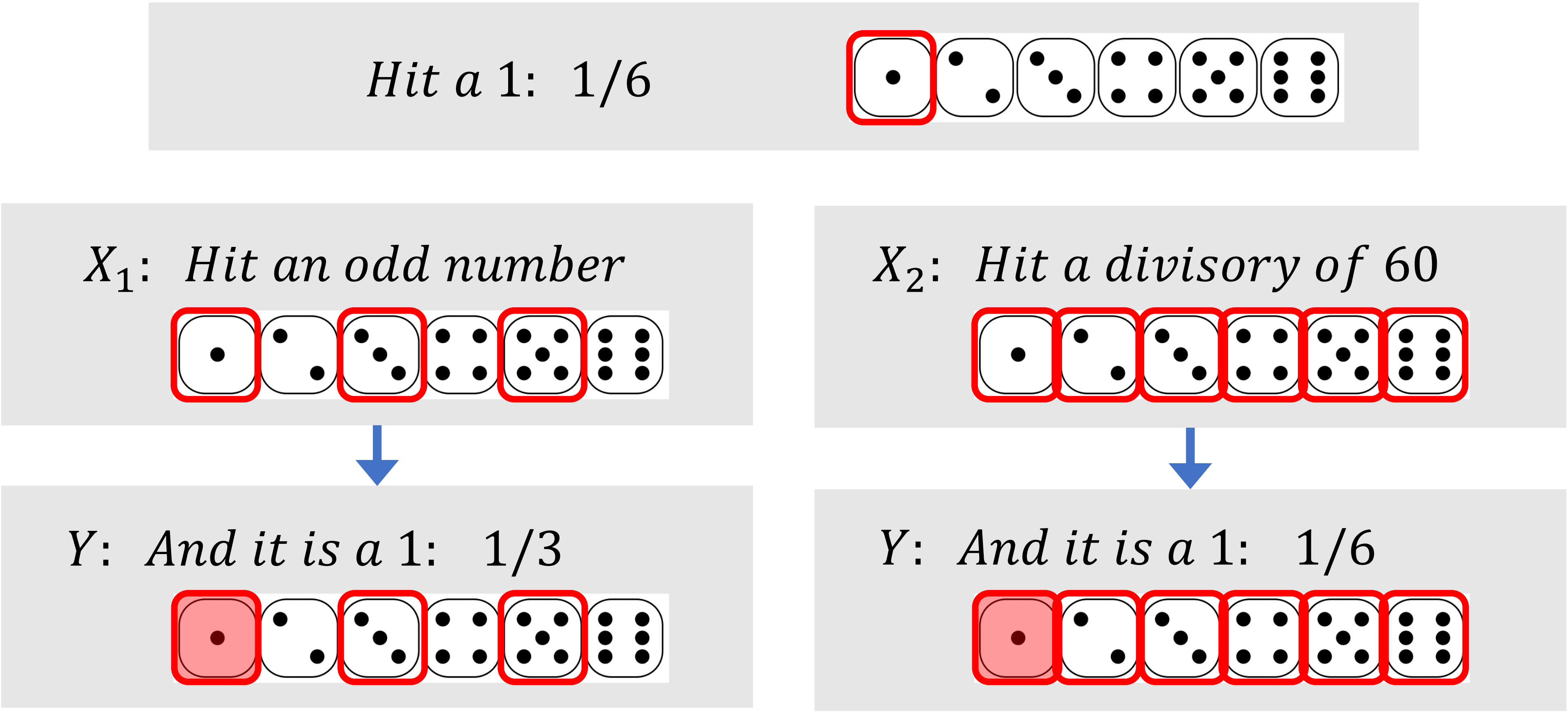
X_2: $Divisor\ of\ 60\ (Antecedent\ 2)$

Y: $1\ (Consequent)$

$$P(Y) = \frac{1}{6}$$

Do X_1 and X_2 increase the possibility of Y occurring?

Example: Roll a Dice (Continued)



$Hit\ a\ 1:\ \ 1/6$

$X_1:\ \ Hit\ an\ odd\ number$

$Y:\ \ And\ it\ is\ a\ 1:\ \ \ 1/3$

$X_2:\ \ Hit\ a\ divisory\ of\ 60$

$Y:\ \ And\ it\ is\ a\ 1:\ \ \ 1/6$

X_1 increased the probability of the occurrence of Y, but X_2 does not.

Example: Lift({detergent} → {water})

$$Lift(X \rightarrow Y) = \frac{freq(X,Y)}{freq(X)} \cdot \frac{1}{Support(Y)}$$

$$\frac{\boxed{freq(detergent, water)}}{\boxed{freq(detergent)}} \times \frac{1}{Support(water)}$$

$$= \frac{2}{3} \times \frac{1}{0.7} = \frac{2}{2.1} = 0.95$$

T1	bread, butter, water
T2	aluminum foil, towel
T3	milk, beef, water
T4	detergent, chicken, water
T5	bread, ham, butter, water
T6	bread, mik, water
T7	detergent, water
T8	bread, bacon, egg, water
T9	detergent, cookie
T10	potato chips, coffee

5. Comprehension

Question. What are the ranges of each metric?

- Support
- Confidence
- Lift

Answer

$$Support(X) = \frac{freq(X)}{N}$$

$freq(X)$ — $0\sim N$

N — N

$$\Rightarrow \left[\frac{0}{N}, \frac{N}{N}\right] \Rightarrow [0, 1]$$

$$Confidence(X \rightarrow Y) = \frac{freq(X,Y)}{freq(X)}$$

$freq(X,Y)$ — $0\sim K$

$freq(X)$ — K

$$\Rightarrow \left[\frac{0}{K}, \frac{K}{K}\right] \Rightarrow [0, 1]$$

$$Lift(X \rightarrow Y) = \frac{Confidence(X,X)}{Support(X)}$$

$Confidence(X,X)$ — $0\sim 1$

$Support(X)$ — $0\sim 1$

$$\Rightarrow \left[\frac{0}{1}, \frac{1}{0}\right] \Rightarrow [0, \infty]$$

Transformation of Formula

- Support(X) $= \dfrac{freq(X)}{N}$

- Confidence(X → Y) $= \dfrac{freq(X,\ Y)}{freq(X)} = \overset{Support(X\&Y)}{\boxed{\dfrac{freq(X,Y)}{N}}} \cdot \overset{\frac{1}{Support(X)}}{\boxed{\dfrac{N}{freq(X)}}} = \dfrac{Support(X\&Y)}{Support(X)}$

- Lift(X → Y) $= \dfrac{Confidence(X \rightarrow\ Y)}{Support(Y)} = \dfrac{Support(X\&Y)}{Support(X)} \cdot \dfrac{1}{Support(Y)}$

$$= \frac{Support(X\&Y)}{Support(X) \cdot Support(Y)}$$

6. Summary

Summary

- Association rule mining is used to find co-occurrence.
- We have metrics to quantify associations.
- The number of association rules can be enormous.
- So we use apriori algorithm to idenfity impotant rules.