Lecture 2: 27 January, 2022

Madhavan Mukund

https://www.cmi.ac.in/~madhavan

Data Mining and Machine Learning
January–May 2022

Market-Basket Analysis

- Set of items $I = \{i_1, i_2, \ldots, i_N\}$
- Set of transactions $T = \{t_1, t_2, \ldots, t_M\}$
  - A transaction is a set $t \subseteq I$ of items
- Identify association rules $X \rightarrow Y$
  - $X, Y \subseteq I$, $X \cap Y = \emptyset$
  - If $X \subseteq t_j$ then it is likely that $Y \subseteq t_j$
- Two thresholds
  - How frequently does $X \subseteq t_j$ imply $Y \subseteq t_j$?
  - How significant is this pattern overall?

Setting thresholds

- For $Z \subseteq I$, $Z.\mathrm{count} = |\{t_j \mid Z \subseteq t_j\}|$
- How frequently does $X \subseteq t_j$ imply $Y \subseteq t_j$?
  - Fix a confidence level χ
  - Want $\dfrac{(X \cup Y).count}{X.count} \geq \chi$
- How significant is this pattern overall?
  - Fix a support level σ
  - Want $\dfrac{(X \cup Y).count}{M} \geq \sigma$
- Given sets of items I and transactions T, with confidence χ and support σ, find all valid association rules $X \to Y$

Frequent itemsets

- $X \rightarrow Y$ is interesting only if $(X \cup Y).\text{count} \geq \sigma \cdot M$
- First identify all frequent itemsets
  - $Z \subseteq I$ such that $Z.\text{count} \geq \sigma \cdot M$
- Naïve strategy: maintain a counter for each Z
  - For each $t_j \in T$
    For each $Z \subseteq t_j$
    Increment the counter for Z
  - After scanning all transactions, keep Z with $Z.\text{count} \geq \sigma \cdot M$
- Need to maintain $2^{|I|}$ counters
  - Infeasible amount of memory
  - Can we do better?

Sample calculation

- Let's assume a bound on each $t_i \in T$
  - No transaction has more than 10 items !!
- Say $N = |I| = 10^6$, $M = |T| = 10^9$, $\sigma = 0.01$

  million = billion =
  - Number of possible subsets to count is $\displaystyle\sum_{i=1}^{10} \binom{10^6}{i}$

Why do we expect to be able to do better?

$X \to Y$

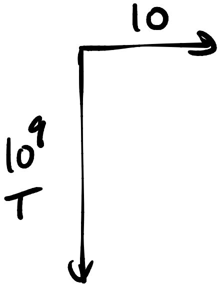
20 7

Sample calculation

- Let's assume a bound on each $t_i \in T$
  - No transaction has more than 10 items
- Say $N = |I| = 10^6$, $M = |T| = 10^9$, $\sigma = 0.01$
  - Number of possible subsets to count is $\displaystyle\sum_{i=1}^{10} \binom{10^6}{i}$
- A singleton subset $\{x\}$ that is frequent is an item x that appears in at least 10^7 transactions $= 0.01 \times 10^9$

$x \in I$

Is $\{x\}$ frequent?

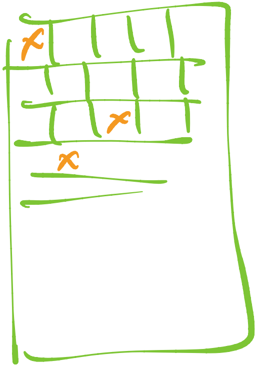


Sample calculation

- Let's assume a bound on each $t_i \in T$
  - No transaction has more than 10 items
- Say $N = |I| = 10^6$, $M = |T| = 10^9$, $\sigma = 0.01$
  - Number of possible subsets to count is $\sum_{i=1}^{10} \binom{10^6}{i}$
- A singleton subset $\{x\}$ that is frequent is an item x that appears in at least 10^7 transactions
- Totally, T contains at most 10^{10} items

$10^9 \times 10$

- At most $10^{10}/10^7 = 1000$ items are frequent!
- How can we exploit this?



Apriori

- Clearly, if Z is frequent, so is every subset $Y \subseteq Z$

$$\frac{X \cup Y.\text{count}}{X.\text{count}} \leq 1$$

Z appears K times

$Y \subseteq Z$ appears $\geq K$ times

$Y.\text{count} \geq Z.\text{count} \geq \sigma M$

Apriori

NOT $\longleftarrow$ NOT

- Clearly, if Z is frequent, so is every subset $Y \subseteq Z$
- We exploit the contrapositive

Apriori observation

$Z \supseteq Y$

If $\cancel{Z}$ Y is not a frequent itemset, no superset $\cancel{Y \supseteq Z}$ can be frequent

Apriori

- Clearly, if Z is frequent, so is every subset $Y \subseteq Z$
- We exploit the contrapositive

Apriori observation

If Z is not a frequent itemset, no superset $Y \supseteq Z$ can be frequent

- For instance, in our earlier example, every frequent itemset must be built from the 1000 frequent items
- In particular, for any frequent pair $\{x, y\}$, both $\{x\}$ and $\{y\}$ must be frequent
- Build frequent itemsets bottom up, size 1,2,...

for each trans
for $Z \subseteq t$
Update
Z count

1000×999 possible
freq pairs

Apriori algorithm

- F_i : frequent itemsets of size i — Level i

Apriori algorithm

- F_i : frequent itemsets of size i — Level i
- F_1: Scan T, maintain a counter for each $x \in I$

Apriori algorithm

- F_i : frequent itemsets of size i — Level i
- F_1: Scan T, maintain a counter for each $x \in I$
- $C_2 = \{\{x, y\} \mid x, y \in F_1\}$: Candidates in level 2

$\{x\}, \{y\} \in F_1$

Apriori algorithm

- F_i : frequent itemsets of size i — Level i
- F_1: Scan T, maintain a counter for each $x \in I$
- $C_2 = \{\{x, y\} \mid x, y \in F_1\}$: Candidates in level 2
- F_2: Scan T, maintain a counter for each $X \in C_2$

Apriori algorithm

- F_i : frequent itemsets of size i — Level i
- F_1: Scan T, maintain a counter for each $x \in I$
- $C_2 = \{\{x, y\} \mid x, y \in F_1\}$: Candidates in level 2
- F_2: Scan T, maintain a counter for each $X \in C_2$
- $C_3 = \{\{x, y, z\} \mid \{x, y\}, \{x, z\}, \{y, z\} \in F_2\}$
- F_3: Scan T, maintain a counter for each $X \in C_3$

F_1 small $\subseteq I$

$F_2 \subseteq C_2$

Restricts $I \times I \times I$ to C_3

T

x,y

y,z

3×10^7 slots

Apriori algorithm

- F_i : frequent itemsets of size i — Level i
- F_1: Scan T, maintain a counter for each $x \in I$
- $C_2 = \{\{x, y\} \mid x, y \in F_1\}$: Candidates in level 2
- F_2: Scan T, maintain a counter for each $X \in C_2$
- $C_3 = \{\{x, y, z\} \mid \{x, y\}, \{x, z\}, \{y, z\} \in F_2\}$
- F_3: Scan T, maintain a counter for each $X \in C_3$
- ...
- C_k = subsets of size k, every $(k-1)$-subset is in F_{k-1}
- F_k: Scan T, maintain a counter for each $X \in C_k$
- ...

$x_1 \quad x_2 \quad \cdots \quad x_n$

Don't have to separately check $\{x\} \in F_1$, $\{y\} \in F_1$, $\{z\} \in F_1$

Apriori algorithm

- C_k = subsets of size k, every $(k-1)$-subset is in F_{k-1}
- How do we generate C_k?

Apriori algorithm

- C_k = subsets of size k, every $(k-1)$-subset is in F_{k-1}
- How do we generate C_k?
- Naïve: enumerate subsets of size k and check each one
  - Expensive!

$|I| = N$ $\binom{N}{k}$

$\frac{N!}{k!\,(N-k)!}$

Apriori algorithm

- C_k = subsets of size k, every $(k-1)$-subset is in F_{k-1}
- How do we generate C_k?
- Naïve: enumerate subsets of size k and check each one
  - Expensive!
- Observation: Any $C'_k \supseteq C_k$ will do as a candidate set

Goal

$F_k \subseteq C_k \subseteq C'_k$

count

don't miss any $x \in F_k$

Apriori algorithm

- C_k = subsets of size k, every $(k-1)$-subset is in F_{k-1}
- How do we generate C_k?
- Naïve: enumerate subsets of size k and check each one
  - Expensive!
- Observation: Any $C'_k \supseteq C_k$ will do as a candidate set
- Items are ordered: $i_1 < i_2 < \cdots < i_N$
- List each itemset in ascending order — canonical representation

Apriori algorithm

- C_k = subsets of size k, every $(k-1)$-subset is in F_{k-1}
- How do we generate C_k?
- Naïve: enumerate subsets of size k and check each one
  - Expensive!
- Observation: Any $C'_k \supseteq C_k$ will do as a candidate set
- Items are ordered: $i_1 < i_2 < \cdots < i_N$
- List each itemset in ascending order — canonical representation
- Merge two $(k-1)$-subsets if they differ in last element
  - $X = \{i_1, i_2, \ldots, i_{k-2}, i_{k-1}\}$
  - $X' = \{i_1, i_2, \ldots, i_{k-2}, i'_{k-1}\}$
  - $\text{Merge}(X, X') = \{i_1, i_2, \ldots, i_{k-2}, i_{k-1}, i'_{k-1}\}$

$k-1$ $\quad k$

$X \subseteq Z$

$Y \subseteq Z$

$k-1$

overlap on $k-2$ elements

i_1' $\quad i_k$

i_1

Apriori algorithm

- $\text{Merge}(X, X') = \{i_1, i_2, \ldots, i_{k-2}, i_{k-1}, i'_{k-1}\}$
  - $X = \{i_1, i_2, \ldots, i_{k-2}, i_{k-1}\}$
  - $X' = \{i_1, i_2, \ldots, i_{k-2}, i'_{k-1}\}$

Apriori algorithm

- $\text{Merge}(X, X') = \{i_1, i_2, \ldots, i_{k-2}, i_{k-1}, i'_{k-1}\}$
  - $X = \{i_1, i_2, \ldots, i_{k-2}, i_{k-1}\}$
  - $X' = \{i_1, i_2, \ldots, i_{k-2}, i'_{k-1}\}$
- $C'_k = \{\text{Merge}(X, X') \mid X, X' \in F_{k-1}\}$

$F_k \subseteq C_k \overset{?}{\subseteq} C'_k$

$F_1 \times F_2 \rightarrow C_2$

$F_{k-1} \times F_{k-1} \rightarrow C'_k$

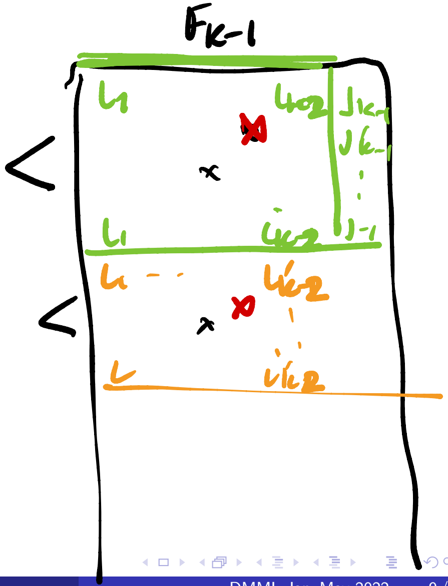
Apriori algorithm

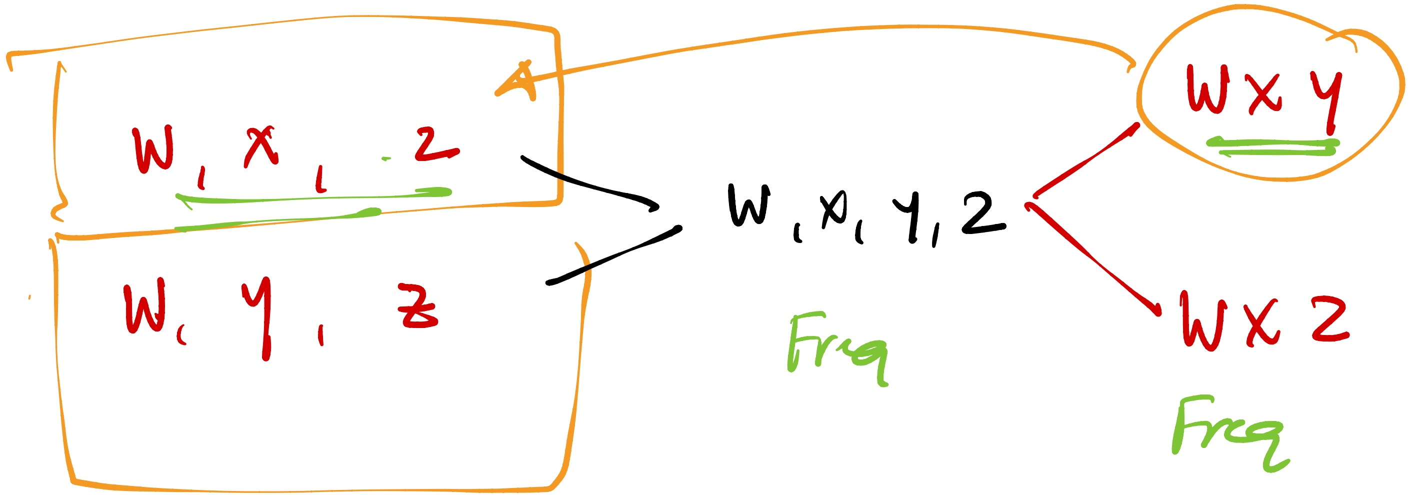
- $\text{Merge}(X, X') = \{i_1, i_2, \ldots, i_{k-2}, i_{k-1}, i'_{k-1}\}$
  - $X = \{i_1, i_2, \ldots, i_{k-2}, i_{k-1}\}$
  - $X' = \{i_1, i_2, \ldots, i_{k-2}, i'_{k-1}\}$
- $C'_k = \{\text{Merge}(X, X') \mid X, X' \in F_{k-1}\}$
- Claim $C_k \subseteq C'_k$
  - Suppose $Y = \{i_1, i_2, \ldots, i_{k-1}, i_k\} \in C_k$

    Every k-1 subset $\in F_{k-1}$
  - $X = \{i_1, i_2, \ldots, i_{k-2}, i_{k-1}\} \in F_{k-1}$ and $X' = \{i_1, i_2, \ldots, i_{k-2}, i_k\} \in F_{k-1}$
  - $Y = \text{Merge}(X, X') \in C'_k$

Apriori algorithm

- Merge$(X, X') = \{i_1, i_2, \ldots, i_{k-2}, i_{k-1}, i'_{k-1}\}$
  - $X = \{i_1, i_2, \ldots, i_{k-2}, i_{k-1}\}$
  - $X' = \{i_1, i_2, \ldots, i_{k-2}, i'_{k-1}\}$
- $C'_k = \{\text{Merge}(X, X') \mid X, X' \in F_{k-1}\}$
- Claim $C_k \subseteq C'_k$
  - Suppose $Y = \{i_1, i_2, \ldots, i_{k-1}, i_k\} \in C_k$
  - $X = \{i_1, i_2, \ldots, i_{k-2}, i_{k-1}\} \in F_{k-1}$ and $X' = \{i_1, i_2, \ldots, i_{k-2}, i_k\} \in F_{k-1}$
  - $Y = \text{Merge}(X, X') \in C'_k$
- Can generate C'_k efficiently
  - Arrange F_{k-1} in dictionary order
  - Split into blocks that differ on last element
  - Merge all pairs within each block





W, X, Z
W, Y, Z
W, X, Y, Z
Freq
W X Y
W X Z
Freq

Apriori algorithm

- $C_1 = \{\{x\} \mid x \in I\}$
- $F_1 = \{Z \mid Z \in C_1, Z.\text{count} \geq \sigma \cdot M\}$
- For $k \in \{2, 3, \ldots\}$
  - $C'_k = \{\text{Merge}(X, X') \mid X, X' \in F_{k-1}\}$
  - $F_k = \{Z \mid Z \in C'_k, Z.\text{count} \geq \sigma \cdot M\}$

1 - k-2 k-1

1 - k-2 $\{x\}$ $\{y\}$ → $\{x, y\}$

x, y
x, z → x, y, z

x, y
y, z ×

Apriori algorithm

- $C_1 = \{\{x\} \mid x \in I\}$
- $F_1 = \{Z \mid Z \in C_1, Z.\text{count} \geq \sigma \cdot M\}$
- For $k \in \{2, 3, \ldots\}$
  - $C'_k = \{\text{Merge}(X, X') \mid X, X' \in F_{k-1}\}$
  - $F_k = \{Z \mid Z \in C'_k, Z.\text{count} \geq \sigma \cdot M\}$
- When do we stop?

Apriori algorithm

- $C_1 = \{\{x\} \mid x \in I\}$
- $F_1 = \{Z \mid Z \in C_1, Z.\text{count} \geq \sigma \cdot M\}$
- For $k \in \{2, 3, \ldots\}$
  - $C'_k = \{\text{Merge}(X, X') \mid X, X' \in F_{k-1}\}$
  - $F_k = \{Z \mid Z \in C'_k, Z.\text{count} \geq \sigma \cdot M\}$
- When do we stop?
- k exceeds the size of the largest transaction
- F_k is empty

Apriori algorithm

- $C_1 = \{\{x\} \mid x \in I\}$
- $F_1 = \{Z \mid Z \in C_1, Z.\text{count} \geq \sigma \cdot M\}$
- For $k \in \{2, 3, \ldots\}$
  - $C'_k = \{\text{Merge}(X, X') \mid X, X' \in F_{k-1}\}$
  - $F_k = \{Z \mid Z \in C'_k, Z.\text{count} \geq \sigma \cdot M\}$
- When do we stop?
- k exceeds the size of the largest transaction
- F_k is empty

Next step: From frequent itemsets to association rules

$|T| \times |T|$

$c \cdot |T|$ smaller ✓

Also if k is "big enough" for application

Association rules

- Given sets of items I and transactions T, with confidence χ and support σ, find all valid association rules $X \rightarrow Y$
  - $X, Y \subseteq I$, $X \cap Y = \emptyset$
  - $\dfrac{(X \cup Y).count}{X.count} \geq \chi$
  - $\dfrac{(X \cup Y).count}{M} \geq \sigma$

Association rules

- Given sets of items I and transactions T, with confidence χ and support σ, find all valid association rules $X \rightarrow Y$
  - $X, Y \subseteq I$, $X \cap Y = \emptyset$
  - $\dfrac{(X \cup Y).count}{X.count} \geq \chi$
  - $\dfrac{(X \cup Y).count}{M} \geq \sigma$
- For a rule $X \rightarrow Y$ to be valid, $X \cup Y$ should be a frequent itemset
- Apriori algorithm finds all $Z \subseteq I$ such that $Z.\text{count} \geq \sigma \cdot M$

Association rules

Naïve strategy

- For every frequent itemset Z
  - Enumerate all pairs $X, Y \subseteq Z$, $X \cap Y = \emptyset$
  - Check $\dfrac{(X \cup Y).count}{X.count} \geq \chi$

Association rules

Naïve strategy

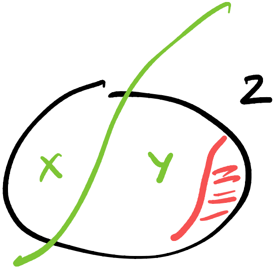
- For every frequent itemset Z
  - Enumerate all pairs $X, Y \subseteq Z$, $X \cap Y = \emptyset$
  - Check $\dfrac{(X \cup Y).count}{X.count} \geq \chi$

Can we do better?

Association rules

Naïve strategy

- For every frequent itemset Z
  - Enumerate all pairs $X, Y \subseteq Z$, $X \cap Y = \emptyset$
  - Check $\dfrac{(X \cup Y).count}{X.count} \geq \chi$



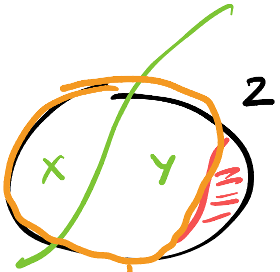
Can we do better?

- Sufficient to check all partitions of Z
  - Suppose $X, Y \subseteq Z$, $X \rightarrow Y$ is a valid association rule, but $X \cup Y$ is a proper subset of Z

Association rules

Naïve strategy

- For every frequent itemset Z
  - Enumerate all pairs $X, Y \subseteq Z$, $X \cap Y = \emptyset$
  - Check $\dfrac{(X \cup Y).count}{X.count} \geq \chi$



Can we do better?

- Sufficient to check all partitions of Z
  - Suppose $X, Y \subseteq Z$, $X \to Y$ is a valid association rule, but $X \cup Y$ is a proper subset of Z
  - $X \cup Y = Z' \subsetneq Z$
  - Z' is also a frequent itemset (a priori)
  - X, Y partitions Z'

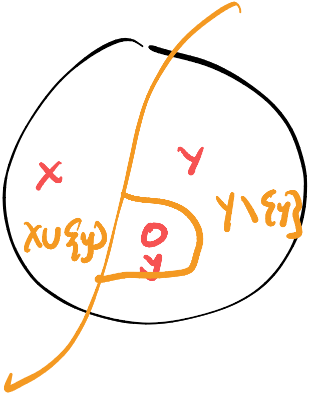
$Z' \subsetneq Z$

Association rules

- Sufficient to check all partitions of Z

  disjoint union

- Suppose $Z = X \uplus Y$, $X \to Y$ is a valid rule and $y \in Y$
- What about $(X \cup \{y\}) \to Y \setminus \{y\}$?



Association rules

- Sufficient to check all partitions of Z
- Suppose $Z = X \uplus Y$, $X \to Y$ is a valid rule and $y \in Y$
- What about $(X \cup \{y\}) \to Y \setminus \{y\}$?
  - Know $\dfrac{(X \cup Y).count}{X.count} \geq \chi$
  - Check $\dfrac{(X \cup Y).count}{(X \cup \{y\}).count} \geq \chi$

$(X \cup \{y\}) \cup (Y \setminus \{y\}) = X \cup Y$

$X \cup \{y\}.count \overset{\leq}{\underset{\geq}{=}} X.count$

Association rules

- Sufficient to check all partitions of Z

- Suppose $Z = X \uplus Y$, $X \to Y$ is a valid rule and $y \in Y$

- What about $(X \cup \{y\}) \to Y \setminus \{y\}$?

  - Know $\dfrac{(X \cup Y).count}{\underline{X.count}} \geq \chi$

  - Check $\dfrac{(X \cup Y).count}{\underline{(X \cup \{y\}).count}} \geq \chi$

  - X.count $\geq (X \cup \{y\})$.count, always

  - Second fraction has smaller denominator, so $(X \cup \{y\}) \to Y \setminus \{y\}$ is also a valid rule

$$\begin{tikzcd} X \arrow[r] & Y \arrow[l, bend left=60, "y"] \end{tikzcd}$$

Association rules

- Sufficient to check all partitions of Z
- Suppose $Z = X \uplus Y$, $X \to Y$ is a valid rule and $y \in Y$
- What about $(X \cup \{y\}) \to Y \setminus \{y\}$?
  - Know $\dfrac{(X \cup Y).count}{X.count} \geq \chi$
  - Check $\dfrac{(X \cup Y).count}{(X \cup \{y\}).count} \geq \chi$
  - $X.\text{count} \geq (X \cup \{y\}).\text{count}$, always
  - Second fraction has smaller denominator, so $(X \cup \{y\}) \to Y \setminus \{y\}$ is also a valid rule

Observation: Can use apriori principle again!

Apriori for association rules

- If $X \to Y$ is a valid rule, and $y \in Y$,
  $(X \cup \{y\}) \to Y \setminus \{y\}$ must also be a valid rule
- If $X \to Y$ is not a valid rule, and $x \in X$,
  $(X \setminus \{x\}) \to Y \cup \{x\}$ cannot be a valid rule

$X \to Y$ not valid

$\{x\}$ not valid

Apriori for association rules

- If $X \to Y$ is a valid rule, and $y \in Y$, $(X \cup \{y\}) \to Y \setminus \{y\}$ must also be a valid rule
- If $X \to Y$ is not a valid rule, and $x \in X$, $(X \setminus \{x\}) \to Y \cup \{x\}$ cannot be a valid rule
- Start by checking rules with single element on the right
  - $Z \setminus \{z\} \to \{z\}$
- For $X \to \{x, y\}$ to be a valid rule, both $(X \cup \{x\}) \to \{y\}$ and $(X \cup \{y\}) \to \{x\}$ must be valid
- Explore partitions of each frequent itemset "level by level"

Association rules for classification

- Classify documents by topic
- Consider the table on the right

Words in document	Topic
student, teach, school	Education
student, school	Education
teach, school, city, game	Education
cricket, football	Sports
football, player, spectator	Sports
cricket, coach, game, team	Sports
football, team, city, game	Sports

Association rules for classification

- Classify documents by topic
- Consider the table on the right
- Items are regular words and topics
- Documents are transactions — set of words and one topic

Words in document	Topic
student, teach, school	Education
student, school	Education
teach, school, city, game	Education
cricket, football	Sports
football, player, spectator	Sports
cricket, coach, game, team	Sports
football, team, city, game	Sports

Association rules for classification

- Classify documents by topic
- Consider the table on the right
- Items are regular words and topics
- Documents are transactions — set of words and one topic
- Look for association rules of a special form
  - $\{$student, school$\} \rightarrow \{$Education$\}$
  - $\{$game, team$\} \rightarrow \{$Sports$\}$

Words in document	**Topic**
student, teach, school	Education
student, school	Education
teach, school, city, game	Education
cricket, football	Sports
football, player, spectator	Sports
cricket, coach, game, team	Sports
football, team, city, game	Sports

Association rules for classification

- Classify documents by topic
- Consider the table on the right
- Items are regular words and topics
- Documents are transactions — set of words and one topic
- Look for association rules of a special form
  - $\{$student, school$\} \rightarrow \{$Education$\}$
  - $\{$game, team$\} \rightarrow \{$Sports$\}$
- Right hand side always a single topic
- Class Association Rules

Words in document	**Topic**
student, teach, school	Education
student, school	Education
teach, school, city, game	Education
cricket, football	Sports
football, player, spectator	Sports
cricket, coach, game, team	Sports
football, team, city, game	Sports

Summary

- Market-basket analysis searches for correlated items across transactions
- Formalized as association rules
- Apriori principle helps us to efficiently
  - identify frequent itemsets, and
  - split these itemsets into valid rules
- Class association rules — simple supervised learning model

Refinements

Multiple classes
Sequences
Varying thresholds

See Liu