Foundation of Data Science and Analytics

Association Analysis: Advance Concepts

Material Adaptation:

Introduction to Data Mining, By Tan, Steinbach, Karpatne, Kumar

Continuous and Categorical Attributes

How to apply association analysis to non-asymmetric binary variables?

Gender	...	Age	Annual Income	No of hours spent online per week	No of email accounts	Privacy Concern
Female	...	26	90K	20	4	Yes
Male	...	51	135K	10	2	No
Male	...	29	80K	10	3	Yes
Female	...	45	120K	15	3	Yes
Female	...	31	95K	20	5	Yes
Male	...	25	55K	25	5	Yes
Male	...	37	100K	10	1	No
Male	...	41	65K	8	2	No
Female	...	26	85K	12	1	No
...	...	...	...	...	...	...

Example of Association Rule:

{Gender=Male, Age $\in$ [21,30)} $\rightarrow$ {No of hours online $\geq$ 10}

Handling Categorical Attributes

- Example: Internet Usage Data

Gender	Level of Education	State	Computer at Home	Online Auction	Chat Online	Online Banking	Privacy Concerns
Female	Graduate	Illinois	Yes	Yes	Daily	Yes	Yes
Male	College	California	No	No	Never	No	No
Male	Graduate	Michigan	Yes	Yes	Monthly	Yes	Yes
Female	College	Virginia	No	Yes	Never	Yes	Yes
Female	Graduate	California	Yes	No	Never	No	Yes
Male	College	Minnesota	Yes	Yes	Weekly	Yes	Yes
Male	College	Alaska	Yes	Yes	Daily	Yes	No
Male	High School	Oregon	Yes	No	Never	No	No
Female	Graduate	Texas	No	No	Monthly	No	No
...	...	...	...	...	...	...	...

{Level of Education=Graduate, Online Banking=Yes}
$\rightarrow$ {Privacy Concerns = Yes}

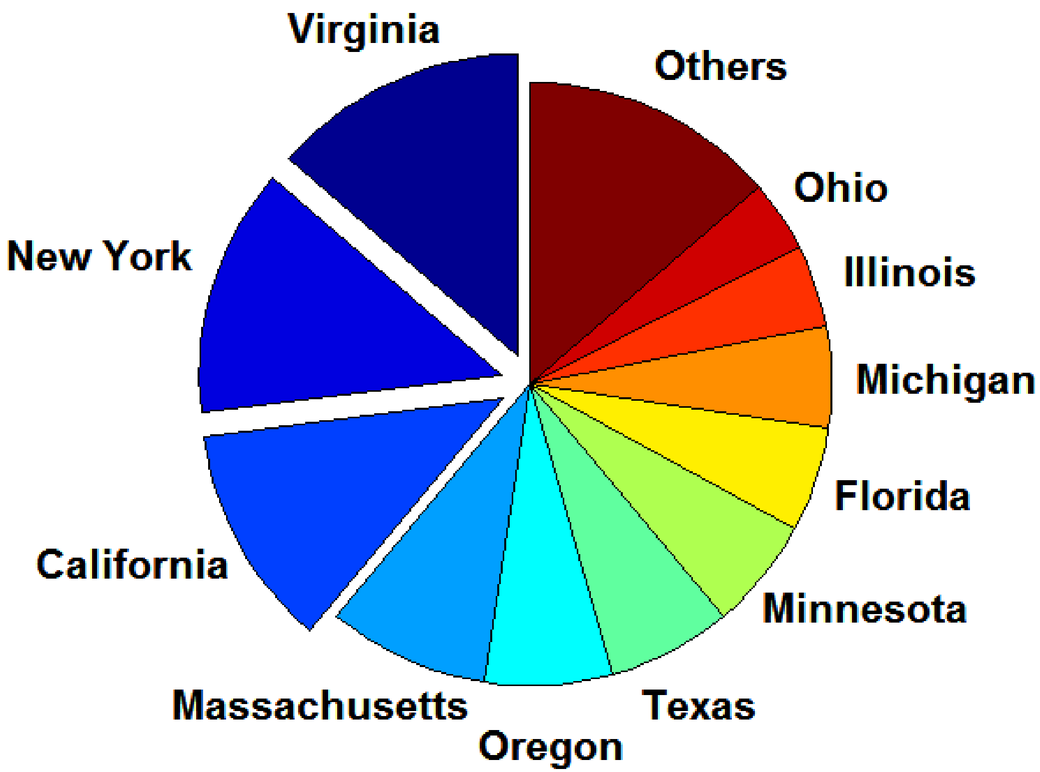
Handling Categorical Attributes

- Introduce a new "item" for each distinct attribute-value pair

Male	Female	Education = Graduate	Education = College	Education = High School	...	Privacy = Yes	Privacy = No
0	1	1	0	0	...	1	0
1	0	0	1	0	...	0	1
1	0	1	0	0	...	1	0
0	1	0	1	0	...	1	0
0	1	1	0	0	...	1	0
1	0	0	1	0	...	1	0
1	0	0	0	0	...	0	1
1	0	0	0	1	...	0	1
0	1	1	0	0	...	0	1
...	...	...	...	...	...	...	...

Handling Categorical Attributes

- Some attributes can have many possible values
  - Many of their attribute values have very low support
    - Potential solution: Aggregate the low-support attribute values



Handling Categorical Attributes

- Distribution of attribute values can be highly skewed
  - Ex.: 85% of survey participants own a computer at home
    - Most records have Computer at home = Yes
    - Computation becomes expensive; many frequent itemsets involving the binary item (Computer at home = Yes)
    - Potential solution:
      - discard the highly frequent items
      - Use alternative measures such as h-confidence
- Computational Complexity
  - Binarizing the data increases the number of items
  - But the width of the “transactions” remain the same as the number of original (non-binarized) attributes
  - Produce more frequent itemsets but maximum size of frequent itemset is limited to the number of original attributes

Handling Continuous Attributes

- Different methods:
  - Discretization-based
  - (Other methods; NOT Discussed)

- Different kinds of rules can be produced:
  - {Age$\in$[21,30), No of hours online$\in$[10,20)} $\rightarrow$ {Chat Online =Yes}
  - {Age$\in$[15,30), Covid-Positive = Yes} $\rightarrow$ Full_recovery

Discretization-based Methods

Gender	···	Age	Annual Income	No of hours spent online per week	No of email accounts	Privacy Concern
Female	···	26	90K	20	4	Yes
Male	···	51	135K	10	2	No
Male	···	29	80K	10	3	Yes
Female	···	45	120K	15	3	Yes
Female	···	31	95K	20	5	Yes
Male	···	25	55K	25	5	Yes
Male	···	37	100K	10	1	No
Male	···	41	65K	8	2	No
Female	···	26	85K	12	1	No
···	···	···	···	···	···	···

Male	Female	··· ···	Age < 13	Age $\in [13, 21)$	Age $\in [21, 30)$	··· ···	Privacy $=$ Yes	Privacy $=$ No
0	1	···	0	0	1	···	1	0
1	0	···	0	0	0	···	0	1
1	0	···	0	0	1	···	1	0
0	1	···	0	0	0	···	1	0
0	1	···	0	0	0	···	1	0
1	0	···	0	0	1	···	1	0
1	0	···	0	0	0	···	0	1
1	0	···	0	0	0	···	0	1
0	1	···	0	0	1	···	0	1
···	···	···	···	···	···	···	···	···

Discretization-based Methods

- Unsupervised:
  - Equal-width binning   <1 2 3> <4 5 6> <7 8 9>
  - Equal-depth binning   <1 2 > <3 4 5 6 7 > < 8 9>
  - Cluster-based

- Supervised discretization

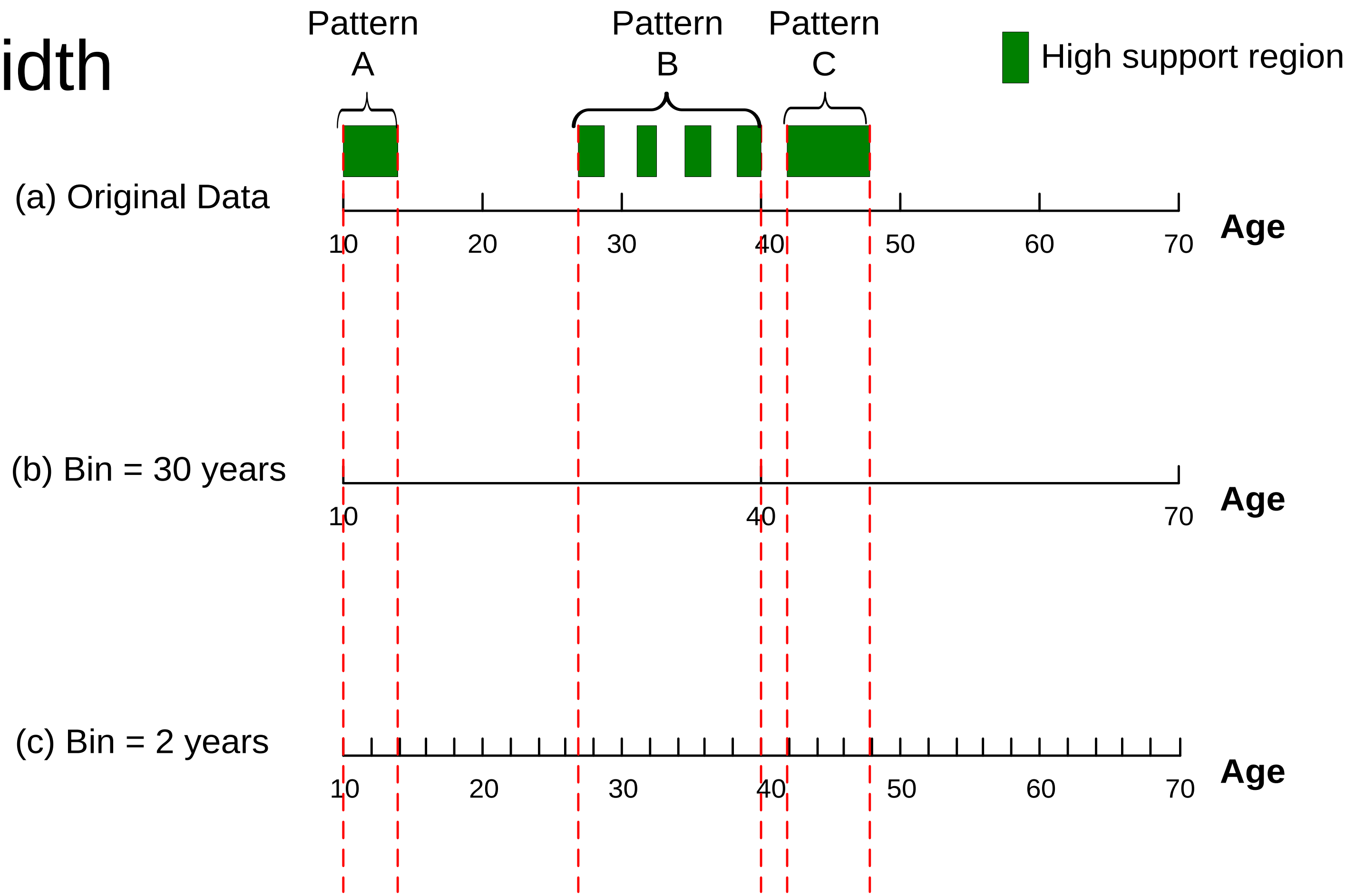
Continuous attribute, v

	1	2	3	4	5	6	7	8	9
Chat Online = Yes	0	0	20	10	20	0	0	0	0
Chat Online = No	150	100	0	0	0	100	100	150	100

bin_1 bin_2 bin_3

Discretization Issues

- Interval width



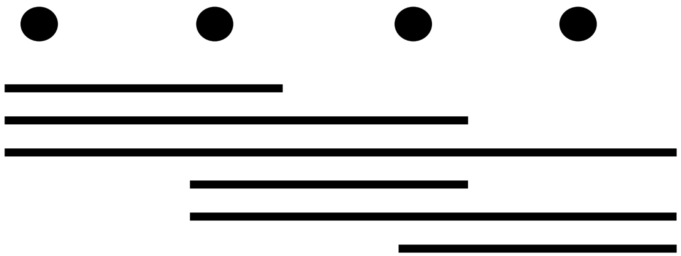
Pattern A: Age $\in [10, 15) \longrightarrow$ Chat Online = Never
Pattern B: Age $\in [26, 41) \longrightarrow$ Chat Online = Never
Pattern C: Age $\in [42, 48) \longrightarrow$ Online Banking = Yes

Discretization Issues

- Interval too wide (e.g., Bin size= 30)
  - May merge several disparate patterns
    - Patterns A and B are merged together
  - May lose some of the interesting patterns
    - Pattern C may not have enough confidence
- Interval too narrow (e.g., Bin size = 2)
  - Pattern A is broken up into two smaller patterns
    - Can recover the pattern by merging adjacent subpatterns
  - Pattern B is broken up into smaller patterns
    - Cannot recover the pattern by merging adjacent subpatterns
  - Some windows may not meet support threshold

Discretization: all possible intervals

Number of intervals = k
Total number of Adjacent intervals = k(k-1)/2

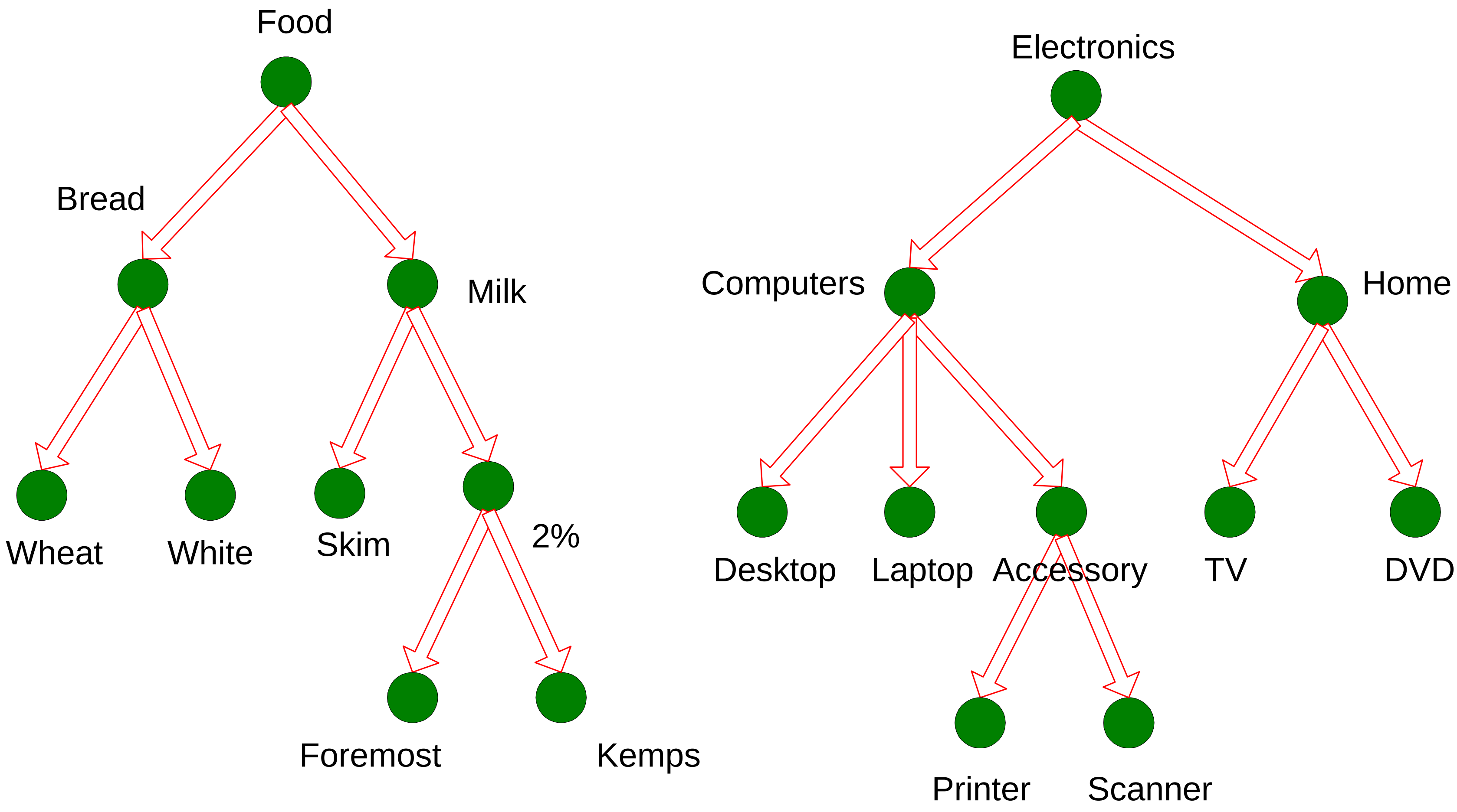


- Execution time
  - If the range is partitioned into k intervals, there are $O(k^2)$ new items
  - If an interval [a,b) is frequent, then all intervals that subsume [a,b) must also be frequent
    - E.g.: if {Age $\in$[21,25), Chat Online=Yes} is frequent,
      then {Age $\in$[10,50), Chat Online=Yes} is also frequent
  - Improve efficiency:
    - Use maximum support to avoid intervals that are too wide

Multi-level Association Rules

- Why should we incorporate concept hierarchy?
  - Rules at lower levels may not have enough support to appear in any frequent itemsets
  - Rules at lower levels of the hierarchy are overly specific
    - e.g., following rules are indicative of association between milk and bread
      - skim milk $\rightarrow$ white bread,
      - 2% milk $\rightarrow$ wheat bread,
      - skim milk $\rightarrow$ wheat bread, etc.
  - Rules at higher level of hierarchy may be too generic
    - e.g., electronics $\rightarrow$ food

Concept Hierarchies



Multi-level Association Rules

- How do support and confidence vary as we traverse the concept hierarchy?

  – If $\sigma(X1 \cup Y1) \geq \text{minsup}$,
  and X is parent of X1, Y is parent of Y1
  then $\sigma(X \cup Y1) \geq \text{minsup}$, $\sigma(X1 \cup Y) \geq \text{minsup}$
  $\sigma(X \cup Y) \geq \text{minsup}$

  – If $\text{conf}(X1 \Rightarrow Y1) \geq \text{minconf}$,
  then $\text{conf}(X1 \Rightarrow Y) \geq \text{minconf}$

Multi-level Association Rules

- Approach 1:
  - Extend current association rule formulation by augmenting each transaction with higher level items

Original Transaction: {skim milk, wheat bread}

Augmented Transaction:
{skim milk, wheat bread, milk, bread, food}

- Issues:
  - Items that reside at higher levels have much higher support counts
    - if support threshold is low, too many frequent patterns involving items from the higher levels
  - Increased dimensionality of the data

Multi-level Association Rules

- Approach 2:
  - Generate frequent patterns at highest level first
  - Then, generate frequent patterns at the next highest level, and so on
- Issues:
  - I/O requirements will increase dramatically because we need to perform more passes over the data
  - May miss some potentially interesting cross-level association patterns