Lecture 2: Market-Basket Analysis

Madhavan Mukund

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

Data Mining and Machine Learning
August–December 2020

Market-Basket Analysis

- People who buy X also tend to buy Y
- Rearrange products on display based on customer patterns
  - The diapers and beer legend
  - The true story, `http://www.dssresources.com/newsletters/66.php`
- Applies in more abstract settings
  - Items are concepts, basket is a set of concepts in which a student does badly
    - Students with difficulties in concept A also tend to do misunderstand concept B
  - Items are words, transactions are documents

Formal setting

- Set of items $I = \{i_1, i_2, \ldots, i_N\}$
- A transaction is a set $t \subseteq I$ of items
- Set of transactions $T = \{t_1, t_2, \ldots, t_M\}$
- 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 $Z \subseteq I$, $Z.\text{count} = |\{t_j \mid Z \subseteq t_j\}|$
- 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 $X \rightarrow Y$

Frequent itemsets

- $X \to Y$ 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 transacation 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 that is frequent is an item that appears in at least 10^7 transactions
- Totally, T contains at most 10^{10} items
- At most $10^{10}/10^7 = 1000$ items are frequent!
- How can we exploit this?