13a_AssciationRuleOverview

January 18, 2024

1 Market Basket Analysis/Association Rules Mining Overview

1.1 What is it?

- Searching for common items(pairs, tuples) in transactions
  - frequent itemsets
- If we see A in a transaction, we also see B x% of the time
  - rules/assciation rules
- Where do we use it
  - Often when *events* can be grouped together
    - Shopping transaction contains many items
      - Eggs, fish, bread, etc.
    - Observing any of these items in a transaction we can think of as an event (probability)
    - We may be interested in understanding which events tend to happen together
      - e.g. Which items tend to be bought together
      - From this, we can derive insights
  - Other examples:
    - Pages visited on a website (in a browsing session)
      - % of people who go to the homepage & the about page in the same session
    - Customer service touchpoints (fo a single query)
      - % who use the chatbot
      - % who phoned up
- Identifying frequent itemsets
  - An itemset is a set of items
  - i.e. some events that occur together
  - e.g. $\{Bread, Cheese\}$
  - We often want to know how many itemsets CONTAIN a specific itemset
- Rules
  - A rule is a statement of the form $\{X\} \rightarrow \{Y\}$
    - $\{X\}$ denotes an itemset containing item X
    - A transaction when shopping may contain many items, e.g. $\{eggs, bread, fish, ...\}$
    - A rule is of the form $\{X\} \rightarrow \{Y\}$
      - To be read, when we see X, we also see Y ... (some % of the time, with some degree of confidence)
    - e.g. $\{eggs\} \rightarrow \{bread\}, 0.8$
      - When we see eggs in a transaction, we also see bread 80% of the time
    - e.g. $\{eggs, butter\} \rightarrow \{bread\}, 0.7$

- When we see eggs AND butter in a transaction, we also see bread 70% of the time

1.1.1 Metrics

1.2 Support

- Proportion of all itemsets containing a specific itemset
  - e.g. $Support(\{X\})$ is the proportion of all itemsets containing the item X
  - e.g. $Support(\{X, Y\})$ is the proportion of all itemsets containing the itemset (i.e. both items) X, Y
- $Support(\{X\}) = \frac{\#\text{ containing X}}{\#\text{ total number of itemsets}}$
- If $Support(\{Bread\}) = 0.8$ then 80% of transaction contain *bread*
- Support = Probability of observing X

1.3 Confidence

- Tells us the proportion of the time that, when we see X, we go on to see Y
  - $Confidence(\{X\} \rightarrow \{Y\})$ is proportion of the time that, when we see X, we go on to see Y
  - e.g. $Confidence(\{Bread\} \rightarrow \{Eggs\})$ is proportion of the time that, when we see $Bread$, we go on to see $Eggs$

$Confidence(\{X\} \rightarrow \{Y\}) = \frac{\#\text{ containing X and Y}}{\#\text{ containing X}} = \frac{Support(\{X,Y\})}{Support(\{X\})}$

- e.g. $Confidence(\{Bread\} \rightarrow \{Eggs\}) = 1$ means that whenever someone purchases Bread, they also purchase eggs
- e.g. $Confidence(\{Cheese\} \rightarrow \{Wine\}) = 0$ means that whenever someone purchases Cheese, they never purchase wine
- e.g. $Confidence(\{Milk\} \rightarrow \{Flour\}) = 0.5$ means that whenever someone purchases Milk, they also purchase flour 50% of the time

1.4 Lift

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

- Numerator = Proportion of the time we actually see X & Y together
- Denomenator = Proportion of the time we would expect to see X & Y together IF they were unrelated (independent)
- $\frac{\text{Actual times we see X, Y}}{\text{Expected amount to see X, Y}}$
- Lift tells us how much more (or less) often we see X and Y together than we would expect to if they were unrelated (independent)

- e.g. $Lift(\{Bread\} \rightarrow \{Eggs\}) = 2$ means that we see $Bread$ and $Eggs$ together 2 times more often than we would expect to if they were unrelated (independent)
- e.g. $Lift(\{Cheese\} \rightarrow \{Wine\}) = 1$ means that we see $Cheese$ and $Wine$ together as often as we would expect to if they were unrelated
- e.g. $Lift(\{Milk\} \rightarrow \{Flour\}) = 0.5$ means that we see $Milk$ and $Flour$ together 0.5 times as often as we would expect to if they were unrelated

1.5 Goals

- Association => Not necessarily causal rule mining
- We need ways to do the following
  - Come up with good rules
  - Come up with sets of items that frequently occur together
  - Evaluate strength of rules
    - Support (= proportion of transaction containing itemset == probability of observing set)
    - Confidence (Evaluate strength of rule == conditional probability)
    - Lift (How much more than randomly do we see a rule happen)

[ ]: