Apriori based & Hybrid RS

Recommendation System

Agenda

- Market basket analysis
- Common terms
- Association rule
- Apriori algorithm
- Hybrid Methods
- Evaluation metrics
- Key points

Recommendation systems (TOC)

Sr. No.	Topic	Scope	Objective
1	Market basket analysis	To understand market basket analysis, uses, examples	
2	Important terms	Itemset, support, support count, confidence, lift	
3	Association rule	Definition, evaluation metrics	
4	Apriori algorithm	Theory and its application	
5	Hybrid model	What is hybrid, how to build a hybrid model, it's application and advantages	
6	Evaluation metrics	To be able to evaluate a recommendation model, rmse, mae, accuracy	

Market basket analysis

- Uncovers association between items.
- Identifies pattern of co-occurrence
- Market basket analysis may provide the retailer with information **to understand the behaviour of a buyer.**

“Customers who bought book A also bought book B”

Examples :

- If a user buys pizza then he is more likely to buys cold drinks also
- One supermarket chain discovered in its analysis that male customers that bought diapers often bought beer as well.

Market basket analysis

- Used to increase profitability through cross-selling, promotions
- Can be used to recommend more relevant item
- Discounts schemes can be used to increase sales.
- Relationship between item is modeled using conditional algorithm
- Applies If-then scenario rules

Important terms

1. Itemset - a collection of items purchased by a customer
   a. Ex - {Pizza, pepsi, garlic bread}
2. Support count (σ)- Frequency of occurrence of an itemset.
   a. Ex- σ(Pizza, pepsi, garlic bread) = 2
3. Support - fraction of transaction that contains itemset
   a. Ex- S(Pizza, pepsi, garlic bread) = ⅖
4. Frequent Itemset – An itemset whose support is greater than or equal to a *min_sup* threshold

ID	Item
1	Pizza, wrap
2	Pizza, garlic bread, pepsi
3	Garlic bread, pizza, cake, pepsi
4	Garlic bread, wrap, cake
5	Pizza, pepsi, cake

Association rule

1. Association rule - An implication expression of the form X → Y, where X and Y are item sets.
   a. If {pizza, pepsi} Then {garlic bread}
2. Support (s) - fraction of transaction that contains both X and Y.
   a. S = (σ(pizza, pepsi, garlic bread))/|T| = 2/5=0.4
3. Confidence (c) - measures how often items in Y appears in transaction that contains X. The probability that a customer will purchase an item on the condition of purchasing another item/items is referred to as the **confidence** of the rule.
   a. confidence(c)=(σ(pizza, pepsi, garlic bread))/(σ(pizza, pepsi))= 2/3=0.66

Association rule

The **lift** of the rule is the ratio of the support of the left-hand side of the rule (pizza, pepsi) co-occurring with the right-hand side (garlic bread), divided by the probability that the left-hand side and right-hand side co-occur if the two are independent.

$$\boldsymbol{lift\ (A \rightarrow B)} = \frac{confidence(A \rightarrow B)}{support\,(B)} = \frac{support\,(A \text{ and } B)}{support\ (A).\,support\,(B)}$$

Apriori algorithm

Idea:

- Set a min. support and confidence
- Take all the subsets in transactions having higher support than min. Support
- Take all the rules of these subsets having higher confidence than min. Confidence
- Sort the rules by decreasing ***lift***

Apriori algorithm example (1/2)

- Items set : {1, 2, 3, 4, 5}
- Combinations:

  {1}, {2}, {3}, {4}, {5}
  {1, 2}, {1, 3}, {1, 4} ……….{4,5}
  {1, 2, 3}, {1, 2, 4}…….
  {1, 2, 3, 4}, {1, 2, 3, 5}……..
  {1, 2, 3, 4, 5}

- Support(1) = 2/4 = 0.5, support(2) = 3/4 = 0.75
- Confidence (1 ->2) = 1/ 2 = 0.5
- Lift(1 ->2) = 0.5/0.75 = 0.6667

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Apriori algorithm example (2/2)

Transaction D

TID	Items
100	1 3 4
200	2 3 5
300	1 2 3 5
400	2 5

Scan D

C_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3

L_1

itemset	sup.
{1}	2
{2}	3
{3}	3
{5}	3

C_2

itemset
{1 2}
{1 3}
{1 5}
{2 3}
{2 5}
{3 5}

Scan D

C_2

itemset	sup
{1 2}	1
{1 3}	2
{1 5}	1
{2 3}	2
{2 5}	3
{3 5}	2

L_2

itemset	sup
{1 3}	2
{2 3}	2
{2 5}	3
{3 5}	2

C_3

itemset
{2 3 5}

Scan D

L_3

itemset	sup
{2 3 5}	2

Min support - 02

Hybrid Algorithm

- Combination of multiple algorithm
- Customized algorithm

Approaches :

1. A common approach is to combine content based approaches and collaborative filtering approaches.
2. Popularity based recommendation can be customized
3. Content based models solve **the cold start and Gray sheep whereas collaborative filtering methods solve diversity and privacy issues.**

Methods

Some typical methods of hybridization include

- **Weighted** – Each system is weighted to calculate final recommendation
- **Switching** – System switches between different recommendation model
- **Mixed** – Recommendations from different models are presented together.
- A common approach is to use **Latent Factor models** for high level recommendation and then improving them using **content based** systems by using information on users or items

Evaluation metrics

- User satisfaction
- Prediction accuracy
- Coverage
- Diversity
- Novelty
- Trust
- Robust
- Real Time

Evaluation metrics (1/ 2)

- **User Satisfaction**
  - Subjective metric
  - Measured by user survey or online experiments
- **Prediction Accuracy**
  - Rating Prediction (MAE, RMSE)
  - Top-N Recommendation (Precision, Recall)
- **Coverage**
  - Ability to recommend long tail items (entropy, gini index)
- **Diversity**
  - Ability to cover user’s different interests

Evaluation metrics (2/2)

- **Novelty -** Ability of Recommendation system to recommend long tail items and new items.
- **Trust -** Trust increases the interaction of user to recommendation system.
    - Transparency, social
- **Robust -** Ability of Recommendation system to prevent attacks.
    - Shilling attack
- **Real Time -** Generate new recommendation when user has new behaviours immediately.

Prediction accuracy metrics

MAE: Mean Absolute Error is the average of the absolute difference between the predictions and actual values.

$$MAE = \frac{1}{N}\sum_{i=1,}^{m}\sum_{j=1}^{n}\left|r_{i,j} - \widehat{r_{i,j}}\right|$$

RMSE: Root Mean Square Error computed by the square root of the average of the difference between predictions and actual values. Lower the RMSE is better the recommendation.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1,}^{m}\sum_{j=1}^{n}\left(r_{i,j} - \widehat{r_{i,j}}\right)^{2}}$$

Classification accuracy metrics

1. Confusion matrix
2. Precision - A measure of exactness, determines the fraction of relevant items retrieved out of all items retrieved.

   Precision= TP/(TP+FP)
1. Recall - A measure of completeness, determines the fraction of relevant items retrieved out of all relevant items.

   Recall= TP/(TP+FN)

1. F-measure -Harmonic mean of precision and recall to get a single value for comparison purpose.

   F-measure = (2 ∗precision ∗recall)/(precision+recall)