Association Rules

Data Mining
Prof. Dawn Woodard
School of ORIE
Cornell University

Reading

Reading is Chap.13 in SPB

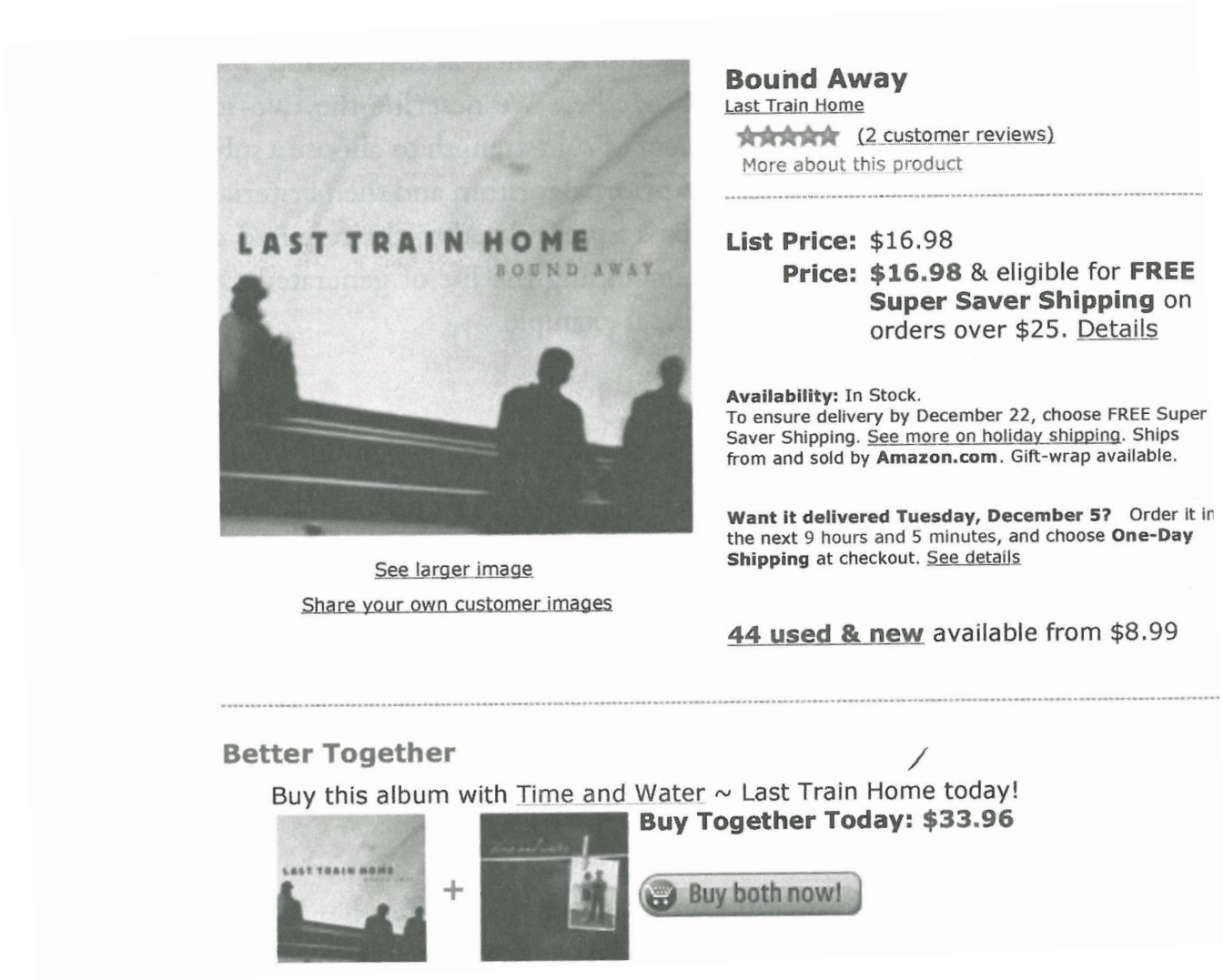
Association Rules

Association Rules:

- Goal: Identify groups of items that are often occur together
- For example, a retailer may wish to identify groups of products that are often purchased together (“**Market basket analysis**”). Can be used for, e.g.:
  -
  -
- Can be used to recommend products to a user (“**Recommender systems**”), as done by Netflix, Amazon.com

Recommender Systems

From Amazon.com (Fig. from SPB text):



Association Rules

- An enormous amount of purchase data is generated automatically, for instance by supermarket scanners and online transaction records
- This can be used to perform market basket analysis for huge numbers of products, using transactions from huge numbers of customers
- Association rules provide information in the form of “if-then” statements
  -

Association Rules

A simple example from SPB:

- Sales data for cell phone faceplates
-

TABLE 13.1 TRANSACTIONS FOR PURCHASES OF DIFFERENT-COLORED CELLULAR PHONE FACEPLATES

Transaction	Faceplate	Colors	Purchased
1	red	white	green
2	white	orange	
3	white	blue	
4	red	white	orange
5	red	blue	
6	white	blue	
7	white	orange	
8	red	white	blue green
9	red	white	blue
10	yellow		

Association Rules

- The “if” part is called the **antecedent**, while the “then” part is called the **consequent**
- Examples of potential association rules for this dataset:
  -
  -

Association Rules

- Ideally, we could consider ALL POSSIBLE antecedent and consequent pairs
- We could then keep the association rules that have the highest value of some estimated quality metric.
- Problems:
  -
  -

Association Rules

A solution:

- Only consider antecedents and consequents with high **support**.
  The support of an item set means the
  proportion of transactions in which ALL those items are purchased.
  - For example:

Association Rules

The "**Apriori**" algorithm of Agrawal et al. (1993)

- Used for finding all item sets with support above some threshold $t \in (0, 1)$
- Very computationally efficient, even for huge numbers of items, provided that:

- Idea:

Association Rules

The **Apriori algorithm**:

1. Calculate the support for all one-item sets, and keep those $> t$. This is relatively fast since:

2. Considering only those items that meet this criterion, find what pairs of those items have support $> t$.
   - These form all the two-item sets with support $> t$, since:

3. Continue in this manner, generating the k-item sets for each k by adding one item to the $(k-1)$-item sets, and checking the support.

4. Stop when:

Association Rules

Now we can consider only association rules for which the UNION of the antecedent and the consequent is one of the high-support sets that we have identified.

We still need a measure of the quality of an association rule. We'll learn about 2:

- **Confidence**
- **Lift ratio**

Association Rules

Confidence:

- Measures how likely the consequent is, given the antecedent
- Specifically:

- For example:

- Confidence is an estimate of the conditional probability:

$$\Pr(\text{consequent}|\text{antecedent}) =$$

Association Rules

An association rule with high confidence suggests that the rule is strong.

However, a rule may have high confidence simply because the consequent is very common (not just common in conjunction with the antecedent).

- For example:

Association Rules

Lift ratio of an association rule:

- Compares the confidence to a benchmark value, namely the frequency of the consequent in the dataset.
- If the two events, {occurrence of the antecedent} and {occurrence of the consequent} are independent, then we should get a lift of about 1.
- Under independence, we have
- So the lift ratio is defined as:

$$\text{Lift ratio} = \frac{\text{confidence}}{\text{benchmark confidence}}$$

where

$$\text{benchmark confidence} =$$

- The higher the lift ratio, the greater the strength of the association.

Association Rules

Cell phone faceplate example: Item sets with support of at least 2/10:

Item Set	Support (Count)
{red}	6
{white}	7
{blue}	6
{orange}	2
{green}	2
{red, white}	4
{red, blue}	4
{red, green}	2
{white, blue}	4
{white, orange}	2
{white, green}	2
{red, white, blue}	2

Association Rules

Cell phone faceplate example: Some association rules with support of at least 2/10 and confidence of at least .7:

Rule #	Conf. %	Antecedent (a)	Consequent (c)	Support(a)	Support(c)	Support(a U c)	Lift Ratio
1	100	green=>	red, white	2	4	2	2.5
2	100	green=>	red	2	6	2	1.666667
3	100	green, white=>	red	2	6	2	1.666667
4	100	green=>	white	2	7	2	1.428571
5	100	green, red=>	white	2	7	2	1.428571
6	100	orange=>	white	2	7	2	1.428571

Association Rules

Charles Book Club case (from SPB):

- We have purchase transaction data for books, separated by category.
- a book club wants to know what types of books it should send to each reader

Association Rules

Charles Book Club case: Some association rules with support of at least 200/2000 and confidence of at least .5:

Rule #	Conf. %	Antecedent (a)	Consequent (c)	Support(a)	Support(c)	Support(a U c)	Lift Ratio
1	**100**	**ItalCook=>**	**CookBks**	**227**	**862**	**227**	**2.320186**
2	62.77	ArtBks, ChildBks=>	GeogBks	325	552	204	2.274247
3	54.13	CookBks, DoItYBks=>	ArtBks	375	482	203	2.246196
4	61.98	ArtBks, CookBks=>	GeogBks	334	552	207	2.245509
5	53.77	CookBks, GeogBks=>	ArtBks	385	482	207	2.230964
6	57.11	RefBks=>	ChildBks, CookBks	429	512	245	2.230842
7	52.31	ChildBks, GeogBks=>	ArtBks	390	482	204	2.170444
8	60.78	ArtBks, CookBks=>	DoItYBks	334	564	203	2.155264
9	58.4	ChildBks, CookBks=>	GeogBks	512	552	299	2.115885
10	54.17	GeogBks=>	ChildBks, CookBks	552	512	299	2.115885
11	57.87	CookBks, DoItYBks=>	GeogBks	375	552	217	2.096618
12	56.79	ChildBks, DoItYBks=>	GeogBks	368	552	209	2.057735
13	52.49	ArtBks=>	ChildBks, CookBks	482	512	253	2.050376
14	52.12	YouthBks=>	ChildBks, CookBks	495	512	258	2.035985
15	50.39	ChildBks, CookBks=>	YouthBks	512	495	258	2.035985
16	57.03	ChildBks, CookBks=>	DoItYBks	512	564	292	2.022385
17	51.77	DoItYBks=>	ChildBks, CookBks	564	512	292	2.022385
18	56.36	CookBks, GeogBks=>	DoItYBks	385	564	217	1.998711
19	52.9	ArtBks=>	GeogBks	482	552	255	1.916832
20	82.19	ArtBks, DoItYBks=>	CookBks	247	862	203	1.906873
21	53.59	ChildBks, GeogBks=>	DoItYBks	390	564	209	1.900346
22	81.89	DoItYBks, GeogBks=>	CookBks	265	862	217	1.899926
23	80.33	CookBks, RefBks=>	ChildBks	305	846	245	1.899004
24	80	ArtBks, GeogBks=>	ChildBks	255	846	204	1.891253
25	81.18	ArtBks, GeogBks=>	CookBks	255	862	207	1.883445
26	79.63	CookBks, YouthBks=>	ChildBks	324	846	258	1.882497

Association Rules

Some of these rules can be ignored or are closely related to other rules. For example:

-
-