14.3 Summary

Association rules (also called market basket analysis) and collaborative filtering are unsupervised methods for deducing associations between purchased items from databases of transactions. Association rules search for generic rules about items that are purchased together. The main advantage of this method is that it generates clear, simple rules of the form “IF *X* is purchased, THEN *Y* is also likely to be purchased.” The method is very transparent and easy to understand.

The process of creating association rules is two-staged. First, a set of candidate rules based on frequent itemsets is generated (the Apriori algorithm being the most popular rule-generating algorithm). Then from these candidate rules, the rules that indicate the strongest association between items are selected. We use the measures of support and confidence to evaluate the uncertainty in a rule. The user also specifies minimal support and confidence values to be used in the rule generation and selection process. A third measure, the lift ratio, compares the efficiency of the rule to detect a real association compared to a random combination.

One shortcoming of association rules is the profusion of rules that are generated. There is therefore a need for ways to reduce these to a small set of useful and strong rules. An important nonautomated method to condense the information involves examining the rules for uninformative and trivial rules as well as for rules that share the same support. Another issue that needs to be kept in mind is that rare combinations tend to be ignored, because they do not meet the minimum support requirement. For this reason, it is better to have items that are approximately equally frequent in the data. This can be achieved by using higher-level hierarchies as the items. An example is to use types of books rather than titles of individual books in deriving association rules from a database of bookstore transactions.

Collaborative filtering is a popular technique used in online recommendation systems. It is based on the relationship between items formed by users who acted similarly on an item, such as purchasing or rating an item highly. User-based collaborative filtering operates on data on item–user combinations, calculates the similarities between users and provides personalized recommendations to users. An important component for the success of collaborative filtering is that users provide feedback about the recommendations provided and have sufficient information on each item. One disadvantage of collaborative filtering methods is that they cannot generate recommendations for new users or new items. Also, with a huge number of users, user-based collaborative filtering becomes computationally challenging, and alternatives such as item-based methods or dimension reduction are popularly used.

Problems

1. **Satellite Radio Customers.** An analyst at a subscription-based satellite radio company has been given a sample of data from their customer database, with the goal of finding groups of customers who are associated with one another. The data consist of company data, together with purchased demographic data that are mapped to the company data (see Table 14.12). The analyst decides to apply association rules to learn more about the associations between customers. Comment on this approach.

Table 14.12 Sample of data on satellite radio customers

Row ID	zipconvert_2	zipconvert_3	zipconvert_4	zipconvert_5	homeowner	NUMCHLD	INCOME	gender	WE
					dummy			dummy	
17	0	1	0	0	1	1	5	1	
25	1	0	0	0	1	1	1	0	
29	0	0	0	1	0	2	5	1	
38	0	0	0	1	1	1	3	0	
40	0	1	0	0	1	1	4	0	
53	0	1	0	0	1	1	4	1	
58	0	0	0	1	1	1	4	1	
61	1	0	0	0	1	1	1	0	
71	0	0	1	0	1	1	4	0	
87	1	0	0	0	1	1	4	1	

100	0	0	0	1	1	1	4	1	
104	1	0	0	0	1	1	1	1	
121	0	0	1	0	1	1	4	1	
142	1	0	0	0	0	1	5	0	

2. **Identifying Course Combinations.** The Institute for Statistics Education at Statistics.com offers online courses in statistics and analytics, and is seeking information that will help in packaging and sequencing courses. Consider the data in the file *CourseTopics.csv*, the first few rows of which are shown in Table 14.13. These data are for purchases of online statistics courses at Statistics.com. Each row represents the courses attended by a single customer. The firm wishes to assess alternative sequencings and bundling of courses. Use association rules to analyze these data, and interpret several of the resulting rules.

   Table 14.13 Data on purchases of online statistics courses

Intro	DataMining	Survey	CatData	Regression	Forecast	DOE	SW
1	1	0	0	0	0	0	0
0	0	1	0	0	0	0	0
0	1	0	1	1	0	0	1
1	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0
0	1	0	0	0	0	0	0
1	0	0	0	0	0	0	0
0	0	0	1	0	1	1	1
1	0	0	0	0	0	0	0
0	0	0	1	0	0	0	0
1	0	0	0	0	0	0	0

3. **Recommending Courses** We again consider the data in *CourseTopics.csv* describing course purchases at Statistics.com (see Problem 14.2 and data sample in Table 14.13). We want to provide a course recommendation to a student who purchased the Regression and Forecast courses. Apply user-based collaborative filtering to the data. You will get a Null matrix. Explain why this happens.

4. **Cosmetics Purchases.** The data shown in Table 14.14 and the output in Table 14.15 are based on a subset of a dataset on cosmetic purchases (*Cosmetics.csv*) at a large chain drugstore. The store wants to analyze associations among purchases of these items for purposes of point-of-sale display, guidance to sales personnel in promoting cross-sales, and guidance for piloting an eventual time-of-purchase electronic recommender system to boost cross-sales. Consider first only the data shown in Table 14.14, given in binary matrix form.

   Table 14.14 Excerpt from data on cosmetics purchases in binary matrix form

Trans. #	Bag	Blush	Nail Polish	Brushes	Concealer	Eyebrow Pencils	Bronzer
1	0	1	1	1	1	0	1
2	0	0	1	0	1	0	1
3	0	1	0	0	1	1	1
4	0	0	1	1	1	0	1
5	0	1	0	0	1	0	1
6	0	0	0	0	1	0	0
7	0	1	1	1	1	0	1
8	0	0	1	1	0	0	1
9	0	0	0	0	1	0	0
10	1	1	1	1	0	0	0

11	0	0	1	0	0	0	1
12	0	0	1	1	1	0	1

Table 14.15 Association rules for cosmetics purchases data

```
   lhs              rhs                 support   confidence       lift
1 {Blush,
   Concealer,
   Mascara,
   Eye.shadow,
   Lipstick}   => {Eyebrow.Pencils}   0.013  0.3023255814  7.198228128
2 {Trans.,
   Blush,
   Concealer,
   Mascara,
   Eye.shadow,
   Lipstick}   => {Eyebrow.Pencils}   0.013  0.3023255814  7.198228128
3 {Blush,
   Concealer,
   Mascara,
   Lipstick}   => {Eyebrow.Pencils}   0.013  0.2888888889  6.878306878
4 {Trans.,
   Blush,
   Concealer,
   Mascara,
   Lipstick}   => {Eyebrow.Pencils}   0.013  0.2888888889  6.878306878
5 {Blush,
   Concealer,
   Eye.shadow,
   Lipstick}   => {Eyebrow.Pencils}   0.013  0.2826086957  6.728778468
6 {Trans.,
   Blush,
   Concealer,
   Eye.shadow,
   Lipstick}   => {Eyebrow.Pencils}   0.013  0.2826086957  6.728778468
```

a. Select several values in the matrix and explain their meaning.

b. Consider the results of the association rules analysis shown in Table 14.15.

   i. For the first row, explain the “confidence” output and how it is calculated.

   ii. For the first row, explain the “support” output and how it is calculated.

   iii. For the first row, explain the “lift” and how it is calculated.

   iv. For the first row, explain the rule that is represented there in words.

c. Now, use the complete dataset on the cosmetics purchases (in the file *Cosmetics.csv*). Using R, apply association rules to these data (use the default parameters).

   i. Interpret the first three rules in the output in words.

   ii. Reviewing the first couple of dozen rules, comment on their redundancy and how you would assess their utility.

5. **Course ratings.** The Institute for Statistics Education at Statistics.com asks students to rate a variety of aspects of a course as soon as the student completes it. The Institute is contemplating instituting a recommendation system that would provide students with recommendations for additional courses as soon as they submit their rating for a completed course. Consider the excerpt from student ratings of online statistics courses shown in Table 14.16, and the problem of what to recommend to student E.N.

   a. First consider a user-based collaborative filter. This requires computing correlations between all student pairs. For which students is it possible to compute correlations with E.N.? Compute them.

   b. Based on the single nearest student to E.N., which single course should we recommend to E.N.? Explain why.

   c. Use R (function *similarity()*) to compute the cosine similarity between users.

   d. Based on the cosine similarities of the nearest students to E.N., which course should be recommended to

d. Based on the cosine similarities of the nearest students to E.N., which course should be recommended to E.N.?

e. What is the conceptual difference between using the correlation as opposed to cosine similarities? [*Hint*: how are the missing values in the matrix handled in each case?]

f. With large datasets, it is computationally difficult to compute user-based recommendations in real time, and an item-based approach is used instead. Returning to the rating data (not the binary matrix), let's now take that approach.

  i. If the goal is still to find a recommendation for E.N., for which course pairs is it possible and useful to calculate correlations?

  ii. Just looking at the data, and without yet calculating course pair correlations, which course would you recommend to E.N., relying on item-based filtering? Calculate two course pair correlations involving your guess and report the results.

g. Apply item-based collaborative filtering to this dataset (using R) and based on the results, recommend a course to E.N.

Table 14.16 Ratings of online statistics courses: 4 = best, 1 = worst, blank = not taken

	SQL	Spatial	PA 1	DM in R	Python	Forecast	R Prog	Hadoop	Regression
L N	4				3	2	4		2
M H	3	4			4				
J H	2	2							
E N	**4**			**4**			**4**		**3**
D U	4	4							
F L		4							
G L		4							
A H		3							
S A			4						
R W			2					4	
B A			4						
M G			4			4			
A F			4						
K G			3						
D S	4			2			4		

Notes

[1] The number of rules that one can generate for *p* items is $3^p - 2^{p+1} + 1$. Computation time therefore grows by a factor for each additional item. For 6 items we have 602 rules, while for 7 items the number of rules grows to 1932.

[2] The concept of confidence is different from and unrelated to the ideas of confidence intervals and confidence levels used in statistical inference.

[3] This section copyright © 2017 Datastats, LLC, Galit Shmueli, and Peter Bruce.

[4] Bell, R. M., Koren, Y., and Volinsky, C., "The BellKor 2008 Solution to the Netflix Prize", www.netflixprize.com/assets/ProgressPrize2008_BellKor.pdf.

[5] Correlation and cosine similarity are popular in collaborative filtering because they are computationally fast for high-dimensional sparse data, and they account both for the rating values and the number of rated items.

[6] "Linden, G., Smith, B., and York J., Amazon.com Recommendations: Item-to-Item Collaborative Filtering", *IEEE Internet Computing*, vol. 7, no. 1, p. 76–80, 2003.

[7] If the rule is "IF milk, THEN cookies and cornflakes" then the association rules would recommend cookies and cornflakes to a milk purchaser, while item-based collaborative filtering would recommend the most popular single item purchased with milk.