

Session 5

Reading Apriori output

This session is a guided reading of one Apriori run so that every number in the output, from the run information to the large itemset counts, can be reproduced by hand.

Objectives

✕ Do not copy. Read for understanding and the viva

- Complete questions 12 to 12 of the manual: reading apriori output
- Prepare the deliverable before the lab and finish it during the session
- Be ready to explain every step in the viva

Questions Covered

✕ Do not copy. Read for understanding and the viva

Question	Requirement	Status
Q12	Perform the following	Complete

Preparation

✕ Do not copy. Read for understanding and the viva

- Contact-lenses has 24 instances and 5 nominal attributes; with default settings WEKA reaches minimum support 0.2 and shows 10 rules.
- For each rule, recompute support and confidence from the counts before and after the arrow using the formula sheet.
- Count the large itemsets of size 1, 2, 3 yourself for the final support level and check them against the output.

Question 12

Problem Statement

■ Write in lab record

Perform the following:

- Use `contactlenses.arff` and load it into WEKA. Check that all attributes are nominal (categorical).
- Change to the Associate Panel. Select Apriori as associator. After pressing the start button, Apriori starts to build its model and writes its output into the output field. The first part of the output (Run information) describes the options that have been set and the data set used. Make sure you understand all the data reported.
- The rules that have been generated are listed at the end of the output. By default, only the 10 most valuable rules according to their confidence level are shown. Each rule consists of some attribute values on a left hand side of the arrow, the arrow sign and the right hand side list of attribute values. Right of the arrow sign are the predicted attribute values. Rules have certain support and confidence values. The number before the arrow sign is the number of instances the rule applies to. The number after the arrow sign is the number of instances predicted correctly. The number in brackets after `conf:` is the confidence of the rule. Analyse the rules mined from the data set. What are their confidence and support values? Examine the number of large itemsets and make sure you understand how this data has been calculated (check that the values you would get manually are correct).

Solution

■ Write in lab record

Steps

1. Explorer, Open file, `data/contact-lenses.arff`. In the Attributes list click each of the five names: the Type field reads Nominal every time, Missing is 0, and the label tables show age 8, 8, 8; spectacle-prescrip 12, 12; astigmatism 12, 12; tear-prod-rate 12, 12; contact-lenses soft 5, hard 4, none 15.
2. Associate tab, Choose, `weka.associations.Apriori`. Click the name, tick `outputItemSets` so the itemsets are printed as well, leave everything else at its default, OK, Start.

3. Read the Run information, then the Associator model, then the rules, against the hand calculations below. The same output comes from `python3 apriori.py contact-lenses.arff -I` (the program from Session 3).

Output

```
1  ≡≡≡ Run information ≡≡≡
2
3  Scheme:      weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M
4  Relation:    contact-lenses
5  Instances:   24
6  Attributes:  5
7              age
8              spectacle-prescrip
9              astigmatism
10             tear-prod-rate
11             contact-lenses
12 ≡≡≡ Associator model (full training set) ≡≡≡
13
14
15 Apriori
16 =====
17
18 Minimum support: 0.2 (5 instances)
19 Minimum metric <confidence>: 0.9
20 Number of cycles performed: 16
21
22 Generated sets of large itemsets:
23
24 Size of set of large itemsets L(1): 11
25
26 Large Itemsets L(1):
27 age=young 8
28 age=pre-presbyopic 8
29 age=presbyopic 8
30 spectacle-prescrip=myope 12
31 spectacle-prescrip=hypermetrope 12
32 astigmatism=no 12
33 astigmatism=yes 12
34 tear-prod-rate=reduced 12
35 tear-prod-rate=normal 12
36 contact-lenses=soft 5
37 contact-lenses=none 15
38
39 Size of set of large itemsets L(2): 21
40
41 Large Itemsets L(2):
42 age=pre-presbyopic contact-lenses=none 5
43 age=presbyopic contact-lenses=none 6
```

```

44 spectacle-prescrip=myope astigmatism=no 6
45 spectacle-prescrip=myope astigmatism=yes 6
46 spectacle-prescrip=myope tear-prod-rate=reduced 6
47 spectacle-prescrip=myope tear-prod-rate=normal 6
48 spectacle-prescrip=myope contact-lenses=none 7
49 spectacle-prescrip=hypermetrope astigmatism=no 6
50 spectacle-prescrip=hypermetrope astigmatism=yes 6
51 spectacle-prescrip=hypermetrope tear-prod-rate=reduced 6
52 spectacle-prescrip=hypermetrope tear-prod-rate=normal 6
53 spectacle-prescrip=hypermetrope contact-lenses=none 8
54 astigmatism=no tear-prod-rate=reduced 6
55 astigmatism=no tear-prod-rate=normal 6
56 astigmatism=no contact-lenses=soft 5
57 astigmatism=no contact-lenses=none 7
58 astigmatism=yes tear-prod-rate=reduced 6
59 astigmatism=yes tear-prod-rate=normal 6
60 astigmatism=yes contact-lenses=none 8
61 tear-prod-rate=reduced contact-lenses=none 12
62 tear-prod-rate=normal contact-lenses=soft 5
63
64 Size of set of large itemsets L(3): 6
65
66 Large Itemsets L(3):
67 spectacle-prescrip=myope tear-prod-rate=reduced contact-lenses=none 6
68 spectacle-prescrip=hypermetrope astigmatism=yes contact-lenses=none 5
69 spectacle-prescrip=hypermetrope tear-prod-rate=reduced contact-lenses=none 6
70 astigmatism=no tear-prod-rate=reduced contact-lenses=none 6
71 astigmatism=no tear-prod-rate=normal contact-lenses=soft 5
72 astigmatism=yes tear-prod-rate=reduced contact-lenses=none 6
73
74 Best rules found:
75
76 1. tear-prod-rate=reduced 12  $\implies$  contact-lenses=none 12 <conf:(1)> lift:
77 2. spectacle-prescrip=myope tear-prod-rate=reduced 6  $\implies$  contact-lenses=non
78 3. spectacle-prescrip=hypermetrope tear-prod-rate=reduced 6  $\implies$  contact-len
79 4. astigmatism=no tear-prod-rate=reduced 6  $\implies$  contact-lenses=none 6 <co
80 5. astigmatism=yes tear-prod-rate=reduced 6  $\implies$  contact-lenses=none 6 <c
81 6. contact-lenses=soft 5  $\implies$  astigmatism=no 5 <conf:(1)> lift:(2) lev:(0
82 7. contact-lenses=soft 5  $\implies$  tear-prod-rate=normal 5 <conf:(1)> lift:(2)
83 8. tear-prod-rate=normal contact-lenses=soft 5  $\implies$  astigmatism=no 5 <con
84 9. astigmatism=no contact-lenses=soft 5  $\implies$  tear-prod-rate=normal 5 <con
85 10. contact-lenses=soft 5  $\implies$  astigmatism=no tear-prod-rate=normal 5 <con

```

Explanation

Run information. The Scheme line is the option string: `-N 10` numRules, `-T 0` metricType confidence (1 lift, 2 leverage, 3 conviction), `-C 0.9` minMetric, `-D 0.05` delta, `-U 1.0` upperBoundMinSupport, `-M 0.1` lowerBoundMinSupport, `-S -1.0` significance test off, `-c -1` class index last (unused because `car` is false). Relation, Instances 24 and the five attribute names come from the ARFF header.

Minimum support 0.2 (5 instances). WEKA starts at $1.0 - 0.05 = 0.95$ and lowers by 0.05 per cycle; cycle 16 is at $0.95 - 15 \times 0.05 = 0.2$, which is why Number of cycles performed is 16. The instance count is $\lfloor 0.2 \times 24 + 0.5 \rfloor = \lfloor 5.3 \rfloor = 5$. It stopped there because 10 rules passed; at 0.3 (7 instances) only one rule passes and at 0.25 (6 instances) only five (the first five in the list), both fewer than numRules.

L(1) = 11. There are 12 distinct items; every count except `contact-lenses=hard` (4) is at least 5:

Attribute	Items and counts	Large
age	young 8, pre-presbyopic 8, presbyopic 8	3
spectacle-prescrip	myope 12, hypermetrope 12	2
astigmatism	no 12, yes 12	2
tear-prod-rate	reduced 12, normal 12	2
contact-lenses	soft 5, hard 4, none 15	2

L(2) = 21. Pairs come from different attributes, so there are $\binom{11}{2} - 7 = 48$ candidates (55 pairs minus the 7 same-attribute pairs). The data is a full factorial of age, spectacle-prescrip, astigmatism and tear-prod-rate, so counts follow from the design:

Pair type	Count	Large pairs
two of spectacle-prescrip, astigmatism, tear-prod-rate	$\frac{24}{2 \times 2} = 6$ each	all 12
age with any of those three	$\frac{24}{3 \times 2} = 4$	none of 18
age with none	young 4, pre-presbyopic 5, presbyopic 6	2
age with soft	2, 2, 1	none
binary attribute with none	myope 7, hypermetrope 8, no 7, yes 8, reduced 12, normal 3	5
binary attribute with soft	myope 2, hypermetrope 3, no 5, yes 0, reduced 0, normal 5	2

$12 + 2 + 5 + 2 = 21$, the 21 itemsets WEKA lists.

L(3) = 6. A triple is a candidate only if all three of its pairs are in L(2). No triple can contain an age item with a binary attribute (those pairs failed), so age drops out. Triples of three binary attributes have count $\frac{24}{8} = 3$, too small. What is left is two binary items plus a lens value:

Triple	Count	Why
myope, reduced, none	6	every reduced row is none
hypermetrope, reduced, none	6	same
no, reduced, none	6	same
yes, reduced, none	6	same
hypermetrope, yes, none	5	6 rows, only the young normal one is hard
no, normal, soft	5	6 rows, only the presbyopic myope one is none
myope, no, none	4	fails: two of the six rows are soft
myope, yes, none	3	fails: three are hard

No L(4). Every 4-itemset would need all four of its triples large; for example myope, no, reduced, none needs myope, no, none, which has count 4. So the algorithm stops and prints nothing for L(4).

The rules. For each of the ten rules, support is the count after the arrow over 24 and confidence is the count after the arrow over the count before it:

Rule	Before, after	Support	Confidence	Lift
1. reduced $\Rightarrow$ none	12, 12	$\frac{12}{24} = 0.5$	$\frac{12}{12} = 1$	$\frac{1}{15/24} = 1.6$
2. myope, reduced $\Rightarrow$ none	6, 6	$\frac{6}{24} = 0.25$	$\frac{6}{6} = 1$	1.6
3. hypermetrope, reduced $\Rightarrow$ none	6, 6	0.25	1	1.6
4. no, reduced $\Rightarrow$ none	6, 6	0.25	1	1.6
5. yes, reduced $\Rightarrow$ none	6, 6	0.25	1	1.6
6. soft $\Rightarrow$ no	5, 5	$\frac{5}{24} = 0.21$	$\frac{5}{5} = 1$	$\frac{1}{12/24} = 2$
7. soft $\Rightarrow$ normal	5, 5	0.21	1	2
8. normal, soft $\Rightarrow$ no	5, 5	0.21	1	2
9. no, soft $\Rightarrow$ normal	5, 5	0.21	1	2
10. soft $\Rightarrow$ no, normal	5, 5	0.21	1	$\frac{1}{6/24} = 4$

Rules that were generated and rejected at this support: `no, normal  $\Rightarrow$  soft` with confidence $\frac{5}{6} = 0.83$, `hypermetrope, yes  $\Rightarrow$  none` with $\frac{5}{6} = 0.83$, `presbyopic  $\Rightarrow$  none` with $\frac{6}{8} = 0.75$, `pre-presbyopic  $\Rightarrow$  none` with $\frac{5}{8} = 0.625$, and `hypermetrope  $\Rightarrow$  none` with $\frac{8}{12} = 0.67$. All fall below 0.9.

The other metrics on rule 1 check the same way. Leverage is $\frac{12}{24} - \frac{12}{24} \times \frac{15}{24} = 0.5 - 0.3125 = 0.19$, and the bracketed 4 is that leverage times 24, truncated (4.5 becomes 4). Conviction is $\frac{12 \times (24 - 15) / 24}{12 - 12 + 1} = 4.5$; WEKA adds 1 to the denominator so that a rule with confidence 1 does not divide by zero.

What the rules mean. Rule 1 is the clinical fact in the data: reduced tear production rules out contact lenses, in every one of the 12 cases. Rules 2 to 5 are the same fact with an extra irrelevant condition on the left; they are not new information, and their lift is the same 1.6. Rules 6 to 10 describe soft-lens wearers: all five have no astigmatism and normal tear production, and rule 10 says both at once with lift 4, because that combination occurs in only 6 of 24 rows. The reverse rule (no astigmatism and normal tears predict soft) is the one a doctor would want, but it has confidence 0.83 and misses the cut, which shows that Apriori rules run from frequent premise to frequent consequent and are not classification rules; for those, set `car` True and `classIndex` to 5.

Viva Questions

× Do not copy. Read for understanding and the viva

Q: Why does the output say 16 cycles when the support went from 1.0 to 0.2? **A:** The first cycle already runs at 0.95; sixteen steps of 0.05 from 0.95 end at 0.2.

Q: How does WEKA turn support 0.2 into 5 instances on 24 rows? **A:** It rounds 0.2 times 24 = 4.8 to the nearest integer.

Q: Why is `contact-lenses=hard` not in L(1)? **A:** It occurs 4 times, below the 5 needed.

Q: Why are there no age items in L(3)? **A:** Every pair of an age value with a binary attribute value has count 4, so no such pair is in L(2), and Apriori only builds triples from large pairs.

Q: What are the two numbers in `tear-prod-rate=reduced 12 ==> contact-lenses=none 12`? **A:** 12 rows match the premise; 12 of them also match the consequent; confidence is their ratio.

Q: Which rule has the highest lift and why? **A:** Rule 10, lift 4: its consequent (no astigmatism and normal tears) covers 25 percent of rows but 100 percent of soft-lens rows.

Q: Why is the useful rule “no astigmatism and normal tears imply soft” missing? **A:** Its confidence is 5 of 6, below 0.9; one presbyopic myope with those values gets no lenses.

Common Mistakes

× Do not copy. Read for understanding and the viva

- Reading the number after the arrow as the confidence; it is a count, confidence is the ratio in brackets.
- Reporting support as the count before the arrow; support uses the joint count after the arrow divided by 24.
- Counting L(2) candidates as 55 and forgetting to drop pairs from the same attribute, which can never occur together.
- Assuming the 10 rules are all the rules with confidence 0.9; they are the top 10 at the final support, and more appear if `numRules` is raised.
- Treating rules 2 to 5 as new findings when they only restate rule 1 with an extra condition.

Formula Sheet

✗ Do not copy. Read for understanding and the viva

Association rules

For a rule $X \Rightarrow Y$ over N transactions:

$$\text{support}(X \Rightarrow Y) = \frac{|X \cup Y|}{N}, \quad \text{confidence}(X \Rightarrow Y) = \frac{|X \cup Y|}{|X|}, \quad \text{lift}(X \Rightarrow Y) = \frac{\text{confidence}(X \Rightarrow Y)}{\text{support}(Y)}$$

WEKA's Apriori starts at the upper bound of minimum support and lowers it by the delta each pass until the requested number of rules is found or the lower bound is reached.

Entropy, information gain and Gini

For a set S with class proportions $p_1, \dots, p_c$:

$$H(S) = - \sum_{i=1}^c p_i \log_2 p_i, \quad \text{Gain}(S, A) = H(S) - \sum_{v \in \text{values}(A)} \frac{|S_v|}{|S|} H(S_v), \quad \text{Gini}(S) = 1 - \sum_{i=1}^c p_i^2$$

ID3 splits on the attribute with the highest gain; J48 (C4.5) uses the gain ratio

$\text{Gain}(S, A) / \text{SplitInfo}(S, A)$ where $\text{SplitInfo}(S, A) = - \sum_v \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|}$.

Classifier evaluation

From the confusion matrix with true positives TP , false positives FP , false negatives FN , true negatives TN :

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad \text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

Kappa compares observed agreement p_o (accuracy) with the agreement expected by chance p_e :

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \quad p_e = \sum_i \frac{(\text{row}_i \text{ total}) (\text{column}_i \text{ total})}{N^2}$$

The ROC curve plots true positive rate $TP / (TP + FN)$ against false positive rate $FP / (FP + TN)$; the area under it (AUC) is 0.5 for guessing and 1.0 for a perfect classifier.

Naive Bayes and k-nearest neighbour

$$P(C | x_1, \dots, x_n) \propto P(C) \prod_{i=1}^n P(x_i | C)$$

k-NN assigns the majority class among the k nearest training records under Euclidean distance

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{i=1}^n (a_i - b_i)^2}$$

after normalising each attribute to $[0, 1]$ with $x' = (x - x_{\min}) / (x_{\max} - x_{\min})$.

Linear regression

$$\hat{y} = \beta_0 + \beta_1 x, \quad \beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}, \quad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

WEKA reports the correlation coefficient, mean absolute error and root mean squared error $\sqrt{\frac{1}{N} \sum (y_i - \hat{y}_i)^2}$.

Clustering

k-means minimises the within-cluster sum of squared errors over clusters $C_1, \dots, C_k$ with centroids μ_j :

$$SSE = \sum_{j=1}^k \sum_{\mathbf{x} \in C_j} \|\mathbf{x} - \mu_j\|^2, \quad \mu_j = \frac{1}{|C_j|} \sum_{\mathbf{x} \in C_j} \mathbf{x}$$

Hierarchical (agglomerative) clustering merges the two closest clusters each step; linkage defines closeness: single $\min d(a, b)$, complete $\max d(a, b)$, average $\frac{1}{|A||B|} \sum d(a, b)$.

DBSCAN calls a point a core point when at least minPts points lie within radius ϵ ; clusters grow from core points, and points reachable from none are noise.

Session Summary

Write in lab record

- Question 12: contact-lenses.arff loaded and confirmed all nominal; Apriori run with defaults and `outputItemSets`; every line of the Run information decoded; the support of 0.2 and 16 cycles, $L(1) = 11$, $L(2) = 21$, $L(3) = 6$, and the support, confidence and lift of all ten rules recomputed by hand and matched to the output

[✎ Edit this page](#)[< Previous](#)
Session 4**Session 6** [Next >](#)