

1. Consider the traffic accident data set shown in Table 7.10 (p. 473)

a) Show a binarized version of the data set.

Attribute names have been shortened due to the size of the table.

Weather = Bad	Weather = Good	Driver = Impaired	Driver = Sober	Violation = None	Violation = Speed Limit	Violation = Stop Sign	Violation = Traffic Sig	Seat Belt = No	Seat Belt = Yes	Crash Sev. = Major	Crash Sev. = Minor
0	1	1	0	0	1	0	0	1	0	1	0
1	0	0	1	1	0	0	0	0	1	0	1
0	1	0	1	0	0	1	0	0	1	0	1
0	1	0	1	0	1	0	0	0	1	1	0
1	0	0	1	0	0	0	1	1	0	1	0
0	1	1	0	0	0	1	0	0	1	0	1
1	0	1	0	1	0	0	0	0	1	1	0
0	1	0	1	0	0	0	1	0	1	1	0
0	1	1	0	1	0	0	0	1	0	1	0
1	0	0	1	0	0	0	1	1	0	1	0
0	1	1	0	0	1	0	0	0	1	1	0
1	0	0	1	0	0	1	0	0	1	0	1

b) What is the maximum width of each transaction in the binarized data?

According to the book, p. 418, “[T]he width of every transaction is the same as the number of attributes in the original data.” Therefore, since there are 5 attribute columns in the original data, the maximum width is 5.

c) Assuming that support threshold is 30%, how many candidate and frequent itemsets are generated?

1-item sets over threshold (at least 4/12):

Item Set	Support
Weather = Good	7/12 = 58%
Weather = Bad	5/12 = 42%
Driver = Impaired	5/12 = 42%
Driver = Sober	7/12 = 58%
Seat Belt = No	4/12 = 33%
Seat Belt = Yes	8/12 = 67%

Crash Severity = Major	8/12 = 67%
Crash Severity = Minor	4/12 = 33%

2-item sets over threshold (at least 4/12 - frequent):

Item Set	Support
Weather = Bad, Driver = Sober	4/12 = 33%
Weather = Good, Driver = Impaired	4/12 = 33%
Weather = Good, Seat Belt = Yes	5/12 = 42%
Weather = Good, Crash Severity = Major	5/12 = 42%
Driver = Impaired, Crash Severity = Major	4/12 = 33%
Driver = Sober, Crash Severity = Major	4/12 = 33%
Driver = Sober, Seat Belt = Yes	5/12 = 42%
Seat Belt = No, Crash Severity = Major	4/12 = 33%
Seat Belt = Yes, Crash Severity = Major	4/12 = 33%
Seat Belt = Yes, Crash Severity = Minor	4/12 = 33%

There are 8 total attributes, and we need to choose 2, so there are 28 candidate item sets, but there are only 10 of them which are frequent.

No 3-item sets were over the 30% threshold, though there were 220 candidate item sets using the brute force method.

d) Create a data set that contains only the following asymmetric binary attributes: (Weather = Bad, Driver's condition = Alcohol-impaired, Traffic violation = Yes, Seat Belt = No, Crash Severity = Major). For Traffic violation, only None has a value of 0. The rest of the attribute values are assigned to 1. Assuming that support threshold is 30%, how many candidate and frequent itemsets will be generated?

Weather = Bad	Driver = Impaired	Violation = Yes	Seat Belt = No	Crash Sev. = Major
0	1	1	1	1
1	0	0	0	0
0	0	1	0	0
0	0	1	0	1
1	0	1	1	1
0	1	1	0	0
1	1	0	0	1
0	0	1	0	1
0	1	0	1	1
1	0	1	1	1
0	1	1	0	1
1	0	1	0	0

Frequent 1-item sets (of 5 candidates):

Item Set	Support
Weather = Bad	5/12 = 42%
Driver = Impaired	5/12 = 42%
Violation = Yes	9/12 = 75%
Seat Belt = No	4/12 = 33%
Crash Severity = Major	8/12 = 67%

Frequent 2-item sets (of 10 candidates):

Item Set	Support
Driver = Impaired, Crash Severity = Major	4/12 = 33%
Violation = Yes, Crash Severity = Major	6/12 = 50%
Seat Belt = No, Crash Severity = Major	4/12 = 33%

No 3-item sets were over the 30% threshold, though there were 10 total candidates using the brute force method.

e) Compare the number of candidate and frequent itemsets generated in parts (c) and (d).

It's pretty obvious that a data set with less attributes yielded more frequent itemsets, and this logically makes sense. Condensing the Violation attribute definitely helped its status as well – breaking the attribute into various types of violation doesn't appear to make much sense in this scenario, as the possible values are too diverse, and aren't as relevant as other attributes being presented.

2.

a) Consider the data set shown in Table 7.11 (p. 474). Suppose we apply the following discretization strategies to the continuous attributes of the data set.

D1: Partition the range of each continuous attribute into 3 equal-sized bins.

D2: Partition the range of each continuous attribute into 3 equal-sized bins, where each bin contains an equal number of transactions.

For each strategy, answer the following questions:

- i) Construct a binarized version of the data set**
- ii) Derive all the frequent itemsets having support $k \geq 30\%$**

Strategy D1:

TID	Temp = 87	Temp = 88-96	Temp = 97	Pres. = 1052	Pres. = 1053-1079	Pres. = 1080	Alarm 1	Alarm 2	Alarm 3
1	0	1	0	0	0	1	0	0	1
2	1	0	0	1	0	0	1	1	0
3	0	0	1	0	0	1	1	1	1
4	0	0	1	0	0	1	1	0	0
5	1	0	0	1	0	0	0	1	1
6	0	0	1	0	0	1	1	1	0
7	1	0	0	1	0	0	1	0	1
8	1	0	0	1	0	0	1	0	0
9	0	0	1	0	0	1	1	1	1

Frequent 1-item sets (9 possible):

Item Set	Support
Temp = 87	4/9 = 44%
Temp = 97	4/9 = 44%
Pres. = p1052	4/9 = 44%
Pres. = p1080	5/9 = 56%
Alarm 1	7/9 = 78%
Alarm 2	5/9 = 56%
Alarm 3	5/9 = 56%

Frequent 2-item sets (36 total possible, 10 possible taking into account exclusivity in Temp and Pressure and infrequent item sets):

Item Set	Support
Temp = 87, Pres. = 1052	4/9 = 44%
Temp = 87, Alarm 1	3/9 = 33%
Temp = 97, Pres. = 1080	4/9 = 44%
Temp = 97, Alarm 1	4/9 = 44%
Pres. = 1080, Alarm 1	4/9 = 44%
Pres. = 1080, Alarm 2	3/9 = 33%
Pres. = 1080, Alarm 3	3/9 = 33%
Alarm 1, Alarm 2	4/9 = 44%
Alarm 1, Alarm 3	3/9 = 33%
Alarm 2, Alarm 3	3/9 = 33%

Frequent 3-item sets (84 total possible with brute-force):

Item Set	Support
Temp = 87, Pres. = 1052, Alarm 1	3/9 = 33%
Temp = 97, Pres. = 1080, Alarm 1	4/9 = 44%
Temp = 97, Pres. = 1080, Alarm 1	3/9 = 33%
Temp = 97, Pres. = 1080, Alarm 2	3/9 = 33%
Pres. = 1080, Alarm 1, Alarm 2	3/9 = 33%

Frequent 4-item sets (126 possible with brute-force):

Item Set	Support
Temp = 97, Pres. = 1080, Alarm 1, Alarm 2	3/9 = 33%

Strategy D2:

TID	Temp = 85	Temp = 86-97	Temp = 98	Pres. = 1038	Pres. = 1039-1084	Pres. = 1085	Alarm 1	Alarm 2	Alarm 3
1	0	1	0	0	0	1	0	0	1
2	1	0	0	0	1	0	1	1	0
3	0	0	1	0	0	1	1	1	1
4	0	1	0	0	1	0	1	0	0
5	1	0	0	1	0	0	0	1	1
6	0	0	1	0	1	0	1	1	0
7	1	0	0	1	0	0	1	0	1
8	0	1	0	1	0	0	1	0	0
9	0	0	1	0	0	1	1	1	1

Frequent 1-item sets:

Item Set	Support
Each temperature attribute	3/9 = 33%
Each pressure attribute	3/9 = 33%
Alarm 1	7/9 = 78%
Alarm 2	5/9 = 56%
Alarm 3	5/9 = 56%

Frequent 2-item sets:

Item Set	Support
Temp = 98, Alarm 1	3/9 = 33%
Temp = 98, Alarm 2	3/9 = 33%
Pres. 1039-1084, Alarm 1	3/9 = 33%

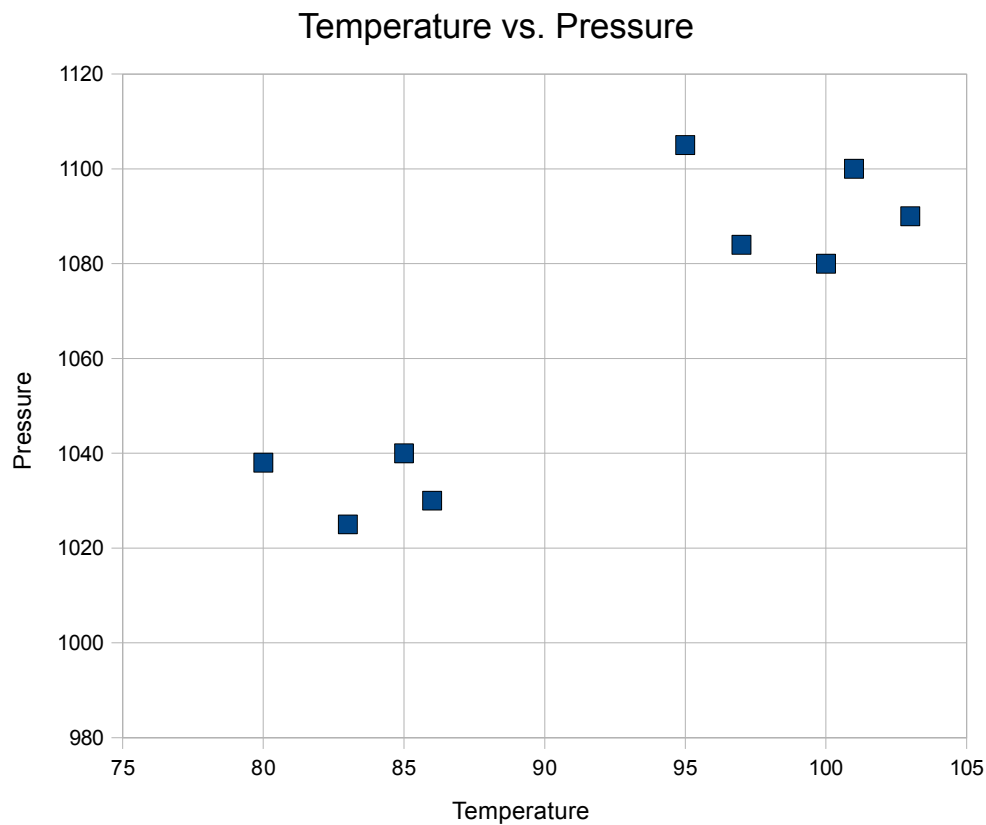
Pres. 1085, Alarm 3	$3/9 = 33\%$
Alarm 1, Alarm 2	$4/9 = 44\%$
Alarm 1, Alarm 3	$3/9 = 33\%$
Alarm 2, Alarm 3	$3/9 = 33\%$

Frequent 3-item sets:

Item Set	Support
Temp = 98, Alarm 1, Alarm 2	$3/9 = 33\%$

b) The continuous attribute can also be discretized using a clustering approach.

i) Plot the graph of temperature vs. pressure for the data points shown in Table 7.11



ii) How many natural clusters do you observe from the graph? Assign a label (C1, C2, etc.) to each cluster in the graph.

There are two obvious clusters on the graph – the one in the mid-lower left (C1), and the one in the top right (C2).

iii) What type of clustering algorithm do you think can be used to identify the clusters?

State your reasons clearly.

Clustering algorithms are the topic of the next chapter, which I haven't gotten to read in-depth yet. However, scanning ahead to p. 495, three techniques are listed: K-means, Agglomerative Hierarchical Clustering, and DBSCAN. Reviewing these three techniques leads me to believe I used a form of DBSCAN when I chose my two clusters; I didn't have a set number of clusters I wished to find, ruling out K-means. AHC seems more complex, leading me to believe I used a form of DBSCAN in my head to find the regions of most density. DBSCAN also specifies dropping points which don't fit into obvious clusters as noise, which I would have done.

iv) Replace the temperature and pressure attributes in Table 7.11 with asymmetric binary attributes C1, C2, etc. Construct a transaction matrix using the new attributes (along with the attributes Alarm 1, Alarm 2, and Alarm 3).

TID	C1	C2	Alarm 1	Alarm 2	Alarm 3
1	0	1	0	0	1
2	1	0	1	1	0
3	0	1	1	1	1
4	0	1	1	0	0
5	1	0	0	1	1
6	0	1	1	1	0
7	1	0	1	0	1
8	1	0	1	0	0
9	0	1	1	1	1

v) Derive all the frequent itemsets having support 30% from the binarized data.

Frequent 1-item sets:

Item Set	Support
C1	4/9 = 44%
C2	5/9 = 56%
Alarm 1	7/9 = 78%
Alarm 2	5/9 = 56%
Alarm 3	5/9 = 56%

Frequent 2-item sets:

Item Set	Support
C1, Alarm 1	3/9 = 33%
C2, Alarm 1	4/9 = 44%
C2, Alarm 2	3/9 = 33%
C2, Alarm 3	3/9 = 33%

Frequent 3-item sets:

Item Set	Support
C2, Alarm 1, Alarm 2	3/9 - 33%