

# Data Mining Assignment 2

September 24, 2025

Viet Huy Duong  
811433146

---

1. Implementing Apriori algorithm (using C++, Java, or any programming language). Your program should be able to accept two parameters with input: filename and a minimal support level. For instance, "myapriori filename 15", where "myapriori" is the execution file, and 15 means a frequent itemset has frequency of 15% of the entire transactions in "filename". The file format is as follows: each line corresponds to a transaction (no transaction id) and each item in the transaction is separated by a space. Your program should output all the frequent itemsets in the input file with the specified minimal support level. (10 pts)
  - (a) A detailed Pseudo code including the necessary data structure for implementation.
  - (b) Source code.
  - (c) The running results (screen captures) of the following input file and minimal support (10%, 20%, 30%, 50%).

(a) A detailed Pseudo code including the necessary data structure for implementation.

---

**Algorithm 1** Support Function

---

```

1: function FINDFREQUENTONE(transactions, min_support_count)
2:   item_counts  $\leftarrow$  DefaultDict(int)
3:   for each transaction in transactions do
4:     for each item in transaction do
5:       item_counts[item]  $\leftarrow$  item_counts[item] + 1
6:     end for
7:   end for
8:   return {frozenset([item]) : count | count  $\geq$  min_support_count}
9: end function
10: function GENERATECANDIDATES(prev_frequent, k)
11:   candidates  $\leftarrow$   $\emptyset$ 
12:   for each pair (set1, set2) in prev_frequent do
13:     union  $\leftarrow$  set1  $\cup$  set2
14:     if len(union) = k then
15:       candidates  $\leftarrow$  candidates  $\cup$  {union}
16:     end if
17:   end for
18:   return candidates
19: end function
20: function COUNTSUPPORTS(candidates, transactions)
21:   supports  $\leftarrow$  DefaultDict(int)
22:   for each transaction in transactions do
23:     for each candidate in candidates do
24:       if candidate  $\subseteq$  transaction then
25:         supports[candidate]  $\leftarrow$  supports[candidate] + 1
26:       end if
27:     end for
28:   end for
29:   return supports
30: end function

```

---

**Algorithm 2** Apriori

---

```

1: procedure APRIORI(transactions, min_support_perc)
2:   total_transactions  $\leftarrow$  len(transactions)
3:   min_support_count  $\leftarrow$  min_support_perc/100  $\times$  total_transactions
4:   frequent_itemsets  $\leftarrow$  FindFrequentOne(transactions,
      min_support_count)
5:   all_frequent  $\leftarrow$  frequent_itemsets
6:   k  $\leftarrow$  2
7:   while frequent_itemsets is not empty do
8:     candidates  $\leftarrow$  GenerateCandidates(frequent_itemsets, k)
9:     if candidates is empty then
10:      break
11:    end if
12:    supports  $\leftarrow$  CountSupports(candidates, transactions)
13:    frequent_itemsets  $\leftarrow$  {itemset:count|count  $\geq$  min_support_count}
14:    all_frequent  $\leftarrow$  all_frequent  $\cup$  frequent_itemsets
15:    k  $\leftarrow$  k + 1
16:  end while
17:  return all_frequent
18: end procedure

```

---

The main idea to solve above problem is base on dynamic programming, first of all, we count and get the set of one item have frequent greater than the threshold, then, the core idea is definition the candidate of k length items set by combine the k-1 length items set, that is because if a set of k elements satisfies the condition, then it proves that every set of k-1 elements in it must satisfy the condition.

For example:

We have 2 set  $\{butter, cake, icecream\}$  and  $\{icecream, meat, cake\}$ , the combination of them is  $\{butter, cake, icecream, meat\}$  and that combination is not insure the set 4 will satisfies the condition, that just be a candidate for the next support count check progress.

After that, we going to recheck the fequent of candidate set, if satisfies, candidate set will be save into the frequent itemsets, which will be used for the next *GenerateCandidate* k value

## (b) Source code.

```
import sys
from collections import defaultdict
import math

def find_frequent_1_itemsets(transactions, min_support_count):
    item_counts = defaultdict(int)
    for transaction in transactions:
        for item in transaction:
            item_counts[item] += 1
    frequent_items = {frozenset([item]):count for item,count in
        ↪ item_counts.items() if count >= min_support_count}
    return frequent_items

def generate_candidates(prev_frequent, k):
    candidates = set()
    prev_items = list(prev_frequent.keys())
    for i in range(len(prev_items)):
        for j in range(i+1, len(prev_items)):
            set1 = prev_items[i]
            set2 = prev_items[j]
            union = set1 | set2
            if len(union) == k:
                candidates.add(union)
    return candidates

def count_supports(candidates, transactions):
    supports = defaultdict(int)
    for transaction in transactions:
        for candidate in candidates:
            if candidate.issubset(transaction):
                supports[candidate] += 1
    return supports

def main():
    filename = sys.argv[1]
    try:
        min_support_perc = float(sys.argv[2])
    except ValueError:
```

```
print("lack of minimum frequent percentage")
sys.exit(1)
try:
    with open(filename, 'r') as f:
        transactions = [set(line.strip().split()) for line in
            ↪ f if line.strip()]
except FileNotFoundError:
    print(f"File {filename} not available")
    sys.exit(1)

if not transactions:
    print("No transactions")
    return

total_transactions = len(transactions)
min_support_count = math.ceil((min_support_perc / 100) *
    ↪ total_transactions)

frequent_itemsets = find_frequent_1_itemsets(transactions,
    ↪ min_support_count)
all_frequent = frequent_itemsets.copy()

k = 2
while frequent_itemsets:
    candidates = generate_candidates(frequent_itemsets, k)
    if not candidates:
        break
    supports = count_supports(candidates, transactions)
    frequent_itemsets = {itemset: count for itemset, count in
        ↪ supports.items() if count >= min_support_count}
    all_frequent.update(frequent_itemsets)
    k += 1

if len(sys.argv) > 3 and sys.argv[3] == "1":
    for itemset in sorted(all_frequent.keys(), key=lambda x:
        ↪ (len(x), sorted(x))):
        print(' '.join(sorted(itemset)) + f" (support:
            ↪ {all_frequent[itemset]} with frequent:
            ↪ {all_frequent[itemset]*100/total_transactions}%)"
            ↪ )
```

```
    else:
        for itemset in sorted(all_frequent.keys(), key=lambda x:
                               ↪ (len(x), sorted(x))):
            print(' '.join(sorted(itemset)))

if __name__ == "__main__":
    main()
```

(c) The running results (screen captures) of the following input file and minimal support (10%, 20%, 30%, 50%)

```

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040

```

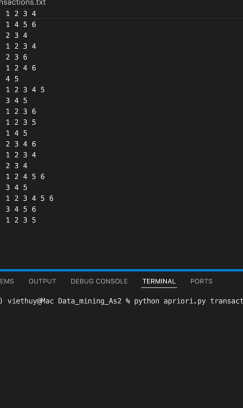
10%

```

E transactions.txt M X
E transactions.txt
1 1 2 3 4
2 1 4 5 6
3 2 3 4
4 1 2 3 4
5 2 3 6
6 1 2 4 6
7 4 5
8 1 2 3 4 5
9 3 4 5
10 1 2 3 4
11 1 2 3 5

```

20%



```

$ transactions.txt M X
$ transactions.txt
1 1
2 2 3 4
3 1 4 5 6
4 2 3 4
5 1 2 3 4
6 2 3 4 5
7 1 4 5
8 1 2 3 4 5
9 3 4 5
10 1 2 3 4
11 1 2 3 5
12 1 4 5
13 2 3 4 5
14 1 2 3 4
15 2 3 4
16 1 2 4 5 6
17 2 4 5
18 1 2 3 4 5 6
19 3 4 5 6
20 1 2 3 5
21

```

30%

```
F transactions.txt M X
F transactions.txt
1 1 2 3 4
2 1 4 5 5
3 2 3 4
4 1 2 3 4
5 2 3 4 6
6 1 2 4 6
7 4 5
8 1 2 3 4 5
9 3 4 5
10 1 2 3 6
11 1 2 3 5
12 1 4 5
13 2 3 4 6
14 1 2 3 4
15 2 3 4
16 1 2 4 5 6
17 3 4 5
18 1 2 3 4 5 6
19 3 4 5 6
20 1 2 3 5
21
```

50%

*Note: I am sorry if the output is too small, i considered to trade off between the clearly output and the evident that i am actually running on my program. It is a pdf embedded, so please zoom in to see clearly. I am very appreciate*

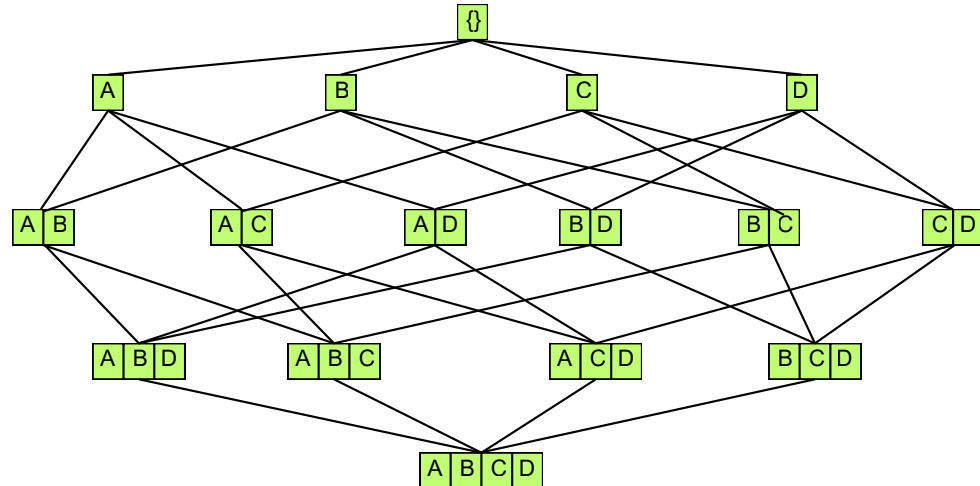
*for that.*

2. Given 4 items, A, B, C, and D, list all possible itemsets in a lattice. (10 pts)

As i understand, lattice is the hierarchical representation of subsets of a set included the empty set. For set  $\{A, B, C, D\}$  with 4 items, the lattice includes all possible subsets, which total  $2^4 = 16$ , include:

- Level 0 (size 0):  $\{\emptyset\}$  (empty set)
- Level 1 (size 1):  $\{A\}, \{B\}, \{C\}, \{D\}$
- Level 2 (size 2):  $\{A, B\}, \{A, C\}, \{A, D\}, \{B, C\}, \{B, D\}, \{C, D\}$
- Level 3 (size 3):  $\{A, B, C\}, \{A, B, D\}, \{A, C, D\}, \{B, C, D\}$
- Level 4 (size 4):  $\{A, B, C, D\}$

And the chart bellow show the hierarchical of them:



3. In the following transaction database, if the minimum support is 7, please list all frequent itemsets and their support counts. (10 pts)



This is the output of Apriori with above transaction and minimum support is 7:

```
(base) viethuy@MacBook-Ben Data_mining_As2 % python apriori.py Next_transactions.txt 70 1
A (support: 8 with frequent: 80.0%)
B (support: 7 with frequent: 70.0%)
```

Note: Because there are 10 transaction so i will use 70 stand for 70% ,parameter 1 at the end of terminal line is for show support count and frequent: 1 shows, None or other value is ignore the output)

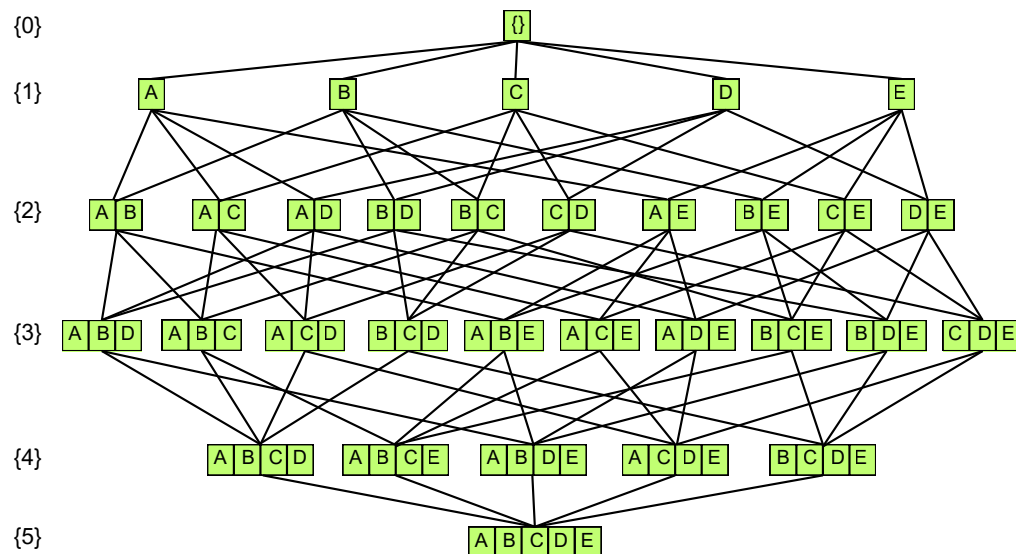
4. In Question 3 above, please answer the following questions:

(10 pts)

- draw the lattice of all itemsets (with their support counts).
- Given minimal support 5, list closed and maximal frequent itemset(s).

**(a) draw the lattice of all itemsets (with their support counts).**

Because the image is too large, so i just represent the lattice with index level, and the support counts will be describe bellow:



The support counts of each itemset:

- Level 0 (size 0):  $\{\emptyset : 10\}$

- Level 1 (size 1):  $\{A : 8\}, \{B : 7\}, \{C : 6\}, \{D : 1\}, \{E : 4\}$
- Level 2 (size 2):  $\{A, B : 5\}, \{A, C : 5\}, \{A, E : 4\}, \{B, C : 3\}, \{B, D : 1\}, \{B, E : 3\}, \{C, E : 2\}, \{A, D : 0\}, \{C, D : 0\}, \{D, E : 0\}$
- Level 3 (size 3):  $\{A, B, C : 2\}, \{A, B, E : 3\}, \{A, C, E : 2\}, \{B, C, E : 1\}, \{A, B, D : 0\}, \{A, C, D : 0\}, \{A, D, E : 0\}, \{B, C, D : 0\}, \{B, D, E : 0\}, \{C, D, E : 0\}$
- Level 4 (size 4):  $\{A, B, C, E : 1\}, \{A, B, C, D : 0\}, \{A, B, D, E : 0\}, \{A, C, D, E : 0\}, \{B, C, D, E : 0\}$
- Level 5 (size 5):  $\{A, B, C, D, E : 0\}$

To have that information, i just run the apriori algorithm with minimum support is 0, and the output will show all the itemsets with their support counts.

```
(base) viethuy@MacBook-Ben Data_mining_As2 % python apriori.py Next_transactions.txt 0 1
A (support: 8 with frequent: 80.0%)
B (support: 7 with frequent: 70.0%)
C (support: 6 with frequent: 60.0%)
D (support: 1 with frequent: 10.0%)
E (support: 4 with frequent: 40.0%)
A B (support: 5 with frequent: 50.0%)
A C (support: 5 with frequent: 50.0%)
A E (support: 4 with frequent: 40.0%)
B C (support: 3 with frequent: 30.0%)
B D (support: 1 with frequent: 10.0%)
B E (support: 3 with frequent: 30.0%)
C E (support: 2 with frequent: 20.0%)
A B C (support: 2 with frequent: 20.0%)
A B E (support: 3 with frequent: 30.0%)
A C E (support: 2 with frequent: 20.0%)
B C E (support: 1 with frequent: 10.0%)
A B C E (support: 1 with frequent: 10.0%)
```

**(b) Given minimal support 5, list closed and maximal frequent itemset(s).**

With minimum support is 5, the frequent itemsets are:

```
(base) viethuy@MacBook-Ben Data_mining_As2 % python apriori.py Next_transactions.txt 50 1
A (support: 8 with frequent: 80.0%)
B (support: 7 with frequent: 70.0%)
C (support: 6 with frequent: 60.0%)
A B (support: 5 with frequent: 50.0%)
A C (support: 5 with frequent: 50.0%)
```

And the closed itemsets:

- $\{A : 8\}$
- $\{B : 7\}$
- $\{C : 6\}$
- $\{A, B : 5\}$
- $\{A, C : 5\}$

And the maximal itemsets:

- $\{A, B : 5\}$
- $\{A, C : 5\}$