Data Mining- Lab Assignment Report

Submitted By- Aaditya Raj Barnwal

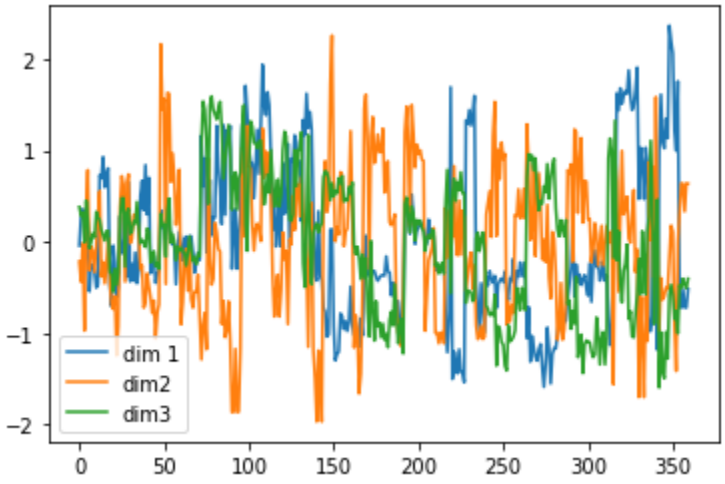
1.PCA:

Step 1: Standardization

Step 2: Covariance Matrix computation

Step 3: Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components

The Eigen values and eigen vectors are the output(Refer to the code).
And after reduction by some value the reduced dimension will be also shown to the output of the code.



This is the plot we get after reducing the dimension.
Here dim1,dim2,dim3 are the different dimensions
of reduced dimension of data,corresponding to 3 maximum eigen values.

2nd Question:

Description of the DataSet:

t20i6d100k Total no. of Transactions : 99922 Total no. of items : 893 Average width of the transaction : 19.899791837633355
Chess Total no. of Transactions : 3196 Total no. of items : 75 Average width of the transaction : 37.0
liquor Total no. of Transactions : 52131 Total no. of items : 4026 Average width of the transaction : 7.876676066064338

1. **Apriori Algorithm**
   uses prior knowledge of frequent itemset properties. We apply an iterative approach or level-wise search where k-frequent itemsets are used to find k+1 itemsets.

   Properties:
   All subsets of a frequent itemset must be frequent(Apriori property).
   If an itemset is infrequent, all its supersets will be infrequent.

Step-1: K=1 (I) generate table of support count of every item present in dataset C1 and compare candidate set items support count with minimum support(L1)

Step-2: K=2

Now Generate candidates set c2 using this L1 such that L(k-1) such a way that it have (k-2) elements in common and check all the subsets of frequent itemset and infrequent itemset and remove that, Now again prune it using minimum support(L2)

Repeat The above step until we are getting the frequent itemsets

(T20i6D100k Dataset)

Minimum support=0.08 Total Time Taken 1028.7339351177216 Memory Usage: 131.5 MiB Total no. of transaction 99922 Total No. of itemsets 893 Average length of transaction 19.899791837633355 total no. Last Level Frequent Itemset 23 The size of maximal freq itemset 1 The No. of maximum frequent itemsets 23	**Minimum support=0.09** Total time taken : 321.23519372940063 Memory Usage: 137.1 MiB Total no. of transaction 99922 Total No. of itemsets 893 Average length of transaction 19.899791837633355 total no. Last Level Frequent Itemset : 13 The size of maximal freq itemset : 1 The No. of maximum frequent itemsets : 13
Minimum support=0.07 Total time taken : 1128..88042497634888 s Memory Usage: 125.1MiB Total no. of transaction 99922 Total No. of itemsets 893 Average length of transaction 19.899791837633355 total no. Last Level Frequent Itemset : 34 The size of maximal freq itemset : 1 The No. of maximum frequent itemsets : 34	**Minimum support=0.6** Total time taken : 1308..88042497634888 s Memory Usage: 115.1MiB Total no. of transaction 99922 Total No. of itemsets 893 Average length of transaction 19.899791837633355 total no. Last Level Frequent Itemset : 61 The size of maximal freq itemset : 1 The No. of maximum frequent itemsets : 61

Chess Dataset

Minimum support=0.92 Total time taken : 805..88042497634888 s Memory Usage: 125.1MiB Total no. of transaction 3196 Total No. of itemsets 75 Average length of transaction 37.0 total no. Last Level Frequent Itemset : 305 The size of maximal freq itemset : 6 The No. of maximum frequent itemsets : 5 **Minimum support = 0.88** Total time taken : 1027..0142497634888 s Memory Usage: 132.1MiB Total no. of transaction 3196 Total No. of itemsets 75 Average length of transaction 37.0 total no. Last Level Frequent Itemset : 1195 The size of maximal freq itemset : 7 The No. of maximum frequent itemsets : 20	**Minimum support = 0.9** Total time taken : 985..88042497634888 s Memory Usage: 130.1MiB Total no. of transaction 3196 Total No. of itemsets 75 Average length of transaction 37.0 total no. Last Level Frequent Itemset : 622 The size of maximal freq itemset : 7 The No. of maximum frequent itemsets : 4 **Minimum support = 0.86** Total time taken : 1027..0142497634888 s Memory Usage: 132.1MiB Total no. of transaction 3196 Total No. of itemsets 75 Average length of transaction 37.0 total no. Last Level Frequent Itemset : 1195 The size of maximal freq itemset : 7 The No. of maximum frequent itemsets : 20

Liquor Dataset

Minimum support=0.8 Total time taken : 1.2671868801116943 Memory Usage: 119.7 MiB Total no. of transaction 52131 Total No. of itemsets 4026 Average length of transaction 7.876676066064338 total no. Last Level Frequent Itemset : 0 The size of maximal freq itemset : 1 The No. of maximum frequent itemsets : 0	**Minimum support=0.05** Total time taken : 1235..78042497685698 s Memory Usage: 128.1MiB Total no. of transaction 52131 Total No. of itemsets 4026 Average length of transaction 7.876676066064338 total no. Last Level Frequent Itemset : 36 The size of maximal freq itemset : 2 The No. of maximum frequent itemsets : 8
Minimum support=0.06	**Minimum support=0.09**

Total time taken : 1152..8820497634888 s Memory Usage: 115.1MiB Total no. of transaction 52131 Total No. of itemsets 4026 Average length of transaction 7.876676066064338 total no. Last Level Frequent Itemset : 23 The size of maximal freq itemset : 2 The No. of maximum frequent itemsets : 4	Total time taken : 805..88042497634888 s Memory Usage: 110.1MiB Total no. of transaction 52131 Total No. of itemsets 4026 Average length of transaction 7.876676066064338 total no. Last Level Frequent Itemset : 12 The size of maximal freq itemset : 2 The No. of maximum frequent itemsets : 3

2. FP-growth(Frequent Pattern Growth Algorithm)

A frequent pattern is generated without the need for candidate generation. FP growth algorithm represents the database in the form of a tree called a frequent pattern tree or FP tree.

Frequent Pattern Algorithm Steps

1.Scan the dataset to find the occurrence of each itemset and now sort each transaction based on this frequency of itemset(dec order)

2.Now construct the FP tree, The root is NULL

3.Scan the file again and examine the transaction. Exame the first transaction and find out the itemset in it and so on ,on all the transactions

4.Now construct the FP tree based on this transaction
Also count of each itemset in the FP tree is increased when we visit that node again and again.

5.Now Mine the fp tree. For this the lowest node is examined first along with the links of the lowest nodes. For this, the lowest node is examined first along with the links of the lowest

nodes. Conditional pattern base is a sub-database consisting of prefix paths in the FP tree occurring with the lowest node (suffix).

6.Now Construct a Conditional FP Tree, which is formed by a count of itemsets in the path. The itemsets meeting the threshold support are considered in the Conditional FP Tree.

Frequent Patterns are generated from the Conditional FP Tree.

(T20i6D100k Dataset)

Minimum support=0.09 The time taken 19.83425784111023 The transaction are 99922 Space - 948.0 MiB Total no. of items 893 Average length of transaction is 19.899791837633355 The no. of frequent itemset 21 Size of maximal frequent itemset : 2	**Minimum support=0.1** The time taken 22.991573095321655 The transaction are 99922 Total no. of items 893 Space - 850.6MiB Average length of transaction is 19.899791837633355 The no. of frequent itemset 13 Size of maximal frequent itemset : 2
Minimum support=0.08 The time taken 28.280821800231934 The transaction are 99922 Space - 989.6MiB Total no. of items 893 Average length of transaction is 19.899791837633355 The no. of frequent itemset 23 Size of maximal frequent itemset : 1	**Minimum support=0.085** The time taken 28.265601873397827 The transaction are 99922 Total no. of items 893 Space - 968.MiB Average length of transaction is 19.899791837633355 The no. of frequent itemset 16 Size of maximal frequent itemset : 1

Chess Dataset

Minimum support=0.7 The time taken 859.83425784111023 The transaction are 3196 Total no. of items 75	**Minimum support=0.75** The time taken 609.83425784111023 The transaction are 3196 Total no. of items 75

Average length of transaction is 37.0 The no. of frequent itemset 48731 Size of maximal frequent itemset : 13	Average length of transaction is 37.0 The no. of frequent itemset 20993 Size of maximal frequent itemset : 11
Minimum support=0.9 The time taken 119.83425784111023 The transaction are 3196 Total no. of items 75 Average length of transaction is 37.0 The no. of frequent itemset 622 Size of maximal frequent itemset : 7	**Minimum support=0.85** The time taken 199.03428554111023 The transaction are 3196 Total no. of items 75 Average length of transaction is 37.0 The no. of frequent itemset 2669 Size of maximal frequent itemset : 8

Liquor Dataset

Minimum support=0.03 The time taken 109.85965784111023 The transaction are 52131 Total no. of items 4026 Average length of transaction is 7.876676066064338 The no. of frequent itemset 94 Size of maximal frequent itemset : 4	**Minimum support=0.05** The time taken 70.85965784111023 The transaction are 52131 Total no. of items 4026 Average length of transaction is 7.876676066064338 The no. of frequent itemset 36 Size of maximal frequent itemset : 2
Minimum support=0.07 The time taken 40.8963784111023 The transaction are 52131 Total no. of items 4026 Average length of transaction is 7.876676066064338 The no. of frequent itemset 23 Size of maximal frequent itemset :2	**Minimum support=0.09** The time taken 21.78585784111023 The transaction are 52131 Total no. of items 4026 Average length of transaction is 7.876676066064338 The no. of frequent itemset 12 Size of maximal frequent itemset : 2

3. Eclat algorithm:

The ECLAT algorithm stands for Equivalence Class Clustering and bottom-up Lattice Traversal

ECLAT algorithm

Step 1 — List the Transaction ID (TID) set of each product

Step 2 — Filter with minimum support

Step 3 — Compute the Transaction ID set of each product pair

Step 4 — Filter out the pairs that do not reach minimum support

Step 5— Continue as long as you can make new pairs above support

.

(T20i6D100k Dataset)

Minimum Support=0.09 The Total Time Taken 259.8486478328705 Total No. of Transaction 99922 Total no. of items 893 Space-966.9 MiB The average length of transactions 19.899791837633355 Total no. of frequent itemsets : 13 Size of maximal frequent itemset : 1	**Minimum Support=0.11** The Total Time Taken 159.69698328705 Total No. of Transaction 99922 Total no. of items 893 Space-905.0MiB The average length of transactions 19.899791837633355 Total no. of frequent itemsets : 4 Size of maximal frequent itemset : 1
Minimum Support=0.13 The Total Time Taken 109.32870585863 Total No. of Transaction 99922 Total no. of items 893 Space - 904.9 MiB The average length of transactions 19.899791837633355 Total no. of frequent itemsets : 4 Size of maximal frequent itemset : 1	**Minimum Support=0.05** The Total Time Taken 329.7832870584864 Total No. of Transaction 99922 Total no. of items 893 Space - 1250.9MiB The average length of transactions 19.899791837633355 Total no. of frequent itemsets : 99 Size of maximal frequent itemset : 1

Chess Dataset

Minimum Support=0.70 The Total Time Taken 1109.38582870563 Total No. of Transaction 3196 Total no. of items 75 The average length of transactions 37.0 Total no. of frequent itemsets : 48731 Size of maximal frequent itemset : 13	**Minimum Support=0.95** The Total Time Taken 9.3750070563 Total No. of Transaction 3196 Total no. of items 75 The average length of transactions 37.0 Total no. of frequent itemsets : 77 Size of maximal frequent itemset : 5
Minimum Support=0.75 The Total Time Taken 759.87053858263 Total No. of Transaction 3196 Total no. of items 75 The average length of transactions 37.0 Total no. of frequent itemsets : 20993 Size of maximal frequent itemset : 11	**Minimum Support=0.80** The Total Time Taken 55.38756858203 Total No. of Transaction 3196 Total no. of items 75 The average length of transactions 37.0 Total no. of frequent itemsets : 2669 Size of maximal frequent itemset : 8

Liquor Dataset

Minimum Support=0.03 The Total Time Taken 159.38587058263 Total No. of Transaction 52131 Total no. of items 4026 The average length of transactions 7.876676066064338 Total no. of frequent itemsets : 84 Size of maximal frequent itemset : 3	**Minimum Support=0.07** The Total Time Taken 209.37858058263 Total No. of Transaction 52131 Total no. of items 4026 The average length of transactions 7.876676066064338 Total no. of frequent itemsets : 23 Size of maximal frequent itemset : 2
Minimum Support=0.09 The Total Time Taken 318.38285870563 Total No. of Transaction 52131 Total no. of items 4026 The average length of transactions 7.876676066064338 Total no. of frequent itemsets : 12 Size of maximal frequent itemset : 2	**Minimum Support=0.04** The Total Time Taken 189.38268587053 Total No. of Transaction 52131 Total no. of items 4026 The average length of transactions 7.876676066064338 Total no. of frequent itemsets : 52 Size of maximal frequent itemset : 3

Observation from the above

(T20i6D100k Dataset)

For lower values of minimum support , Apriori algorithm takes around 4-5 hours to process the association rule, as it scans the dataset again again(however there is some pruning ,but still it takes more time) coming to space complexity it doesn't matter much here,as it is less than other algorithms

For FP growth, Its time taken is less than the Apriori Algorithm, But it takes space more as we have to create the tree structure so it eats a lot of memory.

For Eclat algorithm, It is better than Apriori and FP growth both in terms of space and time

Chess Dataset

Apriori takes very much time in this case (More than 6 hours, for low value of minimum support) as this dataset is quite different

In FP growth, It is faster than other algorithms and ofcourse it takes more space

In Eclat, here also it is better than others

Liquor Dataset

Here all the algorithms are faster ,taking less time as compared to other datasets , this is because of the average transaction is around 7.

Here is also the same case , Apriori takes some time greater than others
FP growth takes less time than apriori
Eclat is best of the other algorithms.

Overall Summary

Apriori Algorithm is slow.It is not efficient ie low minimum support or large itemsets i.e. it is not an efficient approach for large number of datasets.

FP Growth is better , But its implementation is complex, also it space issues
Eclat is overall the best algorithm:
It is faster,Less Memory,Less computation.