PROJECT REPORT

APRIORI ALGORITHM



Course: Data Mining

Prepared By: Vinay Basavaraja

GitHub link: https://github.com/vinayb004/Apriori_Algorithm

Introduction

Apriori Algorithm is the classical algorithm which is used in data mining. It is used for mining frequent item sets and relevant association rules for those item sets. It is devised to operate on a database containing more transactions, for instance, items brought by customers in a store.

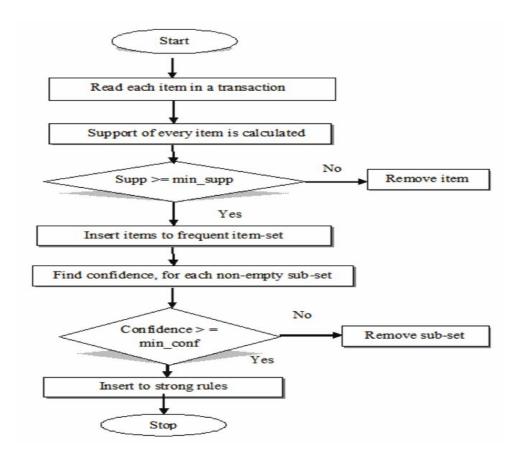
It is important for effective Market Basket Analysis and it helps the users in purchasing their products with more ease which provides more sales in the markets. It has been even used in the healthcare industry for the finding of adverse drug reactions. It produces association rules that shows what all combinations of medications and patient characteristics lead to ADRs.

Implementation

An implementation of Apriori Algorithm to generate Association rules with user input values for support and confidence. The support and confidence values are given in decimal values. It gives associations rules for different datasets: Amazon, BestBuy, KMART, Nike, Grocery.

High values of support and confidence gives an out of index value. Select a numerical value (input) for obtaining the respective data set's association rules. The csv files are uploaded too. The screenshots are attached in the file that is submitted. Check the file for report and for basic software and hardware used to implement the project, the Project is not software or hardware specific.

Implementation Architecture



Software Configuration

- 1. Environment: Anaconda-Navigator
- 2. IDE: Jupiter Notebook 6.1.4
- 3. Python version: 3.8
- 4.Libraries csv (for reader)

Source of Data:

The data used are the ones provided in the dataset examples in the midterm folder. I have also created a custom dataset for grocery store. All the data from these files are written into csv files. The datasets used are:

- 1. Amazon
- 2. BestBuy
- 3. KMART
- 4. Nike
- 5. Custom/Grocery

Hardware Configuration

• Operating System: Windows

• Processor: intel core i-7

• RAM: 16GB

Code

Coding is complete and these are the few functions used in the code

1. append_data: The purpose is to load the data set and return the dataset that contains a list of transactions and further, each transaction contains several items in it.

2. gen_Candidate1: Generate and return a set(dataset) which contains all frequent candidate 1-itemsets.

3. apriori : To check whether a frequent candidate k-itemset satisfy Apriori property and return true/false accordingly.

4. gen_Cank : Generate and return candidate k, a set which contains all frequent candidate k-itemsets.

5. gen_Lisk : Generate and return a list/set which contains all frequent k-itemsets.

6. gen_L : Generate all frequent itemsets by the value of minimum support's input.

7. association_rules: Generate and return association rules.

8. Main: Execution of functions/code and take user input for minimum support and minimum confidence.

9. Encoding used: windows-1252

Difficulties faced:

To implement the understood logic of the functions as this was my first python project.

Screenshots:

Screenshots are attached for different dataset and different user input values for support and confidence.

CSV files.:

CSV files are attached in the zip file.

GitHub Link: https://github.com/vinayb004/Apriori_Algorithm

Source Code:

from csv import reader

```
def append_data():
```

#The purpose is to load the data set and return the dataset that contains a list of transactions and further, each transaction contains several items in it.

print("Hello, welcome to association rule mining\nPlease choose the dataset for which you would like to get the Association rules")

print("Please type $\n1$ for Amazon $\n2$ for BestBuy $\n3$ for KMART $\n4$ for Nike $\n5$ for Custom")

```
while True:
  datachoice = input()
  if (datachoice == '1'):
    data = '/Users/VINAY/midterm_datamining_vinay/amazon.csv'
    print('You selected Amazon Dataset')
    break
  elif(datachoice == '2'):
    data = '/Users/VINAY/midterm_datamining_vinay/bestbuy.csv'
    print('You selected BestBuy Dataset')
    break
  elif(datachoice == '3'):
    data = '/Users/VINAY/midterm_datamining_vinay/kmart.csv'
    print('You selected KMART Dataset')
    break
  elif(datachoice == '4'):
    data = '/Users/VINAY/midterm_datamining_vinay/nike.csv'
    print('You selected Nike Dataset')
    break
  elif(datachoice == '5'):
```

```
data = '/Users/VINAY/midterm_datamining_vinay/grocery.csv'
       print('You selected Custom Dataset')
       break
     else:
       print("Please select a valid choice from the above list")
  with open(data, 'r',encoding="windows-1252") as read_obj:
     # pass the file object to reader() to get the reader object
     csv_reader = reader(read_obj)
     # Pass reader object to list() to get a list of lists
     dataset = list(csv_reader)
     for items in dataset:
       for j in range(0, len(items)):
          for items1 in items:
            if items1 == "":
               items.remove("")
  return dataset
def gen_Candidate1(dataset):
  # Generate and return a set(dataset) which contains all frequent candidate 1-itemsets.
  Can1 = set()
  for t in dataset:
     for item in t:
       item_set = frozenset([item])
       Can1.add(item_set)
  return Can1
```

```
def apriori(Cank_item, Lisk1):
```

#To check whether a frequent candidate k-itemset satisfy Apriori property and return true/false accordingly.

```
for item in Cank_item:
     sub_Cank = Cank_item - frozenset([item])
     if sub_Cank not in Lisk1:
       return False
  return True
def gen_Cank(Lisk1, k):
  #Generate and return candidate k, a set which contains all frequent candidate k-itemsets.
  Cank = set()
  length\_Lisk1 = len(Lisk1)
  listof\_Lisk1 = list(Lisk1)
  for i in range(length_Lisk1):
     for j in range(1, length_Lisk1):
       list1 = list(listof_Lisk1[i])
       list2 = list(listof_Lisk1[j])
       list1.sort()
       list2.sort()
       if list1[0:k-2] == list2[0:k-2]:
          Cank_item = listof_Lisk1[i] | listof_Lisk1[j]
          if apriori(Cank_item, Lisk1):
            Cank.add(Cank_item)
  return Cank
```

```
def gen_Lisk(dataset, Cank, minsup, sup_data):
  #Generate and return a list/set which contains all frequent k-itemsets.
  Lisk = set()
  item_count = {}
  for t in dataset:
     for item in Cank:
       if item.issubset(t):
          if item not in item_count:
            item_count[item] = 1
          else:
            item_count[item] += 1
  t_num = float(len(dataset))
  for item in item_count:
     if (item_count[item] / t_num) >= minsup:
       Lisk.add(item)
       sup_data[item] = item_count[item] / t_num
  return Lisk
def gen_L(dataset, k, minsup):
  #Generate all frequent itemsets by the value of minimum support's input.
  \sup_{a} data = \{ \}
  Can1 = gen_Candidate1(dataset)
  List1 = gen_Lisk(dataset, Can1, minsup, sup_data)
  Lisksub1 = List1.copy()
  List = []
  List.append(Lisksub1)
```

```
for i in range(2, k+1):
     Ci = gen_Cank(Lisksub1, i)
     Li = gen_Lisk(dataset, Ci, minsup, sup_data)
     Lisksub1 = Li.copy()
     List.append(Lisksub1)
  return List, sup_data
def association_rules(List, sup_data, minconf):
  #Generate and return association rules.
  asso_rules = []
  sublist = []
  for i in range(0, len(List)):
     for freq_set in List[i]:
       for sub_set in sublist:
          if sub_set.issubset(freq_set):
            conf = sup_data[freq_set] / sup_data[freq_set - sub_set]
            rule = (freq_set - sub_set, sub_set, conf)
            if conf >= minconf and rule not in asso_rules:
               # print freq_set-sub_set, " => ", sub_set, "conf: ", conf
               asso_rules.append(rule)
       sublist.append(freq_set)
  return asso_rules
if __name__ == "__main__":
  #print("Please enter the min_support")
  #inp = input("Enter the minimum support: ")
```

```
dataset = append_data()
  List, sup_data = gen_L(dataset, k=3, minsup = float(input("Enter the minimum support: ")))
  rules_list = association_rules(List, sup_data, minconf= float(input("Enter the minimum
Confidence: ")))
  for Lisk in List:
     print("-*"*40)
     print("frequent " + str(len(list(Lisk)[0])) + "-itemsets\t\tsupport")
     print("-*"*40)
     for freq_set in Lisk:
       print(freq_set, sup_data[freq_set])
  print("\n")
  print("#-"*40)
  print("Association Rules with Confidence")
  print("#-"*40)
  for item in rules_list:
     print(item[0], "=>", item[1], "confidence: ", item[2])
```

Screenshots:

1.Amazon

```
X
 Command Prompt
                                                                                                                                                                                                                                        C:\Users\VINAY>cd midterm datamining vinay
C:\Users\VINAY\midterm_datamining_vinay>python midterm_apriori.py
 Hello, welcome to association rule mining
Please choose the dataset for which you would like to get the Association rules
Please type
1 for Amazon
2 for BestBuy
3 for KMART
4 for Nike
 5 for Custom
You selected Amazon Dataset
frequent 1-itemsets
                                                              support
                                                                        a
Froguent 1:itemsets support

Frozenset('Java For Dumnies')) 0.6

Frozenset(('Android Programming: The Big Nerd Ranch')) 0.65

Frozenset(('Java: The Complete Reference')) 0.45

Frozenset(('Java: The Complete Reference')) 0.4

Frozenset(('Had First Java 2nd Edition')) 0.4

Frozenset(('Hill. and CSS: Design and Build Websites')) 0.1

Frozenset(('Beginning Programming with Java')) 0.3

Frozenset('Reginning Programming with Java')) 0.5

Frozenset('A Beginner's Guide')) 0.55
 ozenset(('Java For Dummies', 'Java: The Complete Reference', 'A Beginner's Guide')) 0.4

ozenset(('Java For Dummies', 'A Beginner's Guide', 'Android Programming: The Big Nerd Ranch')) 0.25

ozenset(('Java For Dummies', 'Head First Java 2nd Edition', 'Android Programming: The Big Nerd Ranch')) 0.15

ozenset(('Java: The Complete Reference', 'A Beginner's Guide', 'Android Programming: The Big Nerd Ranch')) 0.25

ozenset(('Java For Dummies', 'Java: The Complete Reference', 'Android Programming: The Big Nerd Ranch')) 0.3

ozenset(('Java For Dummies', 'HTML and CSS: Design and Build Websites', 'A Beginner's Guide')) 0.1

ozenset(('Java For Dummies', 'HTML and CSS: Design and Build Websites', 'Java: The Complete Reference')) 0.1

ozenset(('HTML and CSS: Design and Build Websites', 'Java: The Complete Reference', 'A Beginner's Guide')) 0.1

ozenset(('HTML and CSS: Design and Build Websites', 'Java: The Complete Reference', 'A Beginner's Guide')) 0.1
```

2.BestBuy

```
0
```

3.KMART

```
Select Command Prompt
Hello, welcome to association rule mining
Please choose the dataset for which you would like to get the Association rules
 Please type
      for Amazon
2 for BestBuy
3 for KMART
4 for Nike
5 for Custom
  You selected KMART Dataset
"Redect Command Prompt

"Acquent 3 - itemsets support

"Prozenset(('Socks', 'Sweatahirts', 'Modern Pants')) 8.4

"Prozenset(('Hoodies', 'Tech Pants', 'Dry Fit V-Nick')) 8.35

"Prozenset(('Swimming Shirt', 'Tech Pants', 'Dry Fit V-Nick')) 8.35

"Prozenset(('Tech Pants', 'Rash Guard', 'Swimming Shirt')) 8.35

"Prozenset(('Swimming Shirt', 'Rash Guard', 'Dry Fit V-Nick')) 8.4

"Prozenset(('Rounding Shoe', 'Rash Guard', 'Dry Fit V-Nick')) 8.4

"Prozenset(('Rounding Shoe', 'Socks', 'Sweatahirts')) 8.5

"Prozenset(('Rounding Shoe', 'Socks', 'Sweatahirts')) 8.5

"Prozenset(('Hoodies', 'Rash Guard', 'Dry Fit V-Nick')) 8.35
  frozenset(('Socks')) => frozenset(('Modern Pants')) => frozenset(('Socks')) confidence: 0.8

is Select Command Prompt
frozenset('Sickming Shirt', 'Dny Fit V-Nick')) => frozenset(('Sockming Shirt', 'Rash Guard')) => frozenset(('Sickming Shirt', 'Rash Guard')) => frozenset(('Sickming Shirt', 'Rash Guard')) => frozenset(('Sickming Shirt', 'Pash Guard')) => frozenset('Sickming Shirt', 'Pash Guard') => frozenset('Sickming Shirt') => frozenset('Sick') => frozenset('Tach Pants') == (Sick') == (Sick') => frozenset(('Tach Pants', 'Dny Fit V-Nick')) => frozenset('Tach Pants') == (Sick') => frozenset(('Tach Pants', 'Pash Guard')) => frozenset(('Rash Guard')) => froze
```

```
a
          elected Nike Dataset
the minimum support: 0.35
the minimum Confidence: .4
                t l-itemsets support

at(('Swimming Shirt')) 0.55

at(('Tech Pants')) 0.45

at(('Modern Pants')) 0.5

at(('Modern Pants')) 0.5

at(('Modern Pants')) 0.7

at(('Running Shoe')) 0.7

at(('Running Shoe')) 0.7

at(('Indesis')) 0.4

at(('Rach Guard')) 0.65

at(('Pach Guard')) 0.5
            nset(('Dry Fit V-Nick')) 0.5

-1.-temmed's

nset(('Hoodiss', 'Dry Fit V-Nick')) 0.35

nset(('Sudmming Shirt', 'Dry Fit V-Nick')) 0.45

nset(('Sudmming Shirt', 'Rash Guard')) 0.5

nset(('Sweatchirts', 'Modern Pants')) 0.5

nset(('Sweatchirts', 'Modern Pants')) 0.5

nset(('Teah Guard', 'Dry Fit V-Nick')) 0.45

nset(('Teah Grants', 'Rash Guard')) 0.4

nset(('Teah Grants', 'Rash Guard')) 0.4

nset(('Teah Grants', 'Rash Guard')) 0.4

nset(('Teah Grants', 'Rash Guard')) 0.5

nset(('Teah Grants') 0.4

nset(('Rouning Shoe', 'Sweatchirts')) 0.5

nset(('Running Shoe', 'Sweatchirts')) 0.5

nset(('Running Shoe', 'Sweatchirts')) 0.5

nset(('Swimming Shoe', 'Modern Pants')) 0.45

nset('Swimming Shoe', 'Modern Pants')) 0.45

nset('Swimming Shoe', 'Modern Pants')) 0.45

nset('Swimming Shoe', 'Modern Pants')) 0.35

nset('Grants Shoe')
   Select Command Prompt
  Select Command Prompt
                                                                                                                                                                                                                                                                                                                                                                   n.
```

```
C:\Users\VINAY\midterm_datamining_vinay>python midterm_apriori.py
Hello, welcome to association rule mining
Please choose the dataset for which you would like to get the Association rules
Please type
1 for Amazon
2 for BestBuy
3 for WMAET
       frequent 1-itemsets support

frozenset(('BOURNVITA')) 0.2
frozenset(('NAGGI')) 0.35
frozenset(('INAGGI')) 0.35
frozenset(('INAGI')) 0.35
frozenset(('INILK')) 0.4
frozenset(('INILK')) 0.4
frozenset(('INILK')) 0.5
frozenset(('INILK')) 0.1
frozenset(('INILK')) 0.2
frozenset(('INILK')) 0.2
frozenset(('INILK')) 0.1
frozenset(('INILK')) 0.1
frozenset(('INILK')) 0.1
frozenset(('INILK')) 0.2
frozenset(('INILK')) 0.1
frozenset(('INILK')) 0.1
frozenset(('INILK')) 0.2
frozenset(('INILK')) 0.1
             Command Prompt
                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                                 0 :
    Command Prompt

Frozenset(('BREAD', 'BISCUIT')) 0.1

frozenset(('JAM', 'MAGGI')) 0.1

frozenset(('JAM', 'MAGGI')) 0.1

frozenset(('JAM', 'BREAD') 0.1

frozenset(('JAM', 'BREAD') 0.1

frozenset(('TEA', 'BISCUIT')) 0.1

frozenset(('TEA', 'BISCUIT', 'MAGGI')) 0.1

frozenset(('GNRFALMES', 'COCK', 'BISCUIT')) 0.1

frozenset(('TEA', 'BREAD', 'BURNUTA')) 0.1

frozenset(('GFEE', 'CMNFLMES', 'COCK') 0.1

frozenset(('GFFEE', 'CORNFLMES', 'GOK')) 0.1

frozenset(('GFFEE', 'BREAD', 'GORNELAKS'), 'BISCUIT')) 0.1

frozenset(('GFFEE', 'BREAD', 'GORNELAKS'), 'BISCUIT')) 0.1

frozenset(('GFFEE', 'GOKNFLMES', 'BISCUIT')) 0.1

frozenset(('GFFEE', 'GOKNFLMES', 'BISCUIT')) 0.1

frozenset(('GFFEE', 'GOKNFLMES', 'BISCUIT')) 0.1

frozenset(('GFFEE', 'COCK', 'BISCUIT')) 0.1

frozenset(('TEA', 'BREAD', 'MAGGI')) 0.1
```

Command Prompt

đ

```
Command Prompt
   et({'COFFEE'}) => fro
   Ø
Command Prompt
                          0.28571428571428575
 \Users\VINAY\midterm_datamining_vinay>
```

n