МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И.УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЁТ

по лабораторной работе №3
по дисциплине «Машиное обучение»
Тема: Частотный анализ

Студент гр. 6304 _____ Корытов П.В.
Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург
2020

1. Цель работы

Ознакомится с методами частотного анализа из библиотеки `MLxtend`.

2. Выполнение

2.1. Загрузка данных

Произведена загрузка данных. В датасете нашлось 1139 уникальных id покупателей и 38 различных товаров.

2.2. Подготовка данных

Произведено one-hot кодирование датасета с помощью `TransactionEncoder`. Результаты на листинге 1.

Листинг 1. Закодированный датасет

```
      all- purpose  aluminum foil  bagels  ...  vegetables  waffles  yogurt
0             True           True   False  ...        True    False    True
1            False           True   False  ...        True     True    True
2            False          False    True  ...        True    False   False
3             True          False   False  ...       False    False   False
4             True          False   False  ...        True     True    True
...            ...            ...     ...  ...         ...      ...     ...
1134          True          False   False  ...       False    False   False
1135         False          False   False  ...        True    False   False
1136         False          False    True  ...        True    False    True
1137          True          False   False  ...        True     True    True
1138         False          False   False  ...        True    False   False

[1139 rows x 38 columns]
```

2.3. Ассоциативный анализ

1. Применен алгоритм `apriori` с уровнем поддержки 0.3. Результаты (первые 10 строчек) на листинге 2.

   Для наборов длиной 1 алгоритм уровень поддержки — верноятность этого единственного представителя находится в транзакции.

   Для наборов длиной > 1 — вероятность одновременно находится в одной транзакции. Наборы длиной больше 1 представлены на листинге 3.

Листинг 2. Результаты алгоритма

```
    support                                itemsets  length
0  0.374890                  (all- purpose)       1
1  0.384548                 (aluminum foil)       1
2  0.385426                        (bagels)       1
3  0.374890                          (beef)       1
4  0.367867                        (butter)       1
5  0.395961                       (cereals)       1
6  0.390694                       (cheeses)       1
7  0.379280                     (coffee/tea)       1
8  0.388938                   (dinner rolls)       1
```

Листинг 3. Наборы длиной больше 1

```
    support                                    itemsets  length
38  0.310799                (aluminum foil, vegetables)       2
39  0.300263                       (vegetables, bagels)       2
40  0.310799                     (vegetables, cereals)       2
41  0.309043                     (vegetables, cheeses)       2
42  0.308165                 (vegetables, dinner rolls)       2
43  0.306409  (vegetables, dishwashing liquid/detergent)       2
44  0.326602                         (eggs, vegetables)       2
45  0.302897                  (vegetables, ice cream)       2
46  0.309043           (vegetables, laundry detergent)       2
```

2. Определено количество наборов при различных уровнях поддержки. Результаты на рис. 1.

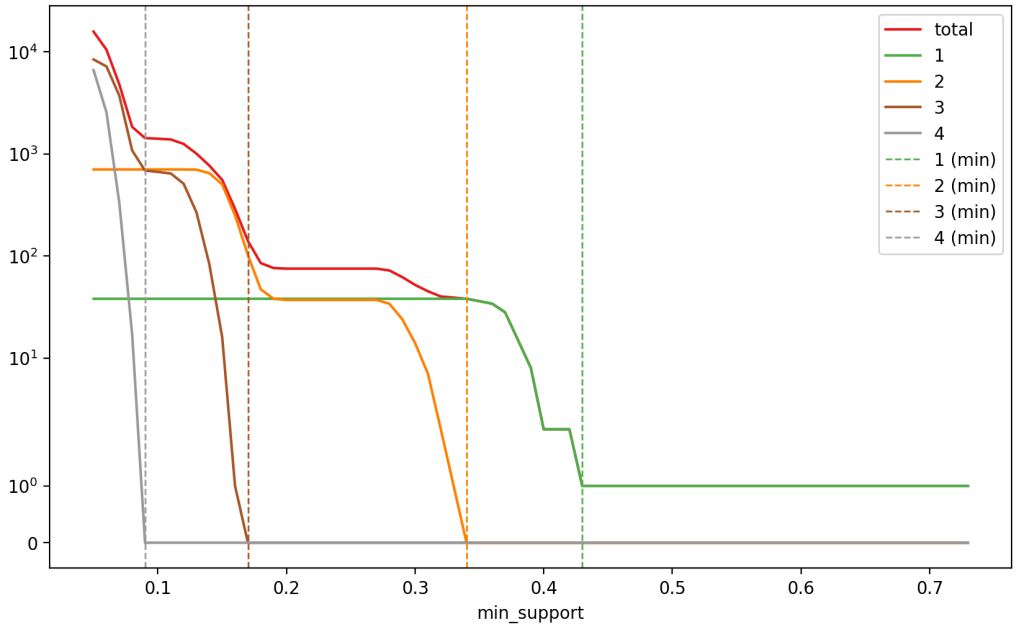

Рисунок 1 – Количество наборов при различных уровнях поддержки

3. В наборе данных оставлены только те элементы, который попадают в наборы размеров 1 при уровне поддержки 0.38. Закодированный набор представлен на листинге 4.

Листинг 4. Набор данных

```
      aluminum foil  bagels  cereals  ...  vegetables  waffles  yogurt
0              True   False    False  ...        True    False    True
1              True   False     True  ...        True     True    True
2             False    True     True  ...        True    False   False
3             False   False     True  ...       False    False   False
4             False   False    False  ...        True     True    True
...             ...     ...      ...  ...         ...      ...     ...
1134          False   False     True  ...       False    False   False
1135          False   False     True  ...        True    False   False
1136          False    True    False  ...        True    False    True
1137          False   False    False  ...        True     True    True
1138          False   False    False  ...        True    False   False

[1139 rows x 15 columns]
```

4. Произведен ассоциативный анализ нового набора данных. Результаты представлены на листинге 5.

Листинг 5. Новые результаты

```
     support                                    itemsets  length
0   0.384548                             (aluminum foil)       1
1   0.385426                                    (bagels)       1
2   0.395961                                   (cereals)       1
3   0.390694                                   (cheeses)       1
4   0.388938                              (dinner rolls)       1
5   0.388060          (dishwashing liquid/detergent)       1
6   0.389816                                     (eggs)       1
7   0.398595                                (ice cream)       1
8   0.395083                               (lunch meat)       1
9   0.380158                                     (milk)       1
10  0.421422                                  (poultry)       1
11  0.390694                                     (soda)       1
12  0.739245                                (vegetables)       1
13  0.394205                                   (waffles)       1
14  0.384548                                    (yogurt)       1
15  0.310799               (aluminum foil, vegetables)       2
16  0.300263                      (vegetables, bagels)       2
17  0.310799                     (vegetables, cereals)       2
18  0.309043                     (cheeses, vegetables)       2
19  0.308165                (vegetables, dinner rolls)       2
20  0.306409  (vegetables, dishwashing liquid/detergent)       2
21  0.326602                        (vegetables, eggs)       2
22  0.302897                   (vegetables, ice cream)       2
23  0.311677                  (vegetables, lunch meat)       2
24  0.331870                     (vegetables, poultry)       2
25  0.305531                        (vegetables, soda)       2
26  0.315189                     (vegetables, waffles)       2
27  0.319579                      (vegetables, yogurt)       2
```

Теперь наборы длиной 1 имееют минимальный уровень поддержки, равный 0.38.

5. Проведен ассоциативный анализ нового набора при уровне поддержки 0.15. Код на листинге 6, результаты на листинге 7.

Листинг 6. Ассоциативный анализ нового набора

```
new2_results = apriori(new_df, min_support=0.15, use_colnames=True)
new2_results['length'] = new2_results['itemsets'].apply(len)
new2_results = new2_results[
    new2_results.apply(
        lambda d: d['length'] > 1 and ('yougurt' in d['itemsets'] or 'waffles' in d['itemsets']), axis=1
    )
].reset_index(drop=True)

with open('output/new2_results.txt', 'w') as f:
    f.write(str(new2_results))

new2_results
```

Листинг 7. Результаты ассоциативного анализа нового набора

```
     support                                itemsets  length
0   0.169447                (aluminum foil, waffles)       2
1   0.159789                       (waffles, bagels)       2
2   0.160667                      (waffles, cereals)       2
3   0.172959                      (cheeses, waffles)       2
4   0.169447                 (waffles, dinner rolls)       2
5   0.175593  (waffles, dishwashing liquid/detergent)       2
6   0.169447                         (eggs, waffles)       2
7   0.172959                    (waffles, ice cream)       2
8   0.184372                   (waffles, lunch meat)       2
9   0.166813                      (waffles, poultry)       2
10  0.177349                         (waffles, soda)       2
11  0.315189                   (vegetables, waffles)       2
12  0.173837                       (waffles, yogurt)       2
13  0.157155       (vegetables, waffles, lunch meat)       3
```

6. Составлен ещё один датасет из элементов, не попавших в датасет п.6 задания. Результаты на листинге 8.

   Как видно, в наборе данных 23 элемента. В наборе на листинге 4 — 15, на листинге 1 сумма — 38.

Листинг 8. Ещё один набор данных

```
      all- purpose   beef  butter  ...  sugar  toilet paper  tortillas
0             True   True    True  ...  False         False      False
1            False  False   False  ...  False          True       True
2            False  False   False  ...  False          True      False
3             True  False   False  ...  False          True      False
4             True  False   False  ...  False          True       True
...            ...    ...     ...  ...    ...           ...        ...
1134          True   True   False  ...   True         False      False
1135         False  False   False  ...  False         False      False
1136         False   True   False  ...   True         False       True
1137          True   True   False  ...   True          True      False
1138         False  False   False  ...  False         False      False

[1139 rows x 23 columns]
```

7. Произведен apriori анализ для нового набора данных с минимальным уровнем поддержки 0.15.

8. Выбраны все наборы, в которых 2 элемента начинаются с “s”. Код представлен на листинге 9, результаты — на листинге 10.

Листинг 9. Правило для выбора наборов с двумя “s“в начале

```
s_results = new3_results[new3_results.apply(
    lambda r: len([e for e in r['itemsets'] if e.startswith('s')]) > 1,
    axis=1
)]

with open('output/s_results.txt', 'w') as f:
    f.write(str(s_results))

s_results
```

Листинг 10. Наборы с двумя “s” в начале

```
      support                           itemsets
248  0.137840    (sandwich bags, sandwich loaves)
249  0.146620             (sandwich bags, shampoo)
250  0.158911                (soap, sandwich bags)
251  0.147498    (sandwich bags, spaghetti sauce)
252  0.131694               (sandwich bags, sugar)
255  0.150132          (shampoo, sandwich loaves)
256  0.158033             (soap, sandwich loaves)
257  0.150132  (spaghetti sauce, sandwich loaves)
258  0.136962            (sugar, sandwich loaves)
261  0.151010                      (soap, shampoo)
262  0.139596          (spaghetti sauce, shampoo)
263  0.147498                     (sugar, shampoo)
266  0.160667             (soap, spaghetti sauce)
267  0.154522                        (soap, sugar)
270  0.144864            (spaghetti sauce, sugar)
```

9. Выбраны все наборы, для которых уровень поддержки лежит в промежутке $[0.1, 0.25]$. Код на листинге 11, результаты — на листинге 12.

Листинг 11. Правило для выбора наборов с уровнем поддержки в промежутке $[0.1, 0.25]$

```
# new3_results[new3_results['support'] > 0.1 | new3_results['support'] < 0.25]
ss_results = new3_results[(new3_results['support'] > 0.1) &
↪  (new3_results['support'] < 0.25)]

with open('output/ss_results.txt', 'w') as f:
    f.write(str(ss_results))

ss_results
```

Листинг 12. Наборы с уровнем поддержки в промежутке $[0.1, 0.25]$

```
      support                          itemsets
23   0.144864             (all- purpose, beef)
24   0.147498           (butter, all- purpose)
25   0.146620       (coffee/tea, all- purpose)
26   0.142230            (all- purpose, flour)
27   0.150132           (fruits, all- purpose)
..        ...                              ...
271  0.151888  (spaghetti sauce, toilet paper)
272  0.148376     (spaghetti sauce, tortillas)
273  0.151888            (sugar, toilet paper)
274  0.147498               (sugar, tortillas)
275  0.156277        (tortillas, toilet paper)

[253 rows x 2 columns]
```

3. Выводы

Изучен метод `apriori` частотного анализа из библиотеки `MLxtend`.

Установлено, что увеличение уровня поддержки для рассмотренного набора данных ведет в первую очередь к уменьшению количества наборов большей длины.