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

ОТЧЕТ

по лабораторной работе №4
по дисциплине «Машинное обучение»
Тема: Ассоциативный анализ

Студентка гр. 8304 _____ Сергеев А. Д.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург
2021

Цель работы

Ознакомиться с методами ассоциативного анализа из библиотеки MLxtend.

Ход работы

1.1 Загрузка данных

1 Был загружен датасет по ссылке, где данные представлены в виде csv таблицы.

2 Данные были переформированы и из них были удалены все значения NaN.

3 Был получен список всех уникальных товаров.

```
Уникальные товары: {'dish cleaner', 'tropical fruit', 'bags', 'salad dressing', 'cake bar', 'nut snack', 'oil', 'kitchen towels', 'sparkl
ing wine', 'liquor', 'decalcifier', 'cream', 'other vegetables', 'Instant food products', 'organic sausage', 'preservation products', 'ch
ewing gum', 'waffles', 'popcorn', 'cling film/bags', 'meat', 'light bulbs', 'softener', 'photo/film', 'beef', 'rubbing alcohol', 'brown b
read', 'sauces', 'seasonal products', 'dishes', 'canned vegetables', 'cleaner', 'newspapers', 'cooking chocolate', 'baking powder', 'ja
m', 'toilet cleaner', 'frozen fish', 'misc. beverages', 'baby cosmetics', 'abrasive cleaner', 'hygiene articles', 'flower soil/fertilize
r', 'dental care', 'salt', 'candy', 'candles', 'turkey', 'butter milk', 'beverages', 'pudding powder', 'house keeping products', 'rum',
'frozen vegetables', 'meat spreads', 'tidbits', 'frozen fruits', 'frankfurter', 'cat food', 'condensed milk', 'pasta', 'white bread', 'fr
uit/vegetable juice', 'curd', 'onions', 'mayonnaise', 'rice', 'specialty fat', 'hair spray', 'liqueur', 'dessert', 'whole milk', 'nuts/pr
unes', 'root vegetables', 'bottled water', 'skin care', 'detergent', 'semi-finished bread', 'snack products', 'hamburger meat', 'pip frui
t', 'packaged fruit/vegetables', 'pork', 'flower (seeds)', 'chocolate', 'frozen meals', 'white wine', 'sweet spreads', 'ham', 'specialty
chocolate', 'cream cheese', 'fish', 'spices', 'frozen chicken', 'cocoa drinks', 'canned fruit', 'make up remover', 'potted plants', 'coff
ee', 'artif. sweetener', 'cookware', 'chocolate marshmallow', 'zwieback', 'spread cheese', 'baby food', 'hard cheese', 'rolls/buns', 'mus
tard', 'specialty bar', 'specialty vegetables', 'prosecco', 'ice cream', 'citrus fruit', 'chicken', 'whipped/sour cream', 'UHT-milk', 'so
ft cheese', 'whisky', 'specialty cheese', 'bottled beer', 'cereals', 'grapes', 'sliced cheese', 'sound storage medium', 'ketchup', 'froze
n dessert', 'liver loaf', 'pastry', 'vinegar', 'canned beer', 'yogurt', 'canned fish', 'pet care', 'red/blush wine', 'finished products',
'butter', 'dog food', 'tea', 'brandy', 'kitchen utensil', 'organic products', 'margarine', 'napkins', 'ready soups', 'long life bakery pr
oduct', 'soda', 'berries', 'roll products', 'instant coffee', 'salty snack', 'herbs', 'female sanitary products', 'sugar', 'curd cheese',
'pickled vegetables', 'honey', 'bathroom cleaner', 'shopping bags', 'domestic eggs', 'processed cheese', 'flour', 'potato products', 'sou
ps', 'sausage', 'male cosmetics', 'soap', 'liquor (appetizer)', 'frozen potato products', 'syrup'}
Всего: 169
```

Рисунок 1 - Уникальные товары

2 FPGrowth и FPMax

1 Данные были преобразованы к виду, удобному для анализа.

2 Был проведен ассоциативный анализ, используя алгоритм FPGrowth при уровне поддержки 0.03.

	support	itemsets
0	0.082766	(citrus fruit)
1	0.058566	(margarine)
2	0.139502	(yogurt)
3	0.104931	(tropical fruit)
4	0.058058	(coffee)
...	...	...
58	0.033249	(whole milk, pastry)
59	0.047382	(root vegetables, other vegetables)
60	0.048907	(root vegetables, whole milk)
61	0.030605	(sausage, rolls/buns)
62	0.032232	(whipped/sour cream, whole milk)

63 rows × 2 columns

Рисунок 2 – FPGrowth при уровне поддержки 0.03

3 Были проанализированы получившиеся варианты. Было определено минимальное и максимальное значения для уровня поддержки для набора из 1,2, и.т.д. объектов.

```
Для набора из 1 элементов:
        максимальное значение - 0.25551601423487547 ['whole milk']
        минимальное - 0.03040162684290798 ['specialty chocolate']

Для набора из 2 элементов:
        максимальное значение - 0.07483477376715811 ['whole milk', 'other vegetables']
        минимальное - 0.030096593797661414 ['pip fruit', 'whole milk']
```

Рисунок 3 – Минимальный и максимальный уровень поддержки для наборов различной длины в алгоритме FPGowth

4 Был проведен аналогичный анализ, используя алгоритм FPMax. Результат представлен на рисунках 4 и 5.

	support	itemsets
0	0.030402	(specialty chocolate)
1	0.031012	(onions)
2	0.032944	(hygiene articles)
3	0.033249	(berries)
4	0.033249	(hamburger meat)
5	0.033452	(UHT-milk)
6	0.033859	(sugar)
7	0.037112	(dessert)
8	0.037417	(long life bakery product)
9	0.037824	(salty snack)
10	0.038434	(waffles)
11	0.039654	(cream cheese)
12	0.042095	(white bread)
13	0.042908	(chicken)
14	0.048094	(frozen vegetables)
15	0.049619	(chocolate)
16	0.052364	(napkins)
17	0.052466	(beef)
18	0.053279	(curd)
19	0.055414	(butter)
20	0.057651	(pork)
21	0.058058	(coffee)

Рисунок 4 - FPMax при уровне поддержки 0.03

5 Были проанализированы получившиеся варианты. Было определено минимальное и максимальное значения для уровня поддержки для набора из 1,2, и.т.д. объектов.

```
Для набора из 1 элементов:
        максимальное значение - 0.09852567361464158 ['shopping bags']
        минимальное - 0.03040162684290798 ['specialty chocolate']

Для набора из 2 элементов:
        максимальное значение - 0.07483477376715811 ['whole milk', 'other vegetables']
        минимальное - 0.030096593797661414 ['whole milk', 'pip fruit']
```

Рисунок 5 - Минимальный и максимальный уровень поддержки для наборов различной длины в алгоритме FPGowth

6 Из рисунков 3 и 5 видно, что отличается только максимальный уровень поддержки для длины набора 1. Так произошло из-за того, что в алгоритме FPMax один набор не может быть частью другого. Наиболее часто встречающиеся наборы длины 1 вошли в наборы длины 2.

7 Была построена гистограмма для каждого товара. Столбцы на гистограмме были упорядочены по уменьшению частоты. Сравнивая результаты гистограммы и рисунка 2, можно прийти к выводу, что уровень поддержки прямо пропорционален количеству товара.

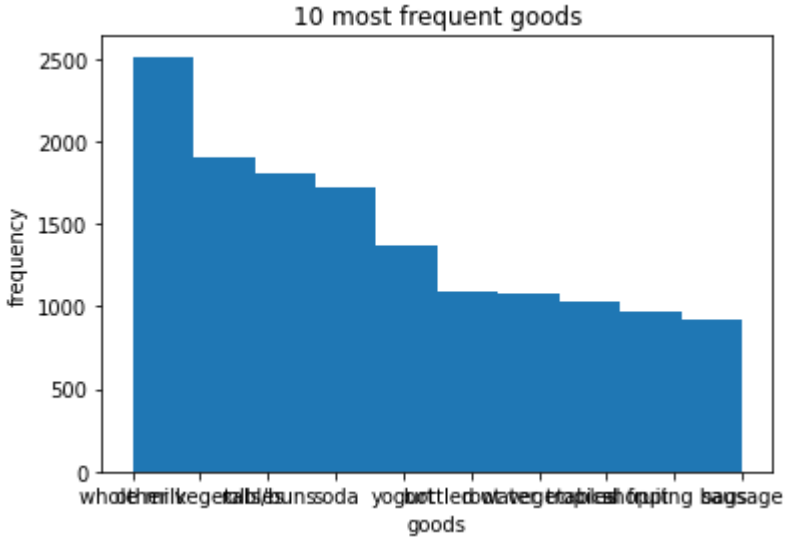

Рисунок 6 - Гистограмма 10 самых часто встречающихся товаров

8 Был преобразован набор данных так, чтобы он содержал ограниченный набор товаров.

9 Был проведен анализ FPGrowth и FPMax для нового набора данных. Т.к. были удалены товары с минимальным уровнем поддержки, то

минимальные значения для FPGrowth и FPMax увеличились. Максимальные остались без изменений.

```
Для набора из 1 элементов:
        максимальное значение - 0.09852567361464158 ['shopping bags']
        минимальное - 0.05765124555160142 ['pork']

Для набора из 2 элементов:
        максимальное значение - 0.07483477376715811 ['whole milk', 'other vegetables']
        минимальное - 0.030503304524656837 ['whole milk', 'citrus fruit']

Для набора из 1 элементов:
        максимальное значение - 0.25551601423487547 ['whole milk']
        минимальное - 0.05765124555160142 ['pork']

Для набора из 2 элементов:
        максимальное значение - 0.07483477376715811 ['whole milk', 'other vegetables']
        минимальное - 0.030503304524656837 ['whole milk', 'citrus fruit']
```

Рисунок 7 и 8 - Анализ для ограниченного списка товаров

10 Был построен график изменения количества получаемых правил от уровня поддержки. Результат представлен на рисунке 8. На графике отдельно отображены кривые для набора товаров 1, 2, и т.д. Количество наборов уменьшается с увеличением уровня минимальной поддержки.

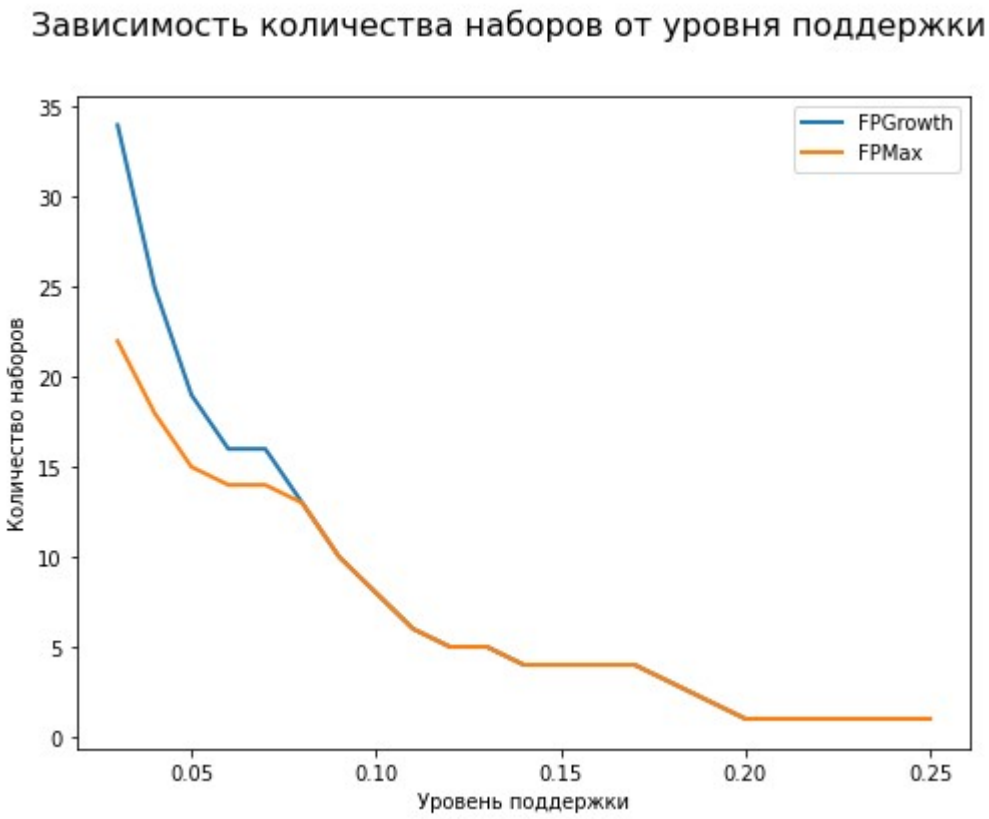

Рисунок 9 - Зависимость количества наборов от уровня поддержки

3 Ассоциативные правила

1 Были получены частоты наборов используя алгоритм FPGrowth.

2 Был проведен ассоциативный анализ.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(citrus fruit)	(whole milk)	0.082766	0.255516	0.030503	0.368550	1.442377	0.009355	1.179008
1	(citrus fruit)	(other vegetables)	0.082766	0.193493	0.028876	0.348894	1.803140	0.012862	1.238674
2	(whole milk, citrus fruit)	(yogurt)	0.030503	0.139502	0.010269	0.336667	2.413350	0.006014	1.297233
3	(citrus fruit, yogurt)	(whole milk)	0.021657	0.255516	0.010269	0.474178	1.855768	0.004736	1.415849
4	(citrus fruit, yogurt)	(other vegetables)	0.021657	0.193493	0.007626	0.352113	1.819773	0.003435	1.244827
...	...	...	...	...	...	...	...	...	...
193	(pork, rolls/buns)	(other vegetables)	0.011286	0.193493	0.005592	0.495495	2.560798	0.003408	1.598613
194	(root vegetables, pork)	(other vegetables)	0.013625	0.193493	0.007016	0.514925	2.661214	0.004379	1.662646
195	(other vegetables, pork)	(root vegetables)	0.021657	0.108998	0.007016	0.323944	2.972002	0.004655	1.317940
196	(root vegetables, pork)	(whole milk)	0.013625	0.255516	0.006812	0.500000	1.956825	0.003331	1.488968
197	(whole milk, pork)	(root vegetables)	0.022166	0.108998	0.006812	0.307339	2.819667	0.004396	1.286347

198 rows × 9 columns

Рисунок 10 – Ассоциативный анализ

3 *Поддержка* — это показатель, показывающий то, насколько часто набор объектов встречается в базе данных. Поддержка набора X определяется как отношение числа транзакций, содержащих набор X, к общему числу транзакций.

Достоверность правила (доверие) — это показатель, насколько часто правило оказывается верным, это условная вероятность того, что транзакция содержит консеквент Y, при условии, что он содержит антецедент X: $conf(X \to Y) = P(Y|X)$.

Лифт определяется как отношение наблюдаемой совместной вероятности X и Y к ожидаемой совместной вероятности, если бы они были статистически независимыми, то есть $lift(X \to Y) = \frac{P(XY)}{P(X) * P(Y)}$. Если правило имеет лифт 1, это означает, что событие в левой части независимо от события в правой части. Если лифт > 1, это позволяет нам знать степень, насколько события связаны друг с другом.

Усиление измеряет разницу между наблюдаемой и ожидаемой совместной вероятностью XY при условии, что X и Y независимы: $leverage(X \to Y) = P(XY) - P(X) * P(Y)$.

Уверенность измеряет ожидаемую ошибку правила, то есть, как часто X встречается в транзакции, а Y - нет. Таким образом, это мера силы правила по отношению к дополнению консеквента.

$$\mathrm{conv}(X \Rightarrow Y) = \frac{1 - \mathrm{supp}(Y)}{1 - \mathrm{conf}(X \Rightarrow Y)}$$

4 Было проведено построение ассоциативных правил для различных метрик (значение min_threshold такое, чтобы выводилось не менее 10 правил).

5 Было рассчитано среднее значение, медиана и СКО для каждой из метрик.

Таблица 1 – Расчет СКО, медианы и среднего значения.

	Antecedent support	**Consequent support**	**Support**	**Confidence (доверие)**	**Lift (лифт)**	**Leverage (усиление)**	**Conviction (уверенность)**
Count	10	10	10	10	10	10	10
Среднее значение	0.0125	0.2183	0.00748	0.60367	2.80204	0.0048	1.9903
СКО	0.00418	0.03203	0.00233	0.04207	0.3158	0.00163	0.20291
Медиана	0.01215	0.19349	0.00732	0.59625	2.88012	0.0045	1.94172

6 Был построен граф. Каждая вершина графа отображает набор товаров. Ширина ребра должна отображать уровень support, а подпись на ребре отображать confidence.

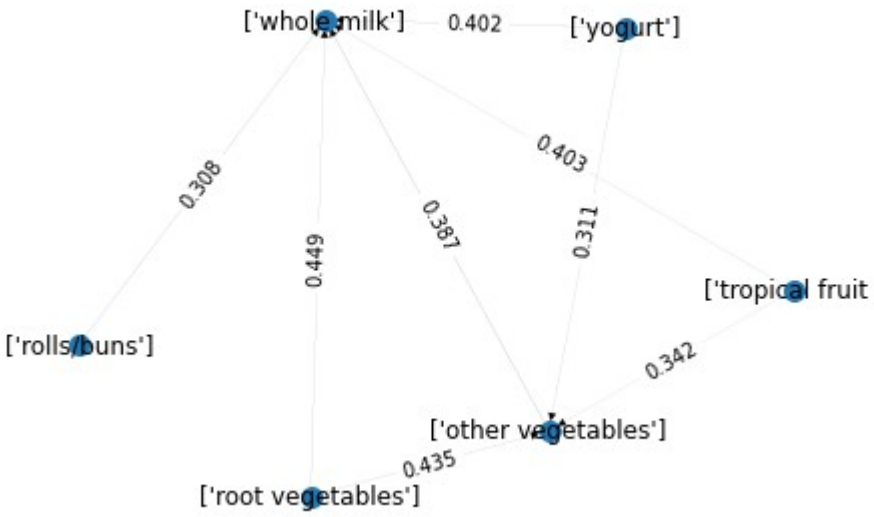

Рисунок 11 - Граф

7 Из графа можно сделать вывод, что при покупке root_vegetables, othher_vegetables tropical_fruit, yogurt с вероятностью примерно 40%, а при покупке rolls/buns — с вероятностью 30%, будет куплено также whole milk. При покупке yogurt и tropical_fruit также присутствует вероятность покупки other_vegetables, а при покупке other_vegetables - root_vegetables.

8 Альтернативный способ представления правил - графический.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(yogurt)	(whole milk)	0.139502	0.255516	0.056024	0.401603	1.571735	0.020379	1.244132
1	(yogurt)	(other vegetables)	0.139502	0.193493	0.043416	0.311224	1.608457	0.016424	1.170929
2	(tropical fruit)	(other vegetables)	0.104931	0.193493	0.035892	0.342054	1.767790	0.015589	1.225796
3	(tropical fruit)	(whole milk)	0.104931	0.255516	0.042298	0.403101	1.577595	0.015486	1.247252
4	(other vegetables)	(whole milk)	0.193493	0.255516	0.074835	0.386758	1.513634	0.025394	1.214013
5	(rolls/buns)	(whole milk)	0.183935	0.255516	0.056634	0.307905	1.205032	0.009636	1.075696
6	(root vegetables)	(other vegetables)	0.108998	0.193493	0.047382	0.434701	2.246605	0.026291	1.426693
7	(root vegetables)	(whole milk)	0.108998	0.255516	0.048907	0.448694	1.756031	0.021056	1.350401

Рисунок 12 - Правила

Выводы

В ходе лабораторной работы изучены методы ассоциативного анализа из библиотеки MLxtend: алгоритмы FPGrowth и FPMax позволяют выделить

часто встречающиеся наборы элементов для заданного минимального уровня поддержки. Различие данных алгоритмов заключается в том, что наборы в FPMax не могут быть частью других наборов большей длины. Ассоциативные правила можно генерировать с помощью алгоритма association_rules, который принимает на вход метрику и ее минимальное значение для расчета.