

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

ОТЧЕТ
по лабораторной работе №3
по дисциплине «Машинное обучение»

Студентка гр. 6307
Преподаватель

Кичерова А. Д.
Жангиров Т. Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами частотного анализа из библиотеки MLxtend.

Ход работы

Загрузка данных

1. Загружен датасет по ссылке. Данные загружены в датафрейм.
2. Получен список всех покупателей и их количество, а также количество и список всех товаров.

```
Количество покупателей (по id): 1139
Количество товаров: 38
```

3. Был сформирован датасет, подходящий для частотного анализа, т.е. все товары одного покупателя были слиты в один список.

Подготовка данных

1. Так как полученный датасет не пригоден для анализа напрямую, нужно закодировать данные так, чтобы их можно было представить в виде матрицы.

	all-purpose	aluminum foil	bagels	beef	butter	cereals	cheeses	coffee/tea	dinner rolls	dishwashing liquid/detergent	...	shampoo	soap	soda	spaghetti sauce	sugar	toilet paper	toi
0	True	True	False	True	True	False	False	False	True	False	...	True	True	True	False	False	False	
1	False	True	False	False	False	True	True	False	False	True	...	True	False	False	False	False	True	
2	False	False	True	False	False	True	True	False	True	False	...	True	True	True	True	False	True	
3	True	False	False	False	False	True	False	False	False	False	...	False	False	True	False	False	True	
4	True	False	False	False	False	False	False	False	True	False	...	False	False	True	True	False	True	
...
1134	True	False	False	True	False	True	True	True	True	True	...	True	True	False	False	True	False	
1135	False	False	False	False	False	True	True	True	True	True	...	False	True	False	True	False	False	
1136	False	False	True	True	False	False	False	False	True	True	...	True	True	False	False	True	False	
1137	True	False	False	True	False	False	True	False	False	False	...	False	True	True	True	True	True	
1138	False	False	False	False	False	False	False	False	False	False	...	True	False	True	False	False	False	

Ассоциативный анализ с использованием алгоритма Apriori

1. Получили наборы товаров, которые покупали не менее чем 30% пользователей.

	support	itemsets
0	0.374890	(all- purpose)
1	0.384548	(aluminum foil)
2	0.385426	(bagels)
3	0.374890	(beef)
4	0.367867	(butter)
5	0.395961	(cereals)
6	0.390694	(cheeses)
7	0.379280	(coffee/tea)
8	0.388938	(dinner rolls)
9	0.388060	(dishwashing liquid/detergent)
10	0.389816	(eggs)

2. Получили наборы товаров, состоящие только из одного элемента

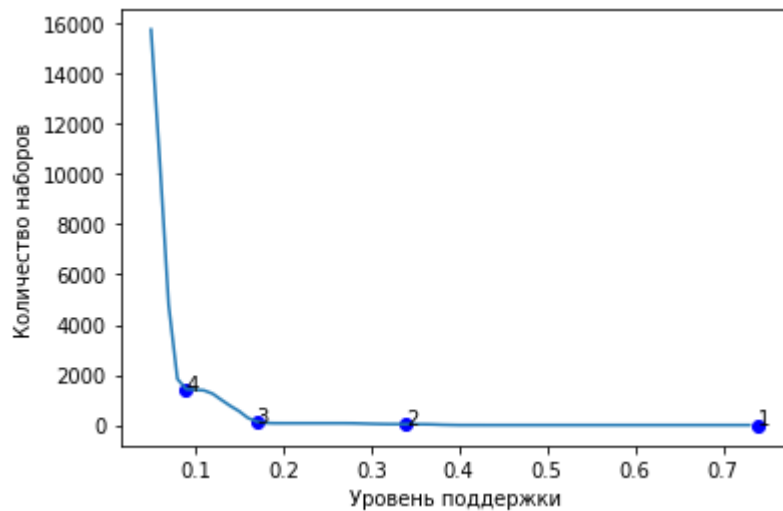
	support	itemsets
0	0.374890	(all- purpose)
1	0.384548	(aluminum foil)
2	0.385426	(bagels)
3	0.374890	(beef)
4	0.367867	(butter)
5	0.395961	(cereals)
6	0.390694	(cheeses)
7	0.379280	(coffee/tea)
8	0.388938	(dinner rolls)
9	0.388060	(dishwashing liquid/detergent)
10	0.389816	(eggs)

3. Посчитали количество и вывели наборы, которые имеют 2 элемента

Количество наборов, которые имеют длину 2: 14

	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	(ice cream, vegetables)	2
46	0.309043	(vegetables, laundry detergent)	2
47	0.311677	(vegetables, lunch meat)	2

4. Посчитали количество наборов при различных уровнях поддержки. А также определили значения уровня поддержки, при котором перестают генерироваться наборы различных размеров.



5. Построили датасет только из тех элементов, которые попадают в наборы с размером 1 при уровне поддержки 0.38. Провели ассоциативный анализ при уровне поддержки 0.3.

support		itemsets
0	0.384548	(aluminum foil)
1	0.385426	(bagels)
2	0.395961	(cereals)
3	0.390694	(cheeses)
4	0.388938	(dinner rolls)
5	0.388060	(dishwashing liquid/detergent)
6	0.389816	(eggs)
7	0.398595	(ice cream)
8	0.395083	(lunch meat)
9	0.380158	(milk)
10	0.421422	(poultry)

Отличие заключается в том, что пропали наборы данных, уровень поддержки которых меньше 0.38. Результат прошлого ассоциативного анализа содержит результат этого.

6. Провели ассоциативный анализ при уровне поддержки 0.15. Вывели все наборы размер, которых больше 1 и в котором есть 'yogurt' или 'waffles'

Количество наборов, подходящих под условие: 30

	support	itemsets
27	0.169447	(aluminum foil, waffles)
28	0.177349	(aluminum foil, yogurt)
40	0.159789	(waffles, bagels)
41	0.162423	(yogurt, bagels)
52	0.160667	(cereals, waffles)
53	0.172081	(yogurt, cereals)
63	0.172959	(waffles, cheeses)
64	0.172081	(yogurt, cheeses)
73	0.169447	(waffles, dinner rolls)
74	0.166813	(yogurt, dinner rolls)
82	0.175593	(waffles, dishwashing liquid/detergent)
83	0.158033	(yogurt, dishwashing liquid/detergent)
90	0.169447	(eggs, waffles)
91	0.174715	(yogurt, eggs)
97	0.172959	(ice cream, waffles)
98	0.156277	(yogurt, ice cream)

7. Построили датасет из элементов, которые не попали в прошлый датасет. И провели ассоциативный анализ.

	support	itemsets
0	0.374890	(all- purpose)
1	0.374890	(beef)
2	0.367867	(butter)
3	0.379280	(coffee/tea)
4	0.352941	(flour)
5	0.370500	(fruits)
6	0.345917	(hand soap)
7	0.375768	(individual meals)
8	0.376646	(juice)
9	0.371378	(ketchup)
10	0.378402	(laundry detergent)
11	0.375768	(mixes)
12	0.362599	(paper towels)
13	0.371378	(pasta)
14	0.355575	(pork)
15	0.367867	(sandwich bags)
16	0.349429	(sandwich loaves)

8. Правило, для вывода всех наборов, в которых хотя бы два элемента начинаются на 's'

	support	itemsets
675	0.137840	(sandwich loaves, sandwich bags)
676	0.146620	(sandwich bags, shampoo)
677	0.158911	(sandwich bags, soap)
678	0.162423	(sandwich bags, soda)
679	0.147498	(sandwich bags, spaghetti sauce)
...
15722	0.064091	(yogurt, vegetables, sugar, soda)
15729	0.058824	(spaghetti sauce, toilet paper, vegetables, su...
15730	0.050044	(spaghetti sauce, vegetables, tortillas, sugar)
15731	0.057946	(spaghetti sauce, vegetables, waffles, sugar)
15732	0.061457	(yogurt, spaghetti sauce, vegetables, sugar)

1275 rows × 2 columns

9. Правило для вывода всех наборов, для которых уровень поддержки изменяется от 0.1 до 0.25

	support	itemsets
38	0.157155	(aluminum foil, all- purpose)
39	0.150132	(all- purpose, bagels)
40	0.144864	(all- purpose, beef)
41	0.147498	(all- purpose, butter)
42	0.151010	(all- purpose, cereals)
...
1401	0.135206	(waffles, toilet paper, vegetables)
1402	0.130817	(yogurt, toilet paper, vegetables)
1403	0.121159	(waffles, vegetables, tortillas)
1404	0.130817	(yogurt, vegetables, tortillas)
1405	0.146620	(yogurt, vegetables, waffles)

Вывод

В результате работы были получены практические навыки по применению методов частотного анализа. В ходе работы был изучен алгоритм Apriori – алгоритм поиска ассоциативных правил из библиотеки MLxtend.