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»

Тема: Частотный анализ

Студент гр. 6307 _____ Трофимов Н.И.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами частотного анализа из библиотеки MLxtend.

Ход работы

1. Загрузка данных.

Был загружен датасет, данные в котором представляют информацию о том, какой покупатель что и когда покупал. Сформируем датасет подходящий для частотного анализа, слив все товары одного чека в список.

2. Подготовка данных.

Закодируем данные в виде матрицы с помощью TransactionEncoder. Получим датафрейм, в котором значение в строке i, столбце j означает, была ли сделана покупка в чеке с id i товара с именованием j. Далее приведена часть датафрейма на рисунке 1:

	all-purpose	aluminum foil	bagels	beef	butter	cereals	cheeses	coffee/tea	dinner rolls	dishwashing liquid/detergent
0	True	True	False	True	True	False	False	False	True	False
1	False	True	False	False	False	True	True	False	False	True
2	False	False	True	False	False	True	True	False	True	False
3	True	False	False	False	False	True	False	False	False	False
4	True	False	False	False	False	False	False	False	True	False

Рисунок 1 Готовый датафрейм

3. Ассоциативный анализ с использованием алгоритма Apriori

Применим алгоритм apriori к подготовленному датафрейму с минимальным уровнем поддержки 0.3. Получим все комбинации товаров, которые встречаются в 30 процентах покупок подготовленного датафрейма. Результаты представлены далее на рисунке 2.

support	itemsets	length
0.374890	(all- purpose)	1
0.384548	(aluminum foil)	1
0.385426	(bagels)	1
0.374890	(beef)	1
0.367867	(butter)	1
0.395961	(cereals)	1
0.390694	(cheeses)	1
0.379280	(coffee/tea)	1
0.388938	(dinner rolls)	1
0.388060	(dishwashing liquid/detergent)	1
0.389816	(eggs)	1
0.352941	(flour)	1
0.370500	(fruits)	1
0.345917	(hand soap)	1
0.398595	(ice cream)	1
0.375768	(individual meals)	1
0.376646	(juice)	1
0.371378	(ketchup)	1
0.378402	(laundry detergent)	1
0.395083	(lunch meat)	1
0.380158	(milk)	1
0.375768	(mixes)	1
0.362599	(paper towels)	1
0.371378	(pasta)	1
0.355575	(pork)	1
0.421422	(poultry)	1
0.367867	(sandwich bags)	1
0.349429	(sandwich loaves)	1
0.368745	(shampoo)	1
0.379280	(soap)	1
0.390694	(soda)	1
0.373134	(spaghetti sauce)	1
0.360843	(sugar)	1
0.378402	(toilet paper)	1
0.369622	(tortillas)	1
0.739245	(vegetables)	1
0.394205	(waffles)	1
0.384548	(yogurt)	1
0.310799	(aluminum foil, vegetables)	2
0.300263	(bagels, vegetables)	2
0.310799	(cereals, vegetables)	2
0.309043	(cheeses, vegetables)	2
0.308165	(vegetables, dinner rolls)	2
0.306409	(vegetables, dishwashing liquid/detergent)	2
0.326602	(eggs, vegetables)	2
0.302897	(vegetables, ice cream)	2
0.309043	(laundry detergent, vegetables)	2
0.311677	(lunch meat, vegetables)	2
0.331870	(vegetables, poultry)	2
0.305531	(soda, vegetables)	2
0.315189	(waffles, vegetables)	2
0.319579	(vegetables, yogurt)	2

Рисунок 2 примененный алгоритм apriori

Затем был применен дважды тот же алгоритм с тем же уровнем поддержки, однако было выставлено значение для максимального размера, равного 1, а потом 2.

Построим график зависимости количества полученных наборов от уровня поддержки. Начнем с уровня поддержки = 0.05, шаг будет равен 0.01. Отметим на графике значения уровней поддержки, при которых перестают генерироваться наборы размеров 1,2,3 и т.д. Результаты представлены на рисунке 3.

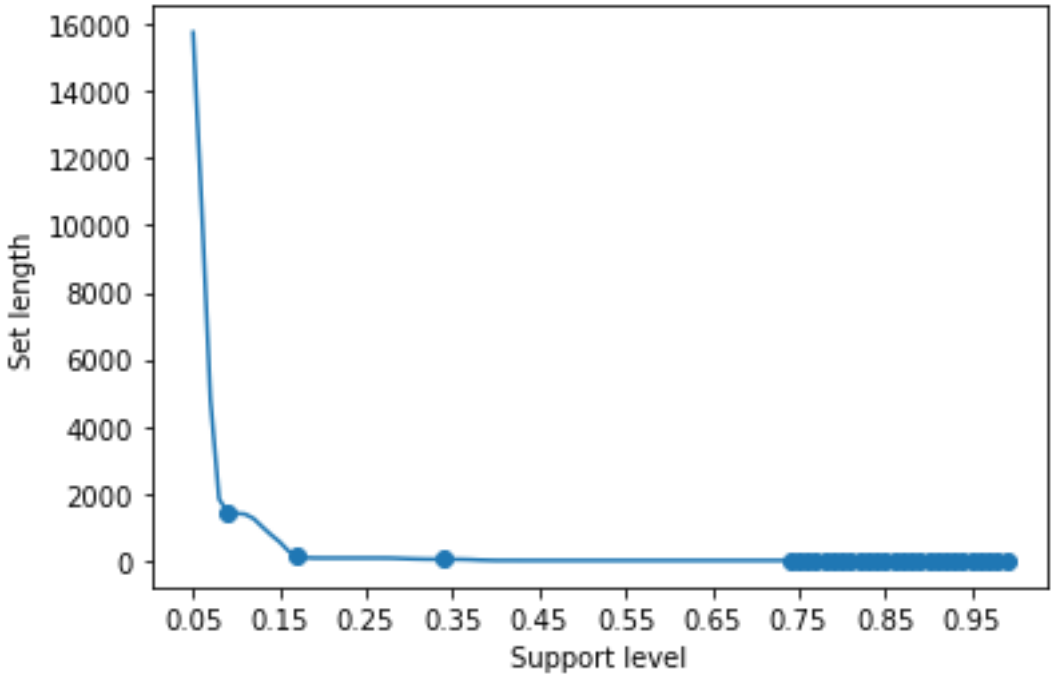

Рисунок 3. Зависимость количества наборов от уровня поддержки

Как можно заметить, начиная с уровня поддержки = 0.74 наборы перестают генерироваться.

Был построен датасет только из тех элементов, которые попадают в наборы размером 1 при уровне поддержки 0.38. Затем полученный датасет был приведен к формату, который можно обработать и проведен ассоциативный анализ при уровне поддержки 0.3. Отличием от исходного датасета является то, что в новом появились наборы состоящие из 1 элемента с минимальным уровнем поддержки = 0.38 (См рис.4).

	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	(bagels, vegetables)	2
17	0.310799	(cereals, vegetables)	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	(eggs, vegetables)	2
22	0.302897	(vegetables, ice cream)	2
23	0.311677	(lunch meat, vegetables)	2
24	0.331870	(vegetables, poultry)	2
25	0.305531	(soda, vegetables)	2
26	0.315189	(waffles, vegetables)	2
27	0.319579	(vegetables, yogurt)	2

	support	itemsets	length
15	0.310799	(aluminum foil, vegetables)	2
16	0.300263	(bagels, vegetables)	2
17	0.310799	(cereals, vegetables)	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	(eggs, vegetables)	2
22	0.302897	(vegetables, ice cream)	2
28	0.309043	(laundry detergent, vegetables)	2
23	0.311677	(lunch meat, vegetables)	2
24	0.331870	(vegetables, poultry)	2
25	0.305531	(soda, vegetables)	2
26	0.315189	(waffles, vegetables)	2
27	0.319579	(vegetables, yogurt)	2

Рисунок 4 Уровень поддержки 0.38 и 0.3 сравнение

Далее был проведен ассоциативный анализ при уровне поддержки 0.15 для нового датасета и выведены все наборы, размер которых больше 1 и в котором есть “yogurt” или “waffles”. Результат представлен на рисунке 5.

	support	itemsets	length	contains
27	0.169447	(aluminum foil, waffles)	2	True
28	0.177349	(aluminum foil, yogurt)	2	True
40	0.159789	(bagels, waffles)	2	True
41	0.162423	(bagels, yogurt)	2	True
52	0.160667	(cereals, waffles)	2	True
53	0.172081	(cereals, yogurt)	2	True
63	0.172959	(cheeses, waffles)	2	True
64	0.172081	(cheeses, yogurt)	2	True
73	0.169447	(waffles, dinner rolls)	2	True
74	0.166813	(yogurt, dinner rolls)	2	True
82	0.175593	(waffles, dishwashing liquid/detergent)	2	True
83	0.158033	(yogurt, dishwashing liquid/detergent)	2	True
90	0.169447	(eggs, waffles)	2	True
91	0.174715	(eggs, yogurt)	2	True
97	0.172959	(waffles, ice cream)	2	True
98	0.156277	(yogurt, ice cream)	2	True
103	0.184372	(lunch meat, waffles)	2	True
104	0.161545	(lunch meat, yogurt)	2	True
108	0.167691	(yogurt, milk)	2	True
111	0.166813	(waffles, poultry)	2	True
112	0.180860	(yogurt, poultry)	2	True
114	0.177349	(soda, waffles)	2	True
115	0.167691	(soda, yogurt)	2	True
116	0.315189	(waffles, vegetables)	2	True
117	0.319579	(vegetables, yogurt)	2	True
118	0.173837	(waffles, yogurt)	2	True
119	0.152766	(aluminum foil, vegetables, yogurt)	3	True
128	0.157155	(eggs, vegetables, yogurt)	3	True
130	0.157155	(lunch meat, waffles, vegetables)	3	True
131	0.152766	(yogurt, vegetables, poultry)	3	True

Рисунок 5. кол-во элементов больше 1 и есть товар

Далее построен датасет из тех элементов, которые не попали в датасет в п. 6. Полученный датасет закодирован в виде матрицы и проведен ассоциативный анализ при минимальном уровне поддержки 0.3. Результаты представлены на рисунке 6.

	support	itemsets	length
0	0.374890	(all- purpose)	1
1	0.374890	(beef)	1
2	0.367867	(butter)	1
3	0.379280	(coffee/tea)	1
4	0.352941	(flour)	1
5	0.370500	(fruits)	1
6	0.345917	(hand soap)	1
7	0.375768	(individual meals)	1
8	0.376646	(juice)	1
9	0.371378	(ketchup)	1
10	0.378402	(laundry detergent)	1
11	0.375768	(mixes)	1
12	0.362599	(paper towels)	1
13	0.371378	(pasta)	1
14	0.355575	(pork)	1
15	0.367867	(sandwich bags)	1
16	0.349429	(sandwich loaves)	1
17	0.368745	(shampoo)	1
18	0.379280	(soap)	1
19	0.373134	(spaghetti sauce)	1
20	0.360843	(sugar)	1
21	0.378402	(toilet paper)	1
22	0.369622	(tortillas)	1

Рисунок 6 Новые данные

Были написаны два правила:

1. Вывод всех наборов, в которых хотя бы два элемента начинаются на “s”.
2. Вывод всех наборов, для которых уровень поддержки изменяется от 0.1 до 0.25.

Результаты представлены на рисунках 7 и 8.

```
results['rule_with_s'] = results['itemsets']\
    .apply(lambda tpl: len([name for name in tpl if name.startswith('s')]) >= 2)
results[results['rule_with_s'] == True]
```

	support	itemsets	length	rule_with_s
675	0.137840	(sandwich bags, sandwich loaves)	2	True
676	0.146620	(sandwich bags, shampoo)	2	True
677	0.158911	(sandwich bags, soap)	2	True
678	0.162423	(sandwich bags, soda)	2	True
679	0.147498	(sandwich bags, spaghetti sauce)	2	True
680	0.131694	(sandwich bags, sugar)	2	True
686	0.150132	(shampoo, sandwich loaves)	2	True
687	0.158033	(soap, sandwich loaves)	2	True
688	0.141352	(soda, sandwich loaves)	2	True
689	0.150132	(sandwich loaves, spaghetti sauce)	2	True
690	0.136962	(sugar, sandwich loaves)	2	True
696	0.151010	(soap, shampoo)	2	True
697	0.150132	(soda, shampoo)	2	True
698	0.139596	(shampoo, spaghetti sauce)	2	True
699	0.147498	(sugar, shampoo)	2	True
705	0.174715	(soda, soap)	2	True
706	0.160667	(soap, spaghetti sauce)	2	True
707	0.154522	(sugar, soap)	2	True
713	0.167691	(soda, spaghetti sauce)	2	True
714	0.162423	(soda, sugar)	2	True
720	0.144864	(sugar, spaghetti sauce)	2	True
1351	0.115013	(sandwich bags, vegetables, sandwich loaves)	3	True
1352	0.122915	(sandwich bags, shampoo, vegetables)	3	True
1353	0.129939	(sandwich bags, soap, vegetables)	3	True
1354	0.129061	(sandwich bags, soda, vegetables)	3	True

Рисунок 7 Правило №1

```
results.query("support >= .1 and support <=.25 ")
```

	support	itemsets	length	rule_with_s
38	0.157155	(aluminum foil, all- purpose)	2	False
39	0.150132	(bagels, all- purpose)	2	False
40	0.144864	(beef, all- purpose)	2	False
41	0.147498	(butter, all- purpose)	2	False
42	0.151010	(cereals, all- purpose)	2	False
...	...	...	...	...
1401	0.135206	(waffles, vegetables, toilet paper)	3	False
1402	0.130817	(yogurt, vegetables, toilet paper)	3	False
1403	0.121159	(tortillas, waffles, vegetables)	3	False
1404	0.130817	(tortillas, vegetables, yogurt)	3	False
1405	0.146620	(waffles, vegetables, yogurt)	3	False

Рисунок 8 Правило №2

Вывод

В ходе работы были изучены методы частотного анализа с использованием библиотеки mlxtend. Основным используемым алгоритмом в данной работе был apriori. Его основная задача – поиск наиболее часто встречающихся наборов значений. Уровень поддержки в этом алгоритме определяет частоту встречаемости набора данных для попадания в итоговую выборку. Чем больше данный параметр, тем меньше наборов будет сгенерировано.