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

ОТЧЕТ

по лабораторной работе №3

по дисциплине «Машинное обучение»

Тема: Частотный анализ

Студент гр. 6307 _____ Мишанов А. А.

Преподаватель _____ 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.

	date	id	product
0	2000-01-01	1	yogurt
1	2000-01-01	1	pork
2	2000-01-01	1	sandwich bags
3	2000-01-01	1	lunch meat
4	2000-01-01	1	all- purpose
...	...	...	...
22338	2002-02-26	1139	soda
22339	2002-02-26	1139	laundry detergent
22340	2002-02-26	1139	vegetables
22341	2002-02-26	1139	shampoo
22342	2002-02-26	1139	vegetables

22343 rows × 3 columns

Рисунок 1. Загруженный датасет

2. Получен список всех id покупателей.
3. Получен список всех товаров.
4. Был сформирован датасет, подходящий для частотного анализа.

Подготовка данных

1. Кодирование данных с использованием TransactionEncoder.

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

1139 rows × 38 columns

Рисунок 2. Закодированный датафрейм

2. Данные теперь представлены в бинарном виде: наличие или отсутствие каждого товара пользователя представлено либо True, либо False.

Ассоциативный анализ с использованием алгоритма Apriori

1. Был применен алгоритм apriori с минимальным уровнем поддержки 0.3. В результате были представлены товары, которые встречаются не реже, чем в 0,3 покупок.

	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
9	0.388060	(dishwashing liquid/detergent)	1

Рисунок 3. Первые 10 элементов датафрейма после применения алгоритма apriori

2. Алгоритм apriori с тем же уровнем поддержки, но с ограничением максимального размера набора единиц.

	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)

Рисунок 4. Первые 10 элементов с ограничением максимального размера набора единиц

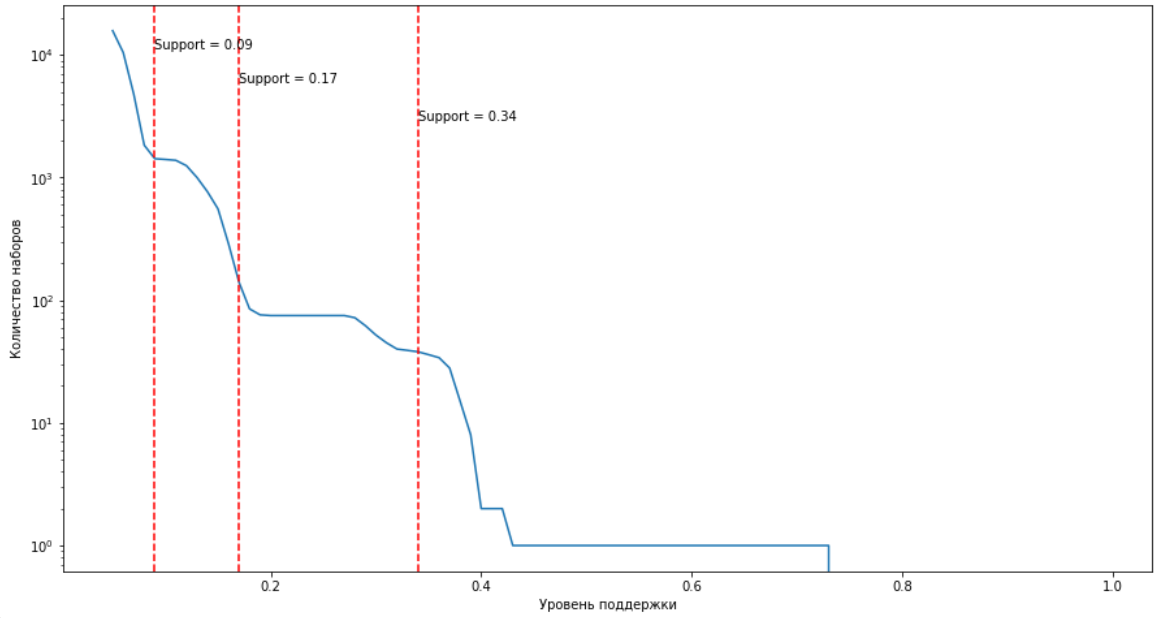
3. Применим алгоритм apriori и выведем только те наборы, которые имеют размер 2, а также количество таких наборов.

	support	itemsets	length
38	0.310799	(vegetables, aluminum foil)	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	(vegetables, eggs)	2
45	0.302897	(ice cream, vegetables)	2
46	0.309043	(vegetables, laundry detergent)	2
47	0.311677	(vegetables, lunch meat)	2
48	0.331870	(poultry, vegetables)	2
49	0.305531	(vegetables, soda)	2
50	0.315189	(waffles, vegetables)	2
51	0.319579	(vegetables, yogurt)	2

Count of result itemstes = 14

Рисунок 5. Наборы размера 2, а также их количество

4. Построим график зависимости количества наборов от уровня поддержки и отметим полученные уровне поддержки на графике.



.

Рисунок 6. Полученный график

5. Был построен датасет только из тех элементов, которые попадают в наборы размером 1 при уровне поддержки 0.38. Полученный датасет был закодирован. Далее к нему был применен алгоритм apriori при уровне поддержки 0.3.

	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

Рисунок 7. Алгоритм apriori при уровне поддержки 0.3. (Первые 10 элементов)

6. Различия между старым и новым датасетом в том, что в новом датасете присутствуют наборы длины 2 с минимальным уровнем поддержки 0.3.
7. Проведем ассоциативный анализ при уровне поддержки 0.15 для нового датасета. Выведем все наборы, размер которых больше 1 и в которых есть 'yogurt' или 'waffles'.

	support	itemsets	length
27	0.169447	(waffles, aluminum foil)	2
28	0.177349	(aluminum foil, yogurt)	2
40	0.159789	(waffles, bagels)	2
41	0.162423	(yogurt, bagels)	2
52	0.160667	(cereals, waffles)	2
53	0.172081	(cereals, yogurt)	2
63	0.172959	(waffles, cheeses)	2
64	0.172081	(yogurt, cheeses)	2
73	0.169447	(waffles, dinner rolls)	2
74	0.166813	(yogurt, dinner rolls)	2

Рисунок 8. Алгоритм apriori при уровне поддержки 0.15 с наборами, размер которых больше 1 и в которых есть 'yogurt' или 'waffles'. (Первые 10 элементов)

8. Построим датасет из тех элементов, которые не попали в датасет в п. 6 и приведем к удобному для анализа виду, а также проведем анализ apriori для полученного датасета.

	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)
17	0.368745	(shampoo)
18	0.379280	(soap)
19	0.373134	(spaghetti sauce)
20	0.360843	(sugar)
21	0.378402	(toilet paper)
22	0.369622	(tortillas)

Рисунок 9. Алгоритм apriori при уровне поддержки 0.3 с наборами, которые не попали в датасет в п. 6.

9. Напишем 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	(shampoo, sandwich bags)
677	0.158911	(soap, sandwich bags)
678	0.162423	(soda, sandwich bags)
679	0.147498	(spaghetti sauce, sandwich bags)
680	0.131694	(sugar, sandwich bags)
686	0.150132	(shampoo, sandwich loaves)
687	0.158033	(sandwich loaves, soap)
688	0.141352	(soda, sandwich loaves)
689	0.150132	(sandwich loaves, spaghetti sauce)

Рисунок 10. Правило для вывода всех наборов, в которых хотя бы два элемента начинаются на 's'. (Первые 10 элементов).

10. Напишем 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	(beef, all- purpose)
41	0.147498	(butter, all- purpose)
42	0.151010	(cereals, all- purpose)
...	...	...
1401	0.135206	(waffles, vegetables, toilet paper)
1402	0.130817	(vegetables, yogurt, toilet paper)
1403	0.121159	(waffles, vegetables, tortillas)
1404	0.130817	(vegetables, yogurt, tortillas)
1405	0.146620	(waffles, vegetables, yogurt)

Рисунок 11. Правило для вывода всех наборов, для которых уровень поддержки изменяется от 0.1 до 0.25.

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