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 _____ Ходос А.А.

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:

	all- purpose	*aluminum foil*	*bagels*	*beef*	*butter*	*cereals*	*cheeses*	
0	*True*	*False*	*True*	*True*	*False*	*False*	*True*	...
1	*False*	*False*	*False*	*False*	*True*	*False*	*False*	...
2	*False*	*True*	*False*	*False*	*True*	*False*	*True*	...
3	*True*	*False*	*False*	*False*	*True*	*False*	*False*	...
4	*True*	*False*	*False*	*False*	*False*	*False*	*True*	...
...	...	...	...	...	...	...	...	...

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:

	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*
…	*…*	*…*	
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*	*(dinner rolls, vegetables)*	*2*
…	*…*	*…*	

Применим алгоритм apriori с тем же уровнем поддержки, но ограничив максимальный размер набора единицей. Результаты представлены далее:

	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)*
…	*…*	*…*

Далее, применив алгоритм apriori, выведем только те наборы, которые имеют размер 2. Количество полученных наборов равно 14. Результаты представлены далее:

	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*	*(dinner rolls, vegetables)*	*2*
	…	*…*	*…*

Построим график зависимости количества полученных наборов от уровня поддержки. Отметим на графике значения уровней поддержки, при которых перестают генерироваться наборы размеров 1,2,3 и т.д. Результаты представлены на рисунке 1.

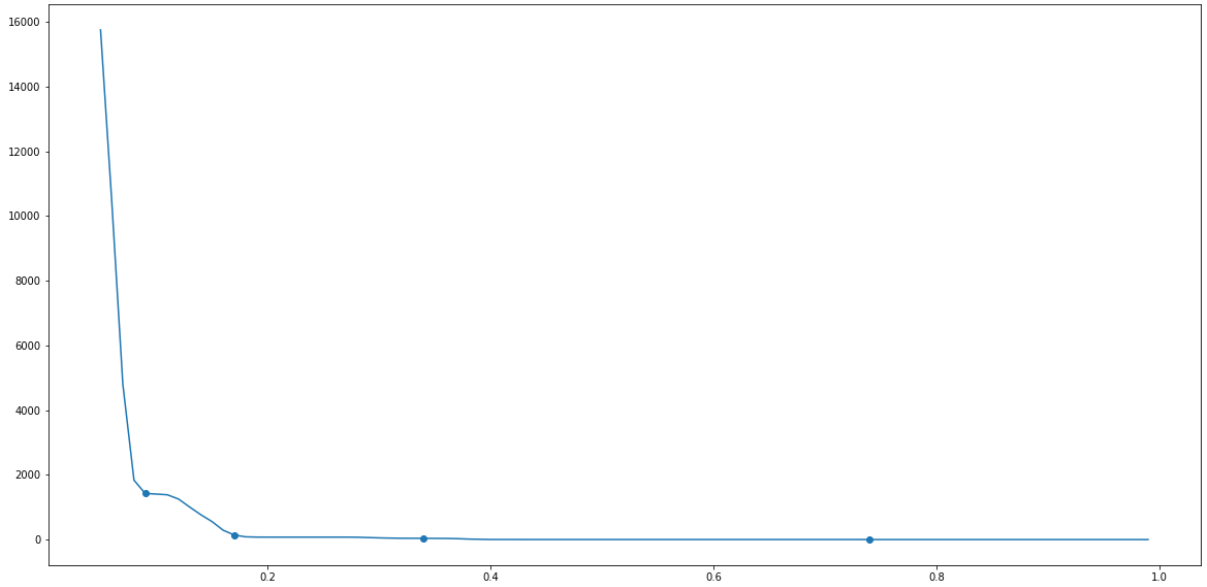

Рисунок 1 - График зависимости количества полученных наборов от уровня поддержки

Построим датасет, состоящих только из тех элементов, которые попадают в наборы размером 1 при уровне поддержки 0.38. Полученный датасет закодируем в виде матрицы и проведем ассоциативный анализ при минимальном уровне поддержки 0.3. Таким образом, новый набор данных будет являться подмножеством старого, исключив наборы включающие продукты, которые встречаются в менее 38 процентах покупок.

	support	*itemsets*	*length*
0	*0.384548*	*(aluminum foil)*	*1*
1	*0.385426*	*(bagels)*	*1*
2	*0.395961*	*(cereals)*	*1*
3	*0.390694*	*(cheeses)*	*1*
…	*…*	*…*	
24	*0.331870*	*(poultry, vegetables)*	*2*
25	*0.305531*	*(vegetables, soda)*	*2*
26	*0.315189*	*(vegetables, waffles)*	*2*
27	*0.319579*	*(vegetables, yogurt)*	*2*

Проведем ассоциативный анализ при уровне поддержки 0.15 для нового датасета. Все наборы, размер которых больше 1 и в котором есть 'yogurt' или 'waffles', представлены далее:

	support	*itemsets*	*length*	*y or w*
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*	*(yogurt, bagels)*	*2*	*True*
52	*0.160667*	*(cereals, waffles)*	*2*	*True*
...	*...*	*...*	*...*	*...*
119	*0.152766*	*(aluminum foil, yogurt, vegetables)*	*3*	*True*
128	*0.157155*	*(vegetables, yogurt, eggs)*	*3*	*True*
130	*0.157155*	*(vegetables, lunch meat, waffles)*	*3*	*True*
131	*0.152766*	*(poultry, yogurt, vegetables)*	*3*	*True*

Далее построим датасет из тех элементов, которые не попали в датасет в п. 6. Полученный датасет закодируем в виде матрицы и проведем ассоциативный анализ при минимальном уровне поддержки 0.3. Результаты представлены далее:

	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*
...	*...*	*...*	*...*
20	*0.360843*	*(sugar)*	*1*
21	*0.378402*	*(toilet paper)*	*1*
22	*0.369622*	*(tortillas)*	*1*

Написав правила вывода, получим все наборы, содержащие хотя бы два элемента, начинающихся с ‘s’, и все наборы, для которых уровень поддержки изменяется от 0.1 до 0.25. Результаты представлены далее:

	support	*itemsets*	*two_s*
675	*0.137840*	*(sandwich loaves, sandwich bags)*	*True*
676	*0.146620*	*(shampoo, sandwich bags)*	*True*

677	*0.158911*	*(sandwich bags, soap)*	*True*
678	*0.162423*	*(sandwich bags, soda)*	*True*
679	*0.147498*	*(sandwich bags, spaghetti sauce)*	*True*
...	*...*	*...*	*...*

	support	*itemsets*	*in_supp_range*
38	*0.157155*	*(aluminum foil, all- purpose)*	*True*
39	*0.150132*	*(all- purpose, bagels)*	*True*
40	*0.144864*	*(all- purpose, beef)*	*True*
41	*0.147498*	*(all- purpose, butter)*	*True*
42	*0.151010*	*(all- purpose, cereals)*	*True*
...	*...*	*...*	*...*

Вывод

Получены навыки работы с методами частотного анализа из библиотеки MLxtend. Изучен алгоритм поиска ассоциативных правил apriori, который перебирает все правила, генерирующиеся на основе входящих в транзакции наборов, удовлетворяющих заданному уровню поддержки.