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

ОТЧЕТ

по лабораторной работе №4

по дисциплине «Машинное обучение»

Тема: Ассоциативный анализ

Студент гр. 6304 Ковынев М.В.

Преподаватель Жангиров Т.Р.

Санкт-Петербург
2020

Цель

Ознакомиться с методами ассоциативного анализа из библиотеки MLxtend

Ход работы

1. Загрузжен датасет по ссылке: https://www.kaggle.com/irfanasrullah/groceries. Данные представлены в виде csv таблицы. Данные представляют собой информацию о купленных вместе товарах.
2. Создан Python скрипт. Загружены данные в датафрейм

```
"C:\Program Files (x86)\Microsoft Visual Studio\Shared\Python37_64\python.exe" "D:/Documents/Google Drive/works/9_sem/mo/lb4/lb4.py"
     Item(s)             Item 1              Item 2  ... Item 30 Item 31 Item 32
0          4        citrus fruit  semi-finished bread  ...     NaN     NaN     NaN
1          3      tropical fruit              yogurt  ...     NaN     NaN     NaN
2          1          whole milk                 NaN  ...     NaN     NaN     NaN
3          4           pip fruit              yogurt  ...     NaN     NaN     NaN
4          4    other vegetables          whole milk  ...     NaN     NaN     NaN
...      ...                 ...                 ...  ...     ...     ...     ...
9830      17           sausage             chicken  ...     NaN     NaN     NaN
9831       1   cooking chocolate                 NaN  ...     NaN     NaN     NaN
9832      10             chicken        citrus fruit  ...     NaN     NaN     NaN
9833       4  semi-finished bread       bottled water  ...     NaN     NaN     NaN
9834       5             chicken       tropical fruit  ...     NaN     NaN     NaN

[9835 rows x 33 columns]
```

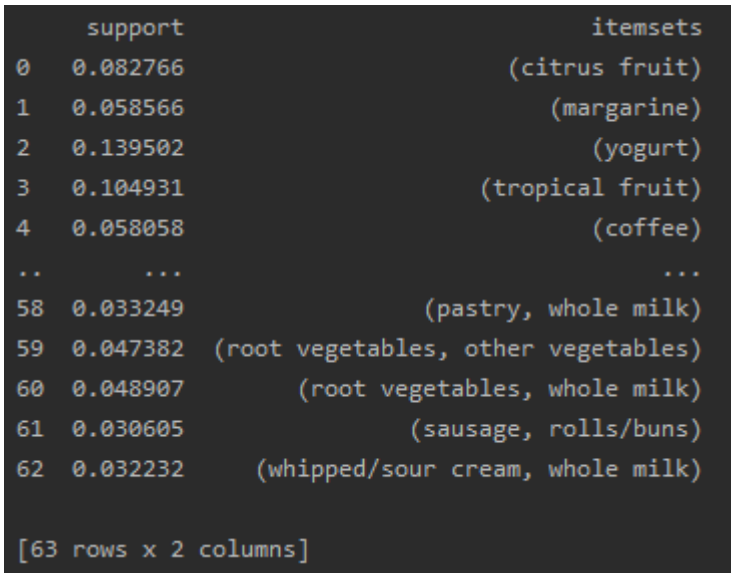
Рисунок 1 — Исходные данные

3. Переформированы данные удалив все значения NaN
4. Получим список всех уникальных товаров
5. Выведен список товаров, а также их количество

```
169 {'dishes', 'curd', 'hamburger meat', 'soft cheese', 'zwieback', 'jam', 'root vegetables', 'coffee', 'butter',
 'specialty vegetables', 'ketchup', 'frankfurter', 'fish', 'rubbing alcohol', 'pastry', 'newspapers', 'liquor',
 'flour', 'citrus fruit', 'potato products', 'UHT-milk', 'cat food', 'flower soil/fertilizer', 'other vegetables',
 'red/blush wine', 'sauces', 'soap', 'sweet spreads', 'potted plants', 'liver loaf', 'vinegar', 'brown bread', 'dish
cleaner', 'cream', 'bathroom cleaner', 'female sanitary products', 'chocolate marshmallow', 'cake bar', 'yogurt',
 'pork', 'detergent', 'finished products', 'semi-finished bread', 'male cosmetics', 'instant coffee', 'skin care',
 'whipped/sour cream', 'chocolate', 'nuts/prunes', 'dog food', 'hard cheese', 'herbs', 'candy', 'baking powder',
 'frozen potato products', 'honey', 'spread cheese', 'sliced cheese', 'salty snack', 'condensed milk', 'frozen
meals', 'grapes', 'syrup', 'flower (seeds)', 'salt', 'ice cream', 'organic products', 'liqueur', 'whole milk',
 'artif. sweetener', 'turkey', 'pet care', 'abrasive cleaner', 'pudding powder', 'oil', 'candles', 'cooking
chocolate', 'napkins', 'chewing gum', 'bottled beer', 'sugar', 'margarine', 'specialty chocolate', 'rum', 'frozen
dessert', 'tea', 'long life bakery product', 'rolls/buns', 'spices', 'domestic eggs', 'butter milk', 'canned fruit',
 'kitchen utensil', 'frozen fish', 'mayonnaise', 'popcorn', 'sound storage medium', 'frozen fruits', 'dental care',
 'specialty cheese', 'mustard', 'decalcifier', 'meat spreads', 'soups', 'soda', 'specialty bar', 'brandy', 'house
keeping products', 'preservation products', 'softener', 'liquor (appetizer)', 'ham', 'snack products', 'kitchen
towels', 'organic sausage', 'Instant food products', 'meat', 'tidbits', 'processed cheese', 'hygiene articles',
 'bottled water', 'shopping bags', 'roll products', 'canned vegetables', 'beverages', 'sparkling wine', 'misc.
beverages', 'light bulbs', 'tropical fruit', 'seasonal products', 'onions', 'chicken', 'packaged fruit/vegetables',
 'berries', 'photo/film', 'cookware', 'cling film/bags', 'baby food', 'toilet cleaner', 'pip fruit', 'dessert', 'make
 up remover', 'salad dressing', 'frozen chicken', 'white bread', 'cereals', 'cream cheese', 'white wine', 'beef',
 'cleaner', 'cocoa drinks', 'hair spray', 'prosecco', 'whisky', 'canned beer', 'rice', 'nut snack', 'baby cosmetics',
 'waffles', 'pickled vegetables', 'bags', 'sausage', 'frozen vegetables', 'canned fish', 'specialty fat',
 'fruit/vegetable juice', 'curd cheese', 'ready soups', 'pasta'}
```

Рисунок 2 — Уникальные товары

6. Преобразуем данные к виду, удобному для анализа
7. Проведем 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	(pastry, whole milk)
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 x 2 columns]

Рисунок 3 — FPGrowth при уровне поддержки 0.03

8. Проанализированы получившиеся варианты. Определите минимальное и максимальное значения для уровня поддержки для набора из 1,2, и.т.д. объектов.
9. Проведен аналогичный анализ используя алгоритм FPMax

Длина набора	FPGrowth	FPMax
1	[0.0304,0.2555]	[0.0304, 0.0985]
2	[0.0300, 0.0748]	[0.0300, 0.0748]

Из таблицы видно, что отличается только максимальный уровень поддержки для длины набора 1. Это происходит потому, что FPMax набор не может быть частью другого набора большей длины. Наиболее часто встречающиеся наборы длины 1 вошли в наборы длины 2.

10. Построена гистограмма для каждого товара. Столбцы на гистограмме были упорядочены по уменьшению частоты.

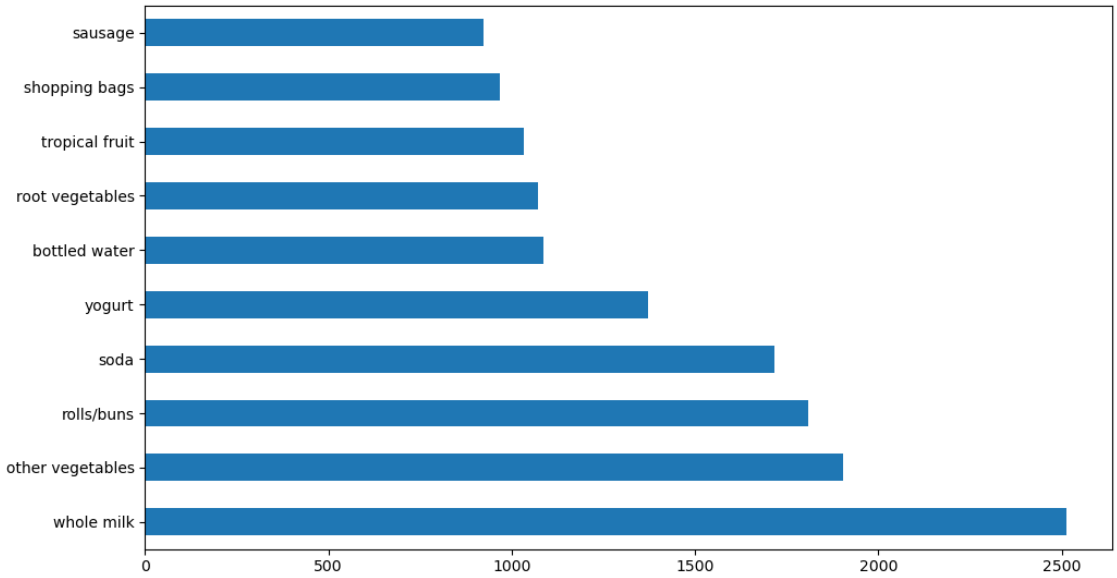


Рисунок 4 — Гистограмма 10 самых часто встречающихся товаров

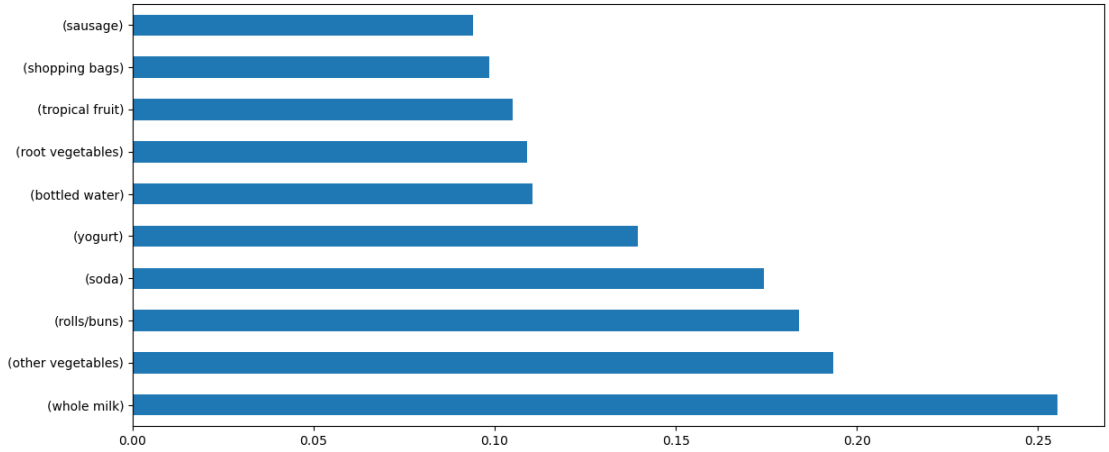


Рисунок 5 — Гистограмма 10 наборов с максимальным уровнем поддержки

11. Преобразуем набор данных, чтобы он содержал ограниченный набор товаров
12. Проведен анализ FPGrowth и FPMax для нового набора данных.

Длина набора	FPGrowth	FPMax
1	[0.0576,0.2555]	[0.0576, 0.0985]
2	[0.0305, 0.0748]	[0.0305, 0.0748]

Т.к. были удалены товары с минимальным уровнем, то минимальное значение для FPGrowth и FPMax увеличились. Максимальные – без изменений.

13. Построены графики изменения количества получаемых правил от уровня поддержки. На графике отдельно отображены кривые для набора товаров 1, 2, и т.д.

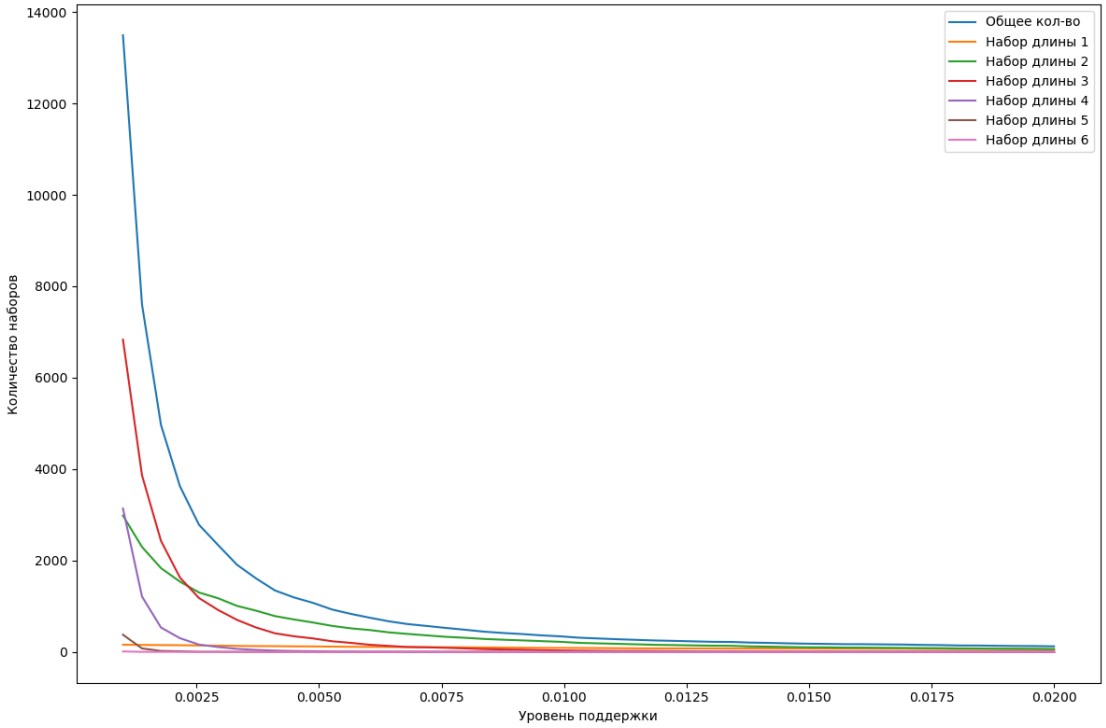


Рисунок 6 — Зависимость уровня поддержки от количества наборов

Количество наборов уменьшается с увеличением уровня минимальной поддержки.

14. Сформируем набор данных из определенных товаров и так, чтобы размер транзакции был 2 и более. Получим частоты наборов используя алгоритм FPGrowth. Проведем ассоциативный анализ

	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	(other vegetables)	(whole milk)	0.193493	0.255516	0.074835	0.386758	1.513634	0.025394	1.214013
2	(rolls/buns)	(whole milk)	0.183935	0.255516	0.056634	0.307905	1.205032	0.009636	1.075696

Поддержка (Support)

Поддержка правила определяется как количество транзакций, содержащих как X, так и Y, то есть

$\sup(X \longrightarrow Y) = sup(XY) = |t(XY)|$

Достоверность (Confidence)

Достоверность правила — это условная вероятность того, что транзакция содержит консеквент Y, при условии, что он содержит антецедент X:

$$conf(X \longrightarrow Y) = P(Y \mid X) = \frac{P(XY)}{P(X)} = \frac{rsup(XY)}{rsup(X)} = \frac{\sup(XY)}{\sup(X)}$$

Лифт (Lift)

Лифт определяется как отношение наблюдаемой совместной вероятности X и Y к ожидаемой совместной вероятности, если бы они были статистически независимыми, то есть

$$lift(X \longrightarrow Y) = \frac{P(XY)}{P(X) \cdot P(Y)} = \frac{rsup(XY)}{rsup(X) \cdot rsup(Y)} = \frac{conf(X \longrightarrow Y)}{rsup(Y)}$$

Усиление (Leverage, Рычаг)

Усиление измеряет разницу между наблюдаемой и ожидаемой совместной вероятностью XY при условии, что X и Y независимы.

$$\text{leverage } (X \longrightarrow Y) = P(XY) - P(X) \cdot P(Y) = \text{rsup}(XY) - \text{rsup}(X) \cdot \text{rsup}(Y)$$

Убежденность (Conviction)

Убежденность 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, определяемого как

$$conv(X \longrightarrow Y) = \frac{P(X) \cdot P(\neg Y)}{P(X \neg Y)} = \frac{1}{lift(X \longrightarrow \neg Y)}$$

15.Рассчитаны среднее значение, медиану и СКО для каждой из метрик

	Antecedent support	**Consequent support**	**Support**	**Confidence**	**Lieft**	**Leverage**	**Conviction**
Count	10	10	10	10	10	10	10
Mean	0.1079	0.2431	0.0431	0.4006	1.6655	0.0168	1.2665
Std	0.0360	0.0261	0.0142	0.0353	0.2374	0.0061	0.0816
Min	0.0716	0.1934	0.0300	0.3420	1.4423	0.0093	1.1790
25%	0.0843	0.2555	0.0324	0.3769	1.5244	0.0155	1.2169
50%	0.1049	0.2555	0.0390	0.3997	1.5746	0.0155	1.2402

75%	0.1089	0.2555	0.0485	0.4268	1.7688	0.0208	1.3246
max	0.1934	0.2555	0.0748	0.4496	2.2466	0.0262	1.4266

16. Построен граф для следующего анализа

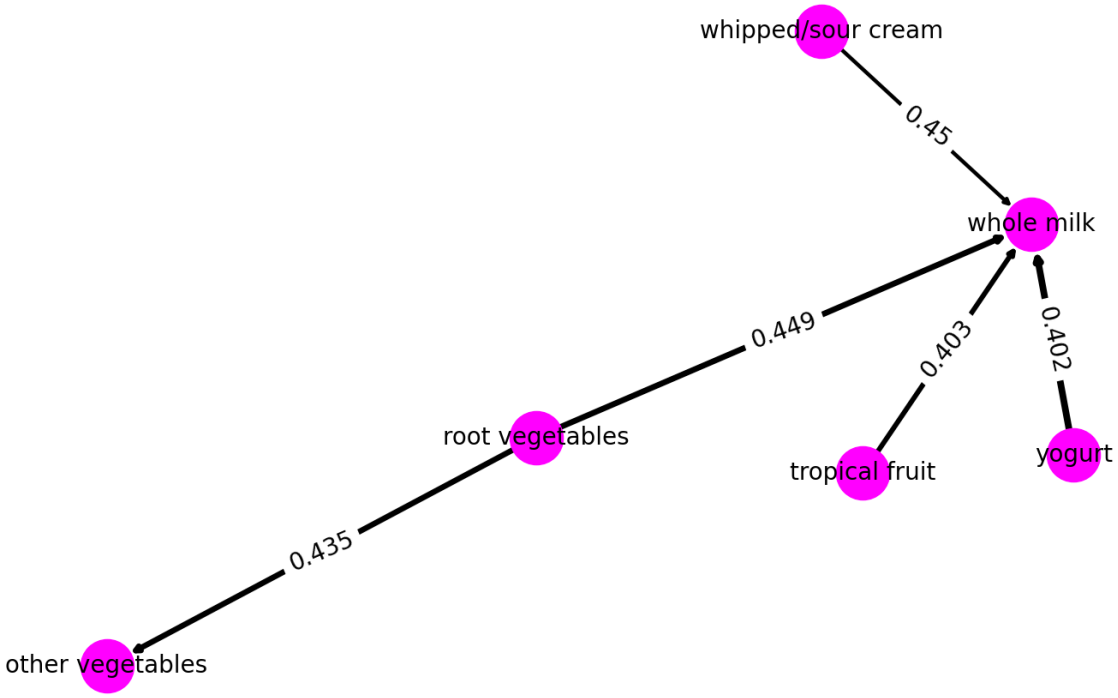


Рисунок 7 — Граф набора товара

Из графа можно сделать вывод, что при покупке root_vegetables, tropical_fruit, yogurt, whipped/sour cream с вероятностью примерно 40% возьмут и whole milk. При покупке root_vegetables возможно еще возьмут и other_vegetables.

17. Альтернативные способы представления

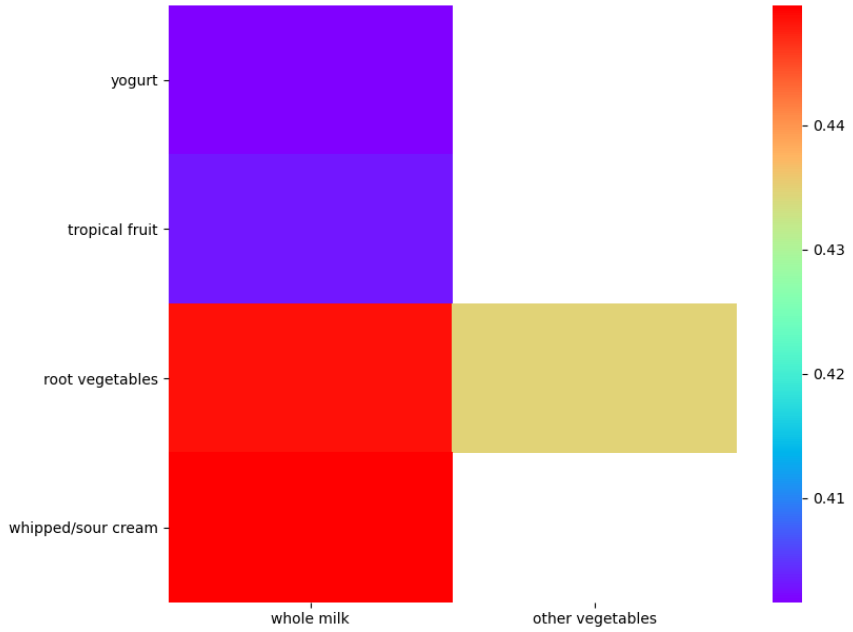


Рисунок 8 — Цветовая карта

	whole milk	other vegetables
yogurt	0.401603	NaN
tropical fruit	0.403101	NaN
root vegetables	0.448694	0.434701
whipped/sour cream	0.449645	NaN

Рисунок 9 — Текстовое представление

Вывод

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.