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

ОТЧЕТ

по лабораторной работе №4

по дисциплине «Машинное обучение»

Тема: Ассоциативный анализ

Студент гр. 6304 _____ Ястребков А. С.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург

2020

Цель работы:

Ознакомиться с методами ассоциативного анализа из библиотеки Mlxtend.

Ход работы

Загрузка данных.

Был загружен датасет groceries.csv (фрагмент исходного датасета показан на рис. 1), он был преобразован к массиву Numpy, были удалены все NaN-значения. Был получен список всех уникальных товаров и их количество — 169 наименований (листинг 1, рис. 2).

	Item(s)	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	...	Item 23	Item 24	Item 25	Item 26	Item 27	Item 28	Item 29	Item 30	Item 31	Item 32
0	4	citrus fruit	semi-finished bread	margarine	ready soups	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	3	tropical fruit	yogurt	coffee	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	1	whole milk	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	4	pip fruit	yogurt	cream cheese	meat spreads	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	4	other vegetables	whole milk	condensed milk	long life bakery product	NaN	NaN	NaN	NaN	NaN	...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Рис. 1. Фрагмент исходного датасета.

Листинг 1. Загрузка данных, обработка и получение списка уникальных товаров.

```
all_data = pd.read_csv('data/groceries - groceries.csv')
np_data = all_data.to_numpy()
np_data = [[elem for elem in row[1:] if isinstance(elem,
str)] for row in np_data]
unique_items = set()

for row in np_data:
    for elem in row:
        unique_items.add(elem)
print(unique_items)
print(len(unique_items))
```

```
B [7]: print(unique_items)
       print(len(unique_items))
```

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

Рис. 2. Список уникальных товаров.

1. Ассоциативный анализ алгоритмами FPGrowth и FPMax

Был проведён ассоциативный анализ для имеющихся данных с помощью алгоритмов FPGrowth и FPMax при минимальном уровне поддержки 0.03. Фрагменты результатов работы алгоритмов представлены на рис. 3-4. В отличие от алгоритма Apriori данные алгоритмы используют деревья шаблонов для поиска часто встречающихся наборов, что даёт линейный рост вычислительной сложности.

	support	itemsets	length
5	0.255516	(whole milk)	1
8	0.193493	(other vegetables)	1
11	0.183935	(rolls/buns)	1
19	0.174377	(soda)	1
2	0.139502	(yogurt)	1
..	...	...	...
43	0.031012	(onions)	1
61	0.030605	(rolls/buns, sausage)	2
44	0.030503	(citrus fruit, whole milk)	2
42	0.030402	(specialty chocolate)	1
50	0.030097	(pip fruit, whole milk)	2

Рис. 3. Фрагмент выходных данных алгоритма FPGrouth.

	support	itemsets	length
28	0.030097	(pip fruit, whole milk)	2
0	0.030402	(specialty chocolate)	1
32	0.030503	(citrus fruit, whole milk)	2
34	0.030605	(rolls/buns, sausage)	2
1	0.031012	(onions)	1
26	0.032232	(whipped/sour cream, whole milk)	2
44	0.032740	(soda, other vegetables)	2

Рис. 4. Фрагмент выходных данных алгоритма FPMax.

В таблице 1 даны уровни поддержки для наборов из 1 и 2 элементов для FPGrowth и FPMax. Разница в единственном значении обсуловлена тем, что второй алгоритм генерирует наборы максимальной длины, то есть, для каждого набора гарантируется, что он не является частью другого часто встречающегося набора. Таким образом, часть наборов длиной один включена в ч=наборы длиной 2.

Таблица 1. Уровни поддержки алгоритмов FPGrowth и FPMax.

эл-тов		FPGrowth	FPMax
1	min	0.030402	0.030402
	max	0.255516	0.098526
2	min	0.030097	0.030097
	max	0.074835	0.074835

На рис. 5 представлена гистограмма 10 наиболее часто встречающихся товаров. Она соответствует наборам из 1 элемента при использовании алгоритма FPGrowth.

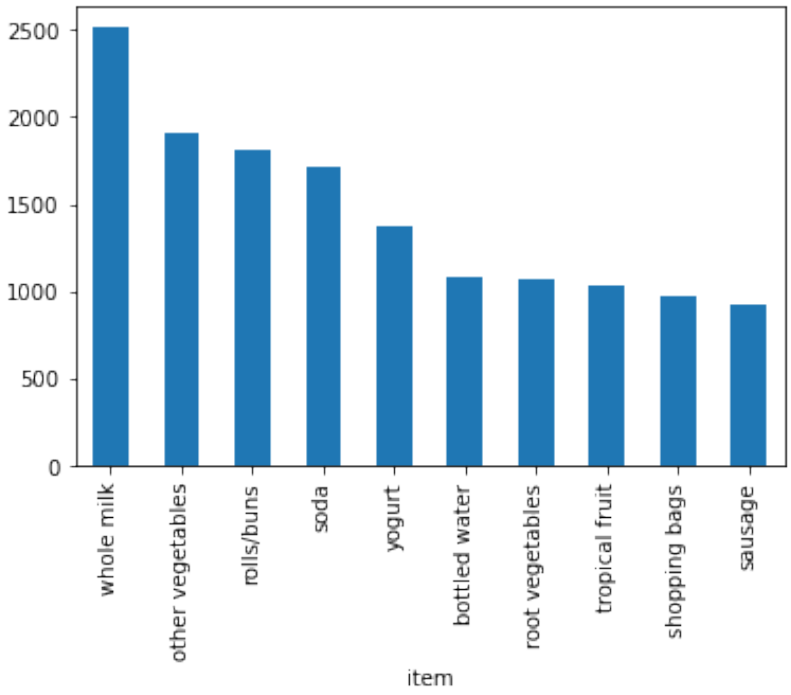

Рис. 5. Гистограмма самых часто встречающихся товаров.

Был проведён анализ алгоритмами FPGrowth и FPMax данных с неполным набором товаров (см. листинг 2). В результате, изменились значения минимального уровня поддержки для наборов, поскольку были удалены товары, ранее имевшие минимальный уровень поддержки. Результаты показаны в таблице 2.

Листинг 2. Анализ неполного набора товаров.

```
items = ['whole milk', 'yogurt', 'soda', 'tropical
    fruit', 'shopping bags', 'sausage', 'whipped/sour
    cream', 'rolls/buns', 'other vegetables', 'root
    vegetables', 'pork', 'bottled water', 'pastry',
    'citrus fruit', 'canned beer', 'bottled beer']
np_data = all_data.to_numpy()
np_data = [[elem for elem in row[1:] if isinstance(elem,
str) and elem in items] for row in np_data]
```

Таблица 2. Уровни поддержки алгоритмов FPGrowth и FPMax для неполного набора данных.

эл-тов		FPGrowth	FPMax
1	min	0.057651	0.057651
	max	0.255516	0.098526
2	min	0.030503	0.030503
	max	0.074835	0.074835

Для лагоритмов FPGrowth и FPMax были построены графики количества наборов от минимального уровня поддержки (рис. 6-7). На них снова видно, что FPMax генерирует меньше наборов длиной 1 и 2, поскольку их элементы входят в наборы с большим числом элементов.

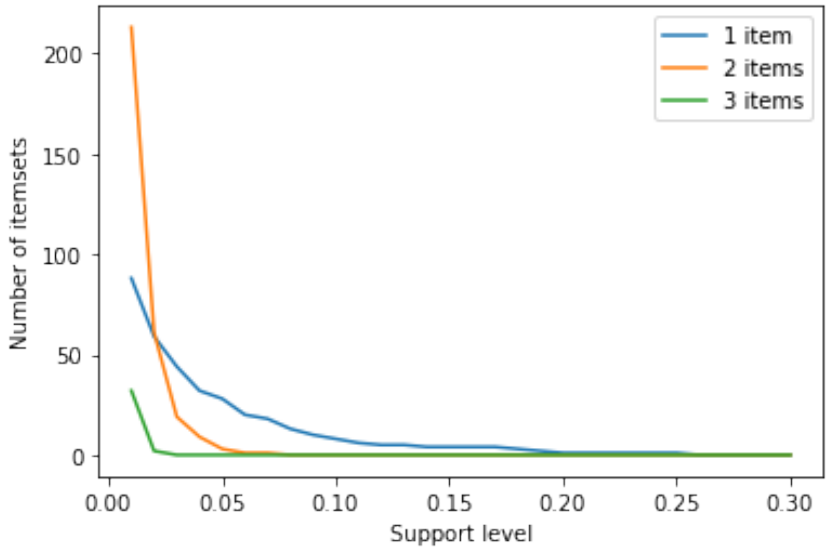

Рис. 6. График зависимости числа наборов от минмиального уровня поддержки алгоритма FPGrowth.

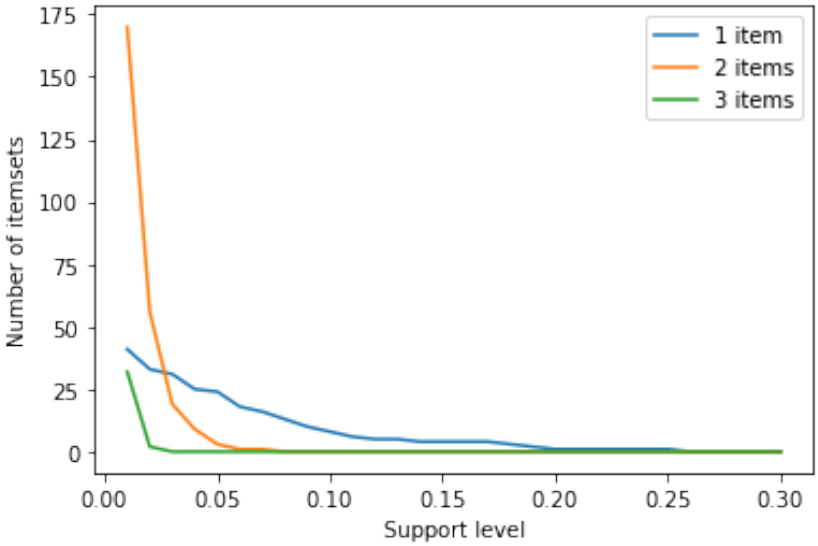

Рис. 7. График зависимости числа наборов от минмиального уровня поддержки алгоритма FPMax.

2. Ассоциативные правила.

Проведён ассоциативный анализ при уровне поддержки 0.3, результат показан на рис. 8.

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction
0	(yogurt)	(whole milk)	0.241240	0.421869	0.110954	0.459933	1.090228	0.009183	1.070481
1	(yogurt)	(other vegetables)	0.241240	0.335079	0.085985	0.356427	1.063713	0.005150	1.033172
2	(tropical fruit)	(yogurt)	0.185864	0.241240	0.057994	0.312026	1.293423	0.013156	1.102890
3	(tropical fruit)	(other vegetables)	0.185864	0.335079	0.071083	0.382449	1.141370	0.008804	1.076706
4	(tropical fruit)	(whole milk)	0.185864	0.421869	0.083770	0.450704	1.068352	0.005359	1.052495
5	(other vegetables)	(whole milk)	0.335079	0.421869	0.148208	0.442308	1.048449	0.006849	1.036649
6	(whole milk)	(other vegetables)	0.421869	0.335079	0.148208	0.351313	1.048449	0.006849	1.025026
7	(rolls/buns)	(whole milk)	0.296214	0.421869	0.112163	0.378654	0.897564	-0.012801	0.930450
8	(bottled water)	(whole milk)	0.185461	0.421869	0.068063	0.366992	0.869921	-0.010177	0.913309
9	(bottled water)	(soda)	0.185461	0.267217	0.057390	0.309446	1.158033	0.007832	1.061153
10	(citrus fruit)	(whole milk)	0.146395	0.421869	0.060411	0.412655	0.978159	-0.001349	0.984313
11	(citrus fruit)	(other vegetables)	0.146395	0.335079	0.057189	0.390646	1.165836	0.008135	1.091192
12	(root vegetables)	(other vegetables)	0.196335	0.335079	0.093838	0.477949	1.426378	0.028050	1.273671
13	(root vegetables)	(whole milk)	0.196335	0.421869	0.096859	0.493333	1.169400	0.014031	1.141049
14	(sausage)	(rolls/buns)	0.167539	0.296214	0.060612	0.361779	1.221342	0.010985	1.102730
15	(sausage)	(whole milk)	0.167539	0.421869	0.059203	0.353365	0.837619	-0.011477	0.894062
16	(sausage)	(other vegetables)	0.167539	0.335079	0.053363	0.318510	0.950552	-0.002776	0.975687
17	(whipped/sour cream)	(whole milk)	0.124245	0.421869	0.063834	0.513776	1.217858	0.011419	1.189023
18	(whipped/sour cream)	(other vegetables)	0.124245	0.335079	0.057189	0.460292	1.373683	0.015557	1.232002
19	(pastry)	(whole milk)	0.150624	0.421869	0.065848	0.437166	1.036260	0.002304	1.027179

Рис. 8. Результат ассоциативного анализа для метрики по умолчанию (confidence).

Значения столбцов (пусть A — антецедент, C — консеквент):

- **antecedent support, consequent support** — значения поддержки для антецедента и консеквента соответственно;
- **support** — поддержка набора из антецедента и консеквента:

$$support(A \rightarrow C) = support(A \vee C)$$

- **confidence** — вероятность получить консеквент в транзакции с антецедентом:

$$confidence(A \rightarrow C) = \frac{support(A \rightarrow C)}{support(A)}$$

- **lift** — показывает, насколько чаще встречаются вместе $A \rightarrow C$, чем если бы они были независимы (единица при независимых антецеденте и консеквенте):

$$lift(A \rightarrow C) = \frac{support(A \rightarrow C)}{support(A)support(B)}$$

- **leverage** — разница между частотой появления $A \rightarrow C$ вместе и частотой, какая была бы при их независимости (ноль, если они действительно независимы):

$$leverage(A \to C) = support(A \to C) - support(A)\,support(B)$$

- **conviction** — метрика, показывающая, насколько сильно консеквент зависит от антецедента (принимает значение единицы при независимости):

$$conviction(A \to C) = \frac{1 - support(A \to C)}{1 - confidence(A \to C)}$$

Для каждой метрики были получены правила. Минимальный порог выбирался так, чтобы количество правил было не менее 10. Для каждой метрики получены значения среднего, СКО и медианы, результаты сведены в таблицу 3.

Таблица 3. Статистические параметры метрик.

	support	confidence	lift	leverage	conviction
уровень поддержки	0.08	0.35	1.1	0.008	1.04
среднее	0.102	0.417	1.2408	0.0133	1.108
СКО	0.021	0.0535	0.0989	0.0059	0.0687
медиана	0.0953	0.4127	1.2179	0.0114	1.084

Для правил, сформированных по метрике confidence с минимальным уровенм поддержки 0.4 был построен граф, показанный на рис. 9. По такому графу можно строить выводы в подобных формулировках: «если в транзакции есть товары citrus fruit, pastry, tropical fruit, то в ней, вероятно, есть и товар whole milk».

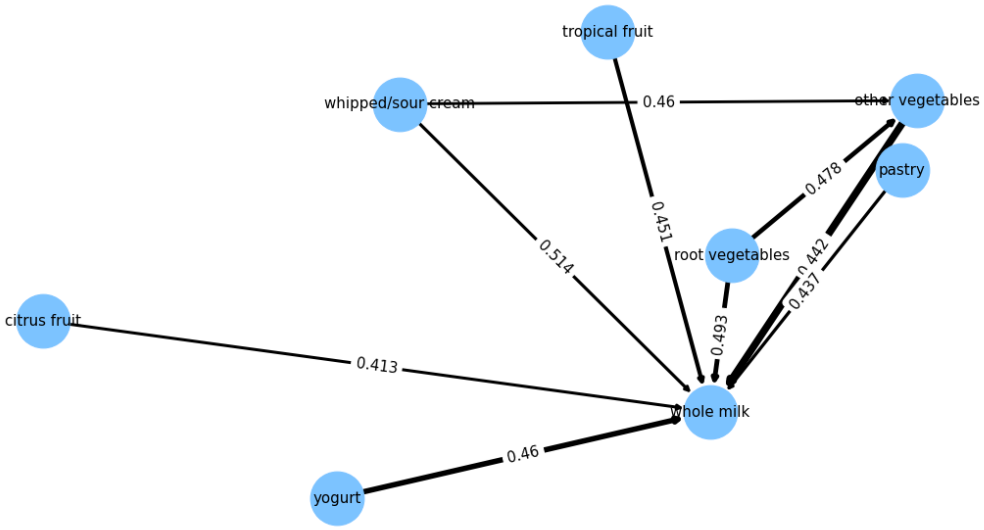

Рис. 9. Граф для сформированных правил.

В качестве альтернативного метода визуализации такой информации можно использовать график типа heat map (тепловая карта), как показано на рис. 10 для тех же правил.

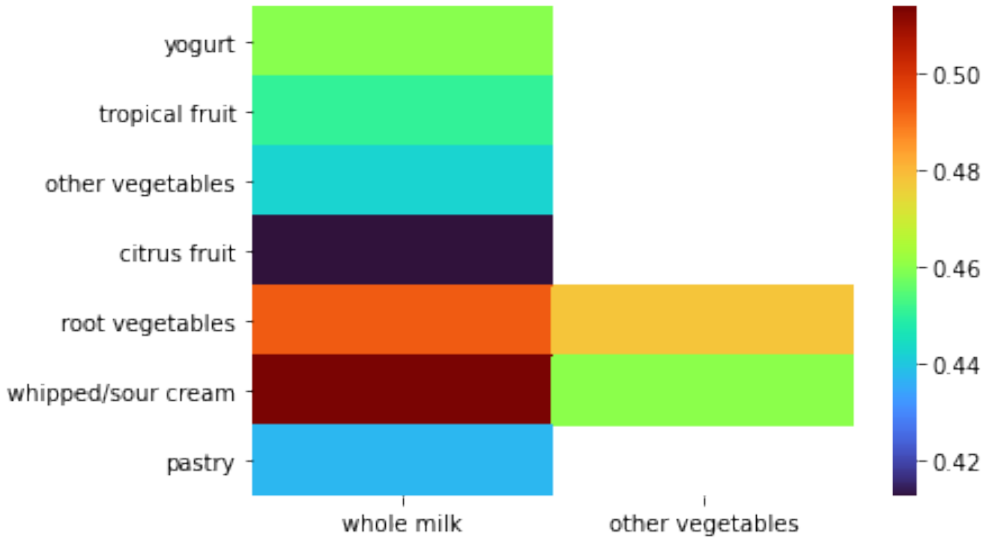

Рис. 10. Визуализация правил с помощью тепловой карты.

Выводы

В ходе выполнения лабораторной работы были изучены принципы работы алгоритмов ассоциативного анализа FRGrowth и FPMax. Было установлено их основное отличие от алгоритма Apriori — работа на базе FP-деревьев, а также

основное различие между собой — алгоритм FPMax пытается построить наборы максимальной длины так, чтобы они не были частью наборов большей длины, из-за чего генерируется меньше наборов. Была изучена работа алгоритма построения ассоциативных правил.