Лабораторная работа №4

Ассоциативный анализ

Цель:

Ознакомиться с методами ассоциативного анализа из библиотеки MLxtend

Выполнение:

Для выполнения лаб. работы необходимо установить библиотеку MLxtend

Для этого в терминале введите:

```
pip install mlxtend
```

Загрузка данных:

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

```
all_data = pd.read_csv('groceries - groceries.csv')
print(all_data) #Видно, что датафрейм содержит NaN значения
```

3. Переформируем данные удалив все значения NaN

```
np_data = all_data.to_numpy()
np_data = [[elem for elem in row[1:] if isinstance(elem,str)] for row in
np_data]
```

4. Получим список всех уникальных товаров

```
unique_items = set()

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

5. Выведите список товаров, а также их количество

FPGrowth и FPMax

1. Преобразуем данные к виду, удобному для анализа

```
from mlxtend.preprocessing import TransactionEncoder

te = TransactionEncoder()
te_ary = te.fit(np_data).transform(np_data)
data = pd.DataFrame(te_ary, columns=te.columns_)
```

2. Проведем ассоциативный анализ используя алгоритм FPGrowth при уровне поддержки 0.03

```
from mlxtend.frequent_patterns import fpgrowth

result = fpgrowth(data, min_support=0.03, use_colnames = True)
print(result)
```

3. Проанализируйте получившиеся варианты. Определите минимальное и максимальное значения для уровня поддержки для набора из 1,2, и.т.д. объектов.
4. Проведите аналогичный анализ используя алгоритм FPMax
5. Сравните полученные результаты для FPGrowth и FPMax. Объясните в чем разница работы алгоритмов
6. Постройте гистограмму для каждого товара. Столбцы на гистограмме должны быть упорядочены по уменьшению частоты. Отобразите результат только для 10 самых встречаемых товаров. Как данная гистограмма коррелируют с результатами?
7. Преобразуем набор данных, чтобы он содержал ограниченный набор 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]
```

8. Проведите анализ FPGrowth и FPMax для нового набора данных. Проанализируйте, что изменилось
9. Постройте график изменения количества получаемых правил от уровня поддержки. На графике отдельно отобразите кривые для набора товаров 1, 2, и.т.д. Какие выводы можно сделать по данному графику?

Ассоциативные правила

1. Сформируем набор данных из определенных товаров и так, чтобы размер транзакции был 2 и более

```
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]
np_data = [row for row in np_data if len(row) > 1]
```

2. Получим частоты наборов используя алгоритм FPGrowth

```
result = fpgrowth(data, min_support=0.05, use_colnames = True)
```

3. Проведем ассоциативный анализ

```
rules = association_rules(result, min_threshold = 0.3)
print(rules)
```

   Объясните, что означает каждая колонка в полученных результатах.

4. Определите, на основе какой метрики проводится расчет
5. Проведите построение ассоциативных правил для различных метрик (значение min_threshold выберите такое, чтобы выводилось не менее 10 правил). Какой смысл несет каждая метрика?
6. Рассчитайте среднее значение, медиану и СКО для каждой из метрик.
7. Постройте граф для следующего анализа

```
rules = association_rules(result, min_threshold = 0.4, metric='confidence')
```

   Каждая вершина графа должна отображать набор товаров. Граф должен быть ориентирован от антецедента к консеквенту. Ширина ребра должна отображать уровень support, а подпись на ребре отображать confidence.

   Для отрисовка графа можно использовать библиотеку NetworkX

8. Проанализируйте полученный граф, какую информацию можно из него извлечь?
9. Предложите свои способы визуализации полученных правил