Лабораторная работа №6

Кластеризация (DBSCAN, OPTICS)

Цель:

Ознакомиться с методами кластеризации модуля Sklearn

Выполнение:

Загрузка данных:

1. Загрузить датасет по ссылке: https://www.kaggle.com/arjunbhasin2013/ccdata . Данные представлены в виде csv файла. Датасет содержит пропущенные значения
2. Создать Python скрипт. Загрузить данные в датафрейм, убрав столбец с метками и откинув наблюдения с пропущенными значениями

```
import pandas as pd
import numpy as np

data = pd.read_csv('CC GENERAL.csv').iloc[:,1:].dropna()
```

DBSCAN

1. Проведем кластеризацию методов k-средних

```
from sklearn.cluster import KMeans

k_means = KMeans(init='k-means++', n_clusters=3, n_init=15)
k_means.fit(no_labeled_data)
```

2. Так как разные признаки лежат в разных шкалах, то стандартизируем данные

```
from sklearn import preprocessing

data = np.array(data, dtype='float')

min_max_scaler = preprocessing.StandardScaler()
scaled_data = min_max_scaler.fit_transform(data)
```

3. Проведем кластеризацию методов DBSCAN при параметрах по умолчанию. Выведем метки кластеров, количество кластеров, а также процент наблюдений, которые кластеризовать не удалось

```
clustering = DBSCAN().fit(scaled_data)

print(set(clustering.labels_))
print(len(set(clustering.labels_)) - 1)
print(list(clustering.labels_).count(-1) /  len(list(clustering.labels_)))
```

Опишите все параметры, которые принимает DBSCAN

4. Постройте график количества кластеров и процента не кластеризованных наблюдений в зависимости от максимальной рассматриваемой дистанции между наблюдениями. Минимальное значение количества точек образующих, кластер оставить по умолчанию
5. Постройте график количества кластеров и процента не кластеризованных наблюдений в зависимости от минимального значения количества точек, образующих кластер. Максимальную рассматриваемую дистанцию между наблюдениями оставьте по умолчанию
6. Определите значения параметров, при котором количество кластеров получается от 5 до 7, и процент не кластеризованных наблюдений не превышает 12%.
7. Понизьте размерность данных до 2 при используя метод главных компонент. Визуализируйте результаты кластеризации полученные в пункте 6 (метки должны быть получены на данных до уменьшения размерности). гайд по визуализации

OPTICS

1. Опишите параметры метода OPTICS, а также какими атрибутами он обладает
2. Найдите такие параметры метода OPTICS (*max_eps *и *min_samples*) при которых, чтобы получить результаты близкие к результатам DBSCAN из пункта 6

   В чем отличия от метода OPTICS от метода DBSCAN
3. Визуализируйте полученный результат, а также постройге график достижимости (reachable plot) гайд
4. Исследуйте работу метода OPTICS с использованием различных метрик (выберите не менее 5 метрик)