Задание 4. PCA, tSNE, UMAP

[1] Сгенерировать двумерные датасеты df_a (размерности 4x4, 16x16, 256x256), в которых в каждой строке x начиная с колонки y(=x) значения следующих (a/2) ячеек равны «1» *(если достигнут конец строки датасета, сверх размера датасета не заполняем, на новую строку не переходим)*, а оставшиеся ячейки датасета равны «0». Выполнить визуализацию датасетов в пространстве X-Y.

[2] Используя метод главных компонент (основанный на корреляциях), для каждого датасета df_a отдельно визуализировать:
- Scree plot;
- Отображения отдельных объектов-строк в двумерных пространствах первых компонент от 1 до a (a на выбор от 4 до 8) – без отображения старых векторов-переменных;
- Отображения отдельных объектов-строк в двумерных пространствах первых компонент от 1 до a (a на выбор от 4 до 8) – с отображением старых векторов-переменных.
- Отображения отдельных объектов-строк в двумерных пространствах первых компонент от c до c+a (a на выбор от 4 до 8, c – любая со 129) – без отображения старых векторов-переменных;
- Отображения отдельных объектов-строк в двумерных пространствах первых компонент от c до c+a (a на выбор от 4 до 8, c – любая со 129)– с отображением старых векторов-переменных.

[3] Используя методы UMAP, PacMAP и tSNE, для каждого датасета df_a отдельно визуализировать проекции с нескольких рендом-стартов, а также после предвращения датасета методом PCA.

[4] Используя make_blobs, сгенерировать на основе номеров строк (параметр x), столбцов (параметр y) и значений («0», «1») ячеек датасета df_256 новый датасет df_256v, в котором в каждой точке с центром (x,y) и дисперсией d (d – на выбор, в диапазоне от 5 до 20) сформировано по 10 случайных точек класса «1» или «0».

[5] Используя метод главных компонент (основанный на корреляциях), для датасета df_256v выполнить визуализации аналогично первым трем подпунктам пункта [2].

[6] Для датасета df_256v повторить пункт [3].

[7] Взять датасет с Kaggle, в котором есть >=10 переменных, >=10000 объектов и несколько классов. Для каждого датасета не забываем повторять шаги с задания 1 - ключевые характеристики датасета, корреляции, визуализация на всех парах var.

[8] Используя метод главных компонент (основанный на корреляциях), для датасета из пункта [7] выполнить визуализации аналогично первым трем подпунктам пункта [2]. Определить различными способами (метод Кайзера, метод ломанной трости), сколько компонент необходимо оставлять для регрессионной модели.

[9] Для датасета из пункта [7] повторить пункт [3].

https://umap-learn.readthedocs.io/en/latest/
https://github.com/YingfanWang/PaCMAP