

Лабораторная №5. Кластеризация

Набор данных

- Раздобудьте где-нибудь набор данных для задачи классификации. Он должен содержать не менее трёх классов.
- Приведите его в порядок: заполните пропуски, векторизуйте, нормализуйте.
- Спроецируйте данные на двухмерное пространство при помощи PCA/t-SNE/UMAP.
- Визуализируйте набор данных. Если в нём визуально не наблюдаются кластеры, то лучше выбрать другой набор данных.

Задание

- Выберите два алгоритма кластеризации: первый с возможностью явно задать число кластеров, а второй — без.
- Выберите две меры оценки качества кластеризации: внешнюю и внутреннюю.
- Для первого и второго алгоритма измерьте, как зависит внешняя и внутренняя мера от числа кластеров. Для второго алгоритма требуется выбрать гиперпараметр, от которого зависит число кластеров и варьировать его.