Кластеризация

Цели работы:

1. Реализовать алгоритм кластеризации и две меры качества.
2. Реализовать алгоритм сокращения размерности.
3. Анализ результатов.

Задание

Реализуйте любой алгоритм кластеризации на выбор:

1) K-Means;
2) DBSCAN;
3) иерархический алгоритм.

Также реализуйте две метрики качества кластеризации: одну внешнюю и одну внутреннюю. Внутренняя мера качества **должна** учитывать как межкластерные так и внутрикластерные расстояния.

Варианты внешних мер (формулы представлены в http://neerc.ifmo.ru/wiki/index.php?title=%D0%9E%D1%86%D0%B5%D0%BD%D0%BA%D0%B0_%D0%BA%D0%B0%D1%87%D0%B5%D1%81%D1%82%D0%B2%D0%B0_%D0%B2_%D0%B7%D0%B0%D0%B4%D0%B0%D1%87%D0%B5_%D0%BA%D0%BB%D0%B0%D1%81%D1%82%D0%B5%D1%80%D0%B8%D0%B7%D0%B0%D1%86%D0%B8%D0%B8):

1) Rand index
2) Jaccard index
3) F-мера

Варианты внутренних мер:

1) Силуэтный индекс
2) метрика Calinski-Harabasz
3) метрика Dunn

Нарисуйте два графика: набор данных с реальными метками и с метками полученными в результате кластеризации. Постарайтесь выбрать такие гиперпараметры алгоритма кластеризации, чтобы результат кластеризации был как можно более похож на реальные метки. Для отрисовки многомерных данных используйте алгоритм сокращения размерности (например PCA или tSNE), но кластеризацию по прежнему проводите в многомерном пространстве.

Постройте график зависимости выбранных метрик качества кластеризации от числа кластеров при выборе K-Means или иерархического алгоритма. В случае DBSCAN постройте график зависимости выбранных метрик качества кластеризации от радиуса шара.

Наборы данных

Возьмите любой набор данных из первой лабораторной работы про KNN. Не забудьте векторизовать и нормализовать набор данных.