
# Машинное обучение

НИЯУ МИФИ, Кафедра финансового мониторинга

Лабораторный практикум.

В.Ю. Радыгин, Д.Ю. Куприянов

Семестр 2. Лабораторная работа 3

## Лабораторная работа 3 (2 семестр)

Лабораторная работа 3 рассчитана на два занятия и работу дома. Её целью является изучение основ кластеризации данных с помощью метода агломеративных методов и метода k-means.

**Задание 1**

1. Загрузите с сайта https://sci2s.ugr.es/keel/datasets.php набор статистических данных, указанный в вашем варианте. Разберитесь, какие данные приведены в наборе и какой атрибут является меткой класса.
2. На основе загруженного файла создайте Pandas DataFrame, подобрав правильные типы данных столбцов.
3. Выполните стандартизацию полученного дата фрейма.
4. Оцените число кластеров для метода K-means с помощью методов локтя на основе инерции и искажения и с помощью метрики силуэта. Сравните рекомендуемые числа кластеров с реальным числом классов.
5. Выполните кластеризацию методом K-means с реальным числом кластеров. Постройте кросс-таблицу для сравнения оригинальных и предсказанных классов.
6. Постройте дендрограмму по методу Уорда для вашего набора данных. Оцените, соответствует ли дендрограмма реальному числу классов.
7. Выполните кластеризацию методом Уорда с реальным числом кластеров.
8. Постройте три раза проекцию по двум первым координатам точек набора данных, раскрасив их в различные цвета в соответствии с реальными классами, классами, предсказанными k-means, и классами, предсказанными методом Уорда.
9. Примените к исходному набору данных (после стандартизации) метод главных компонент, выбрав компоненты, соответствующие собственным числам, большим 1.
10. Повторите для модифицированного набора данных шаги 4-9.
11. Дайте оценку, стала ли кластеризация точнее после применения метода главных компонент.

**Варианты**

**Задание 1**

1. https://sci2s.ugr.es/keel/dataset.php?cod=210
2. https://sci2s.ugr.es/keel/dataset.php?cod=209
3. https://sci2s.ugr.es/keel/dataset.php?cod=107
4. https://sci2s.ugr.es/keel/dataset.php?cod=72