



Машинное обучение

НИЯУ МИФИ, Кафедра финансового мониторинга

Лабораторный практикум

В.Ю. Радыгин, Т.А. Манаенкова

Семестр 1. Лабораторная работа 4

Лабораторная работа 4

Лабораторная работа 4 рассчитана на два занятия и работу дома. Её целью является изучение основ классификации данных с помощью древовидных структур.

Задание 1

1. Зарегистрируйтесь на сайте <https://www.kaggle.com/datasets> и загрузите с него набор статистических данных, указанный в вашем варианте. Разберитесь, какие данные приведены в наборе и какой атрибут является меткой класса.
2. На основе загруженного CSV-файла создайте Pandas DataFrame, подобрав правильные типы данных столбцов.
3. Выполните стандартизацию полученного дата фрейма.
4. Разделите дата фрейм на обучающую и тестовую выборки в соотношении 6 к 4.
5. Меняя значение параметра альфа ([0.005, 0.01, 0.015, 0.02, 0.025, 0.03, 0.035, 0.2, 0.8]) и критерий классификации ([Entropy, Gini]) обосновано подберите наиболее удачное дерево классификации для подготовленных выборок.

Варианты датасетов

Задание 1

1. <https://www.kaggle.com/uciml/glass/data>
2. https://www.kaggle.com/raghupalem/bill_authentication
3. <https://www.kaggle.com/uciml/glass/data>
4. https://www.kaggle.com/raghupalem/bill_authentication

Либо

1. <https://sci2s.ugr.es/keel/dataset/data/classification/winequality-red.zip>
2. <https://sci2s.ugr.es/keel/dataset/data/classification/winequality-white.zip>
3. <https://sci2s.ugr.es/keel/dataset/data/classification/segment.zip>
4. <https://sci2s.ugr.es/keel/dataset.php?cod=72>

Сильная группа (по очень большому желанию)

- Выполните все задания также для алгоритмов C4.5 и ID3.