Московский авиационный институт (национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №2 по курсу Машинное обучение

Студент: С. Я. Симонов
Преподаватель: Ахмед Самир Халид
Группа: М8о-406Б-18

Дата:
Оценка:
Подпись:

Москва, 2021

1. Условие задачи:

Необходимо реализовать алгоритмы машинного обучения. Применить данные алгоритмы на наборы данных, подготовленных в первой лабораторной работе. Провести анализ полученных моделей, вычислить метрики классификатора. Произвести тюнинг параметров в случае необходимости. Сравнить полученные результаты с моделями реализованными в scikit-learn. Аналогично построить метрики классификации. Показать, что полученные модели не переобучились. Также необходимо сделать выводы о применимости данных моделей к вашей задаче. Задачи со звездочкой бьются по вариантам: N по списку

1) ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ
2) *SVM - ПЕРВЫЙ ВАРИАНТ
3) ДЕРЕВО РЕШЕНИЙ
4) *RANDOM FOREST - ВТОРОЙ ВАРИАНТ

Т.к. у меня второй вариант, был реализован RANDOM FOREST.

2. Метод решения

Для классификации я взял дата сет из первой лабораторной работы с информацией о грибах. Классификация бинарная: съедобные и несъедобные грибы.

В отличие от обычной регрессии, в методе логистической регрессии не производится предсказание значения числовой переменной исходя из выборки исходных значений. Вместо этого, значением функции является вероятность того, что данное исходное значение принадлежит к определенному классу.

Основная задача при построении дерева решений — последовательно и рекурсивно разбить обучающее множество на подмножества с применением решающих правил в узлах. Этот процесс продолжается до того, пока все узлы в конце ветвей не станут листами.

Random Forest – модель, состоящая из множества деревьев решений. Вместо того,чтобы просто усреднять прогнозы разных деревьев, эта модель использует две ключевые концепции, которые и делают этот лес случайным:

1) Случайная выборка образцов из набора данных при построении деревьев.
2) При разделении узлов выбираются случайные наборы параметров.

3. Вывод

В ходе данной лабораторной работы я познакомился с тремя алгоритмами для классификации данных. Самым интересным и интуитивно понятным алгоритмом оказалось дерево решений. Также данный алгоритм хорошь тем, что можно явно сказать почему было предсказанно именно это значение. Random Forest оказлся для меня самым сложным алгоритмом. Крайне тяжело воспринимается случайное разбиение данных.