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: В. В. Косогоров
Преподаватель: Ахмед Самир Халид
Группа: М8О-306Б-18

Дата:
Оценка:
Подпись:

Москва, 2021

Лабораторная работа №2

Задача:
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 % 2 + 1.
1) ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ
2) *SVM - ПЕРВЫЙ ВАРИАНТ
3) ДЕРЕВО РЕШЕНИЙ
4) *RANDOM FOREST - ВТОРОЙ ВАРИАНТ

1 Метод решения

Для классификации я взял датасет из первой лабораторной работы с информацией о крушении Титаника. Класификация здесь бинарная: нужно предсказать, пережил ли человек крушение.

Для анализа результатов классификации я написал функцию, которая возвращает confusion matrix для реальных и предсказанных классов. Также я написал функцию metrics, которая возвращает accuracy, precision и recall, взяв данные для значение TP, FP, FN, TN из confusion matrix.

Гиперпараметры для моих моделей я подбирал, просто проходясь по списку из списков параметров. Для SVM классификатора из sklearn я попробовал подобрать гиперпараметры с помощью GridSearchCV.

2 Выводы

Я считаю, что модели применимы к моей задаче. Для применимости логистической регрессии важно не иметь большого числа признаков. Из-за того, что эта модель предсказывает вероятности, а не классы, то можно изменять значения порога для предсказания какого-либо класса. В моей задаче recall составил около 67%. То есть, модель довольно часто говорила, что выжившие люди погибнут. Если бы мы хотели бы проводить классификацию и было бы крайне не желательно ошибаться в эту сторону, то можно просто изменить порог вероятности для предсказания выживания. SVM справляется с нелинейными границами разделения классов и ему не страшно большое количество признаков. Но, в отличие от Decision tree и линейной регрессии, его результаты сложны в интерпретации.
Главный плюс Decision tree – это его прозрачность в интерпретации. Минус – это непараметрическая модель, и поэтому её легко переобучить.