

Московский авиационный институт
(национальный исследовательский университет)

Факультет информационных технологий и прикладной
математики

Кафедра вычислительной математики и программирования

Лабораторная работа №2 по курсу «Искусственный интеллект»

Студент: А. Е. Максимов
Преподаватель: А. С. Халид
Группа: М8О-406Б
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.

1. ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ
2. KNN
3. SVM
4. Decision Tree
5. Random Forest

1 Описание решения

Для выполнения задачи был использован датасет, обработанный в первой лабораторной работе:

1. Задача классификации: <https://www.kaggle.com/deepu1109/star-dataset>

Поскольку большинство алгоритмов решают задачи бинарной классификации, я провёл дополнительное изменение в датасете, уменьшив количество возможных классов до двух.

В качестве основы алгоритмов я использовал открытые описания в интернете.

Для унификации взаимодействия с классификаторами я написал каждому основной набор функций, обеспечивая удобную работу. У коллеги я увидел функцию, объединяющую обучение, проверку результата и вычисление метрик и решил использовать такую же.

Проводя анализ результатов, я пришёл к выводу, что используемый датасет слишком мал и прост для полноценных исследований. Многие классификаторы обучались до 100% точности, вероятно из-за малого количества данных, а также из-за того, что классификация звёзд изначально могла строиться с использованием линейных функций.

Основные проблемы были с KNN-алгоритмом. структура данных и выбранный метод измерения расстояния до соседей создали очень плохую комбинацию, предсказывающую случайные события, а не класс звезды.

Для моего первого датасета модели подходили относительно неплохо, не считая того, что пришлось самостоятельно уменьшить количество классов, но задачу регрессии (для второго датасета) они не решали, поэтому второй датасет был не востребован.

2 Выводы

В этой лабораторной работе я глубже разобрался в основных классификаторах. Интересная система самообучения с помощью оптимизаций различных функций была для меня нова, сложна. В конце концов, я разобрался в особенностях классификаторов и смог завершить работу.