Московский авиационный институт
(национальный исследовательский университет)

Факультет информационных технологий и прикладной математики

Кафедра вычислительной математики и программирования

Лабораторная работа №2 по курсу Искусственный интеллект (Машинное обучение)

Студент: Д. О. Петрухин
Преподаватель: Ахмед Самир Халид
Группа: М8О-301Б
Дата:
Оценка:
Подпись:

Москва, 2021

Общая постановка 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.

Вариант по списку: 11.

Алгоритмы:

1. Логистическая регрессия.
2. Дерево решений.
3. Random Forest.

Логистическая регрессия

Теория.

Гипотеза:

$$h(x) = \sigma(w^T x) = \frac{1}{1 + e^{-w^T x}}$$

$$P(Y = y_i | x_i) = h(x_i)^{y_i} * (1 - h(x_i))^{1-y_i}$$

$$lnP(Y = y_i | x_i) = y_i ln h(x_i) + (1 - y_i) ln(1 - h(x_i))$$

Функция ошибок:

$$L(w) = -\sum_{i=1}^{l} lnP(Y = y_i | x_i) \to \min_w$$

Используется метод градиентного спуска:

$$w = w - \alpha \nabla L$$

Вычисление градиента:

$$\nabla L = x^T \cdot (h(x) - y)/y.size$$

Результаты работы.

```
Результат собственной реализации логистической регресси на обучающей выборке: 75.1972230
Результат собственной реализации логистической регрессии на тестовой выборке: 72.3484848
Результат sklearn реализации логистической регрессии: 72.72727272727273%
```

Дерево решений

Алгоритм построения дерева:

1. Создаем корень дерева. Находим наилучшее разбиение путём вычисления для каждого значения атрибута индекса Джини и выбора наименьшего из них.
2. Рекурсивно продолжаем разбиение для левой и правой ветвей, пока энтропия не окажется достаточно малой величиной или не будет достигнуто ограничение по высоте дерева.

Результаты работы:

```
Результат собственной реализации дерева решений на обучающей выборке: 82.4234774376775%
Результат собственной реализации дерева решений на тестовой выборке: 73.73737373737373%
Результат sklearn реализации дерева решений: 73.23232323232324%
```

Random Forest

Алгоритм:

1. Сгенерировать случайную подвыборку с повторениями размером N из обучающей выборки.
2. Построить решающее дерево, классифицирующее образцы данной подвыборки, причём в ходе создания очередного узла дерева будет выбираться набор признаков, на основе которых производится разбиение (не из всех M признаков, а лишь из m случайно выбранных).
3. Дерево строится до полного исчерпания подвыборки.

Результаты работы:

```
Результат собственной реализации random forest на обучающей выборке: 76.39633953928684%
Результат собственной реализации random forest на тестовой выборке: 72.09595959595958%
Результат sklearn реализации random forest: 72.47474747474747%
```