Московский Авиационный Институт

(Национальный Исследовательский Университет)

Кафедра 806 «Вычислительная информатика и программирование»

Факультет: «Информационные технологии и прикладная математика»

Лабораторная работа

Дисциплина: «Машинное Обучение»

6 семестр

Задание 2

|  |  |
| --- | --- |
| Группа: | М8О-308Б-18, №6 |
| Студент: | Васильева Василиса Евгеньевна |
| Преподаватель: | Ахмед Самир Халид |
| Оценка: |  |
| Дата: | 06.06.2021 |

Москва, 2021

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. Задачи со звездочкой бьются по вариантам: N по списку % 2 + 1.

1) ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

2) \*SVM - ПЕРВЫЙ ВАРИАНТ

3) ДЕРЕВО РЕШЕНИЙ

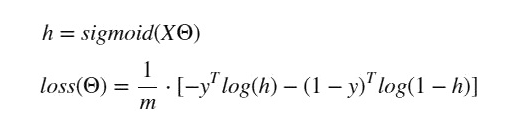
4) \*RANDOM FOREST - ВТОРОЙ ВАРИАНТ

Логистическая регрессия

Логистическая регрессия – классический метод, который используется в основном для работы с двоичными классификациями. Первой составляющей логистической регрессии является сигмоида.

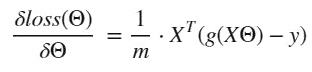
res = 1/(1 + np.exp(-z)) - возвращает вероятность принадлежности объекта классу.

Потом определяется функция потери.

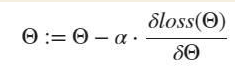


Ее необходимо оптимизировать, для этого ищутся подходящие коэффициенты. Обычно это w и b.

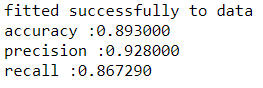
Для поиска коэффициентов используем градиентный спуск.



Обновление весов



Результат запуска программы:



Дерево решений и Random Forest

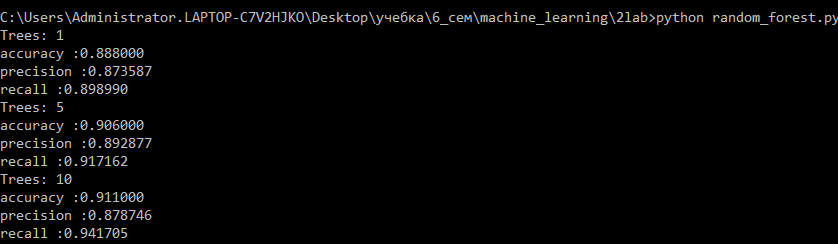
Индекс Джини — это показатель, определяющий, как часто случайно выбранный элемент будет неверно идентифицирован. Это означает, что атрибут с более низким индексом Джини должен быть предпочтительнее.



Мы берем данные, смотрим на некоторую фичу, выбираем значение и разделяем датасет на две части. В одну идут значения, меньшие выбранного, в другую – большие. В узле хранится правило разделения. Все дерево делится на два типа узлов, листья – в них объекты, внутренние – в них правила. Число для правила выбираем такое, чтобы индекс Джини был как можно меньше. Устанавливаем ограничение на глубину дерева и количество объектов в листьях, чтобы избежать переобучения.

В случае с лесом, берем датасет, из него делаем новый с рандомными объектами из первого. Объекты могут повторяться. Для этого рандомного датасета строим дерево, потом создаем еще датасет и дерево. Следовательно, деревья будут разные. Потом берем все эти деревья и усредняем их предсказания.

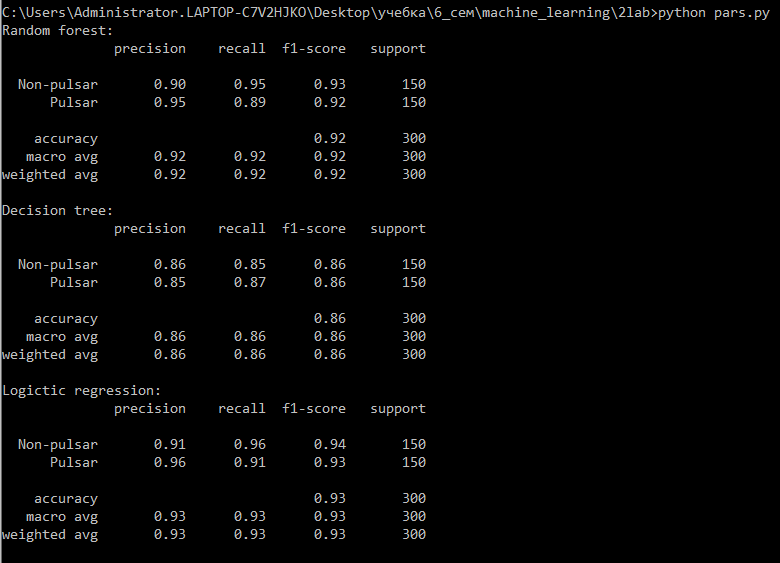
Результат запуска:



Библиотечные алгоритмы

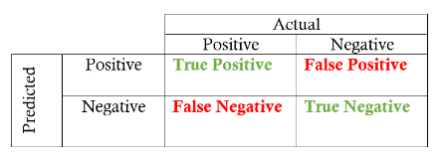
Чтобы сравнить наши алгоритмы с библиотечными, тоже пишем новый код, где запускаем наши выборки в алгоритмы sklearn.

Их результаты:



Выводы

Для подсчета метрик используем матрицу путаницы



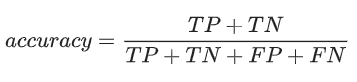
Истинный Позитив (TP): Это относится к случаям, в которых мы предсказывали «ДА и наш прогноз был на самом деле ПРАВДА  
(Например, у пациента на самом деле диабет, и вы предсказали, что это правда)

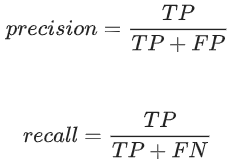
Правда отрицательный (TN): это относится к случаям, в которых мы предсказывали «НЕТ и наш прогноз был на самом деле ПРАВДА  
(Например, у пациента нет диабета, и вы предсказали то же самое.)

Ложно положительный (FP): это относится к случаям, в которых мы предсказывали «ДА», Но наш прогноз оказался ЛОЖНЫЙ  
(У пациента не было диабета, но наша модель предсказала, что он / она диабетик)

Ложный Отрицательный (FN): это относится к случаям, в которых мы предсказывали «НЕТ, но наш прогноз оказался ЛОЖНЫЙ

Для оценки качества работы алгоритма на каждом из классов считаем метрики accuracy(точноть), precision (точность) и recall (полнота).





Для этого пишем отдельную функцию, которая будет подсчитывать все четыре переменные из матрицы и потом считать метрики по формулам.

При прогонке на наших данных библиотечных алгоритмов и своих мы получили сравнимо высокие точности ответов, что говорит о корректности реализуемых нами алгоритмов и удачно подобранных данных в датасете.