Московский авиационный институт
(Национальный исследовательский университет)
Факультет прикладной математики и физики
Кафедра вычислительной математики и программирования

Курсовая работа

по курсу «Средства и технологии мультимедиа»

Студент: Чурсина Н.А.
Группа: 80-408б
Преподаватель: Вишняков Б. В.
Оценка:

Москва, 2021

1. Введение

С развитием компьютерных технологий все более актуальным становится вопрос совершенствования и упрощения работы с компьютерным кодом. Согласно книге [1], в настоящее время мы сталкиваемся с машинным обучением каждый день, даже не подозревая об этом. Сейчас компьютеры можно не просто программировать, но и настраивать их так, чтобы они обучались сами. Данная технология основана на оценке ранее сделанных действий и предположении, какие бы действия совершил человек в данной ситуации. При этом существуют критерии, по которым можно понять, насколько правильно принял решение алгоритм. Так как алгоритмов машинного обучения огромное множество, именно эти критерии и являются основным показателем, по которому алгоритмы и модели сравнивают между собой. Алгоритмы машинного обучения широко применяются в области компьютерного зрения, распознания речи, фильтрации спама - то есть там, где создание обычного алгоритма трудно или даже невозможно.

Машинное обучение это одна из областей искусственного интеллекта, занимающаяся алгоритмами, способными автоматически улучшаться на основе опыта или данных, также называемых обучающими данными, чтобы самостоятельно принимать решения или предсказания.

В данной работе рассматривается один из таких алгоритмов - алгоритм логистической регрессии - для классификации грибов на съедобные и несъедобные.

2. Постановка задачи

Требуется обучить алгоритм логистической регрессии и с использованием обученной модели предсказать выбранный признак из тестовых данных.

3. Описание данных

Каждый фрагмент данных из набора содержит описание формы, текстуры, цвета и других характеристик гриба (рис.1).

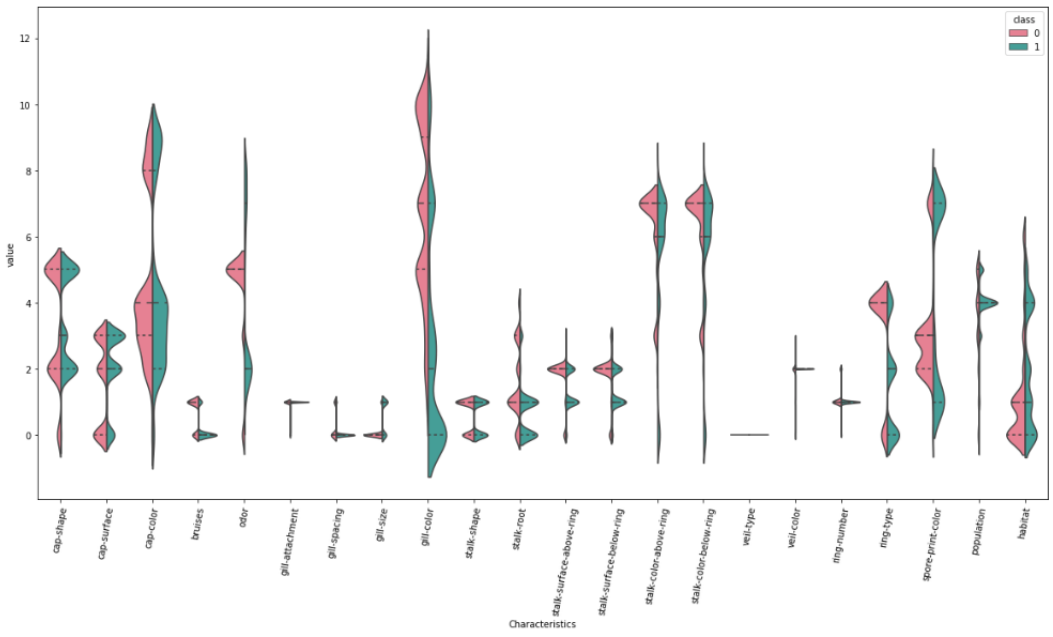

Рис. 1 – распределение данных по характеристикам

Тестовая выборка содержит 8124 записи с описанием 22 характеристик грибов двух классов: 4208 съедобных (e) и 3916 несъедобных (p) грибов.

4. Описание алгоритма

Логистическая регрессия используется для задач классификации. В частности, логистическая регрессия используется для предсказания классов задач с бинарной классификацией из-за особенности применяемой функции – она трансформирует выход линейной регрессии из области $(-\infty, +\infty)$ в область $(-1,1)$ (единицы могут достигаться скорее из-за плавающей точки вычислений).

При вычислении минимизируется следующая функция стоимости:

$$\min_{w,c} \frac{1}{2} w^T w + C \sum_{i=1}^{n} \log(\exp(-y_i(X_i^T w + c)) + 1)$$

5. Препроцессинг данных

В выбранном датасете нет пропущенных данных, согласно [2] missing = 0%. Значит, представляется возможным стандартизировать значения в промежуток [0;1]. Данные по корреляции по признакам представлены на рис.2.

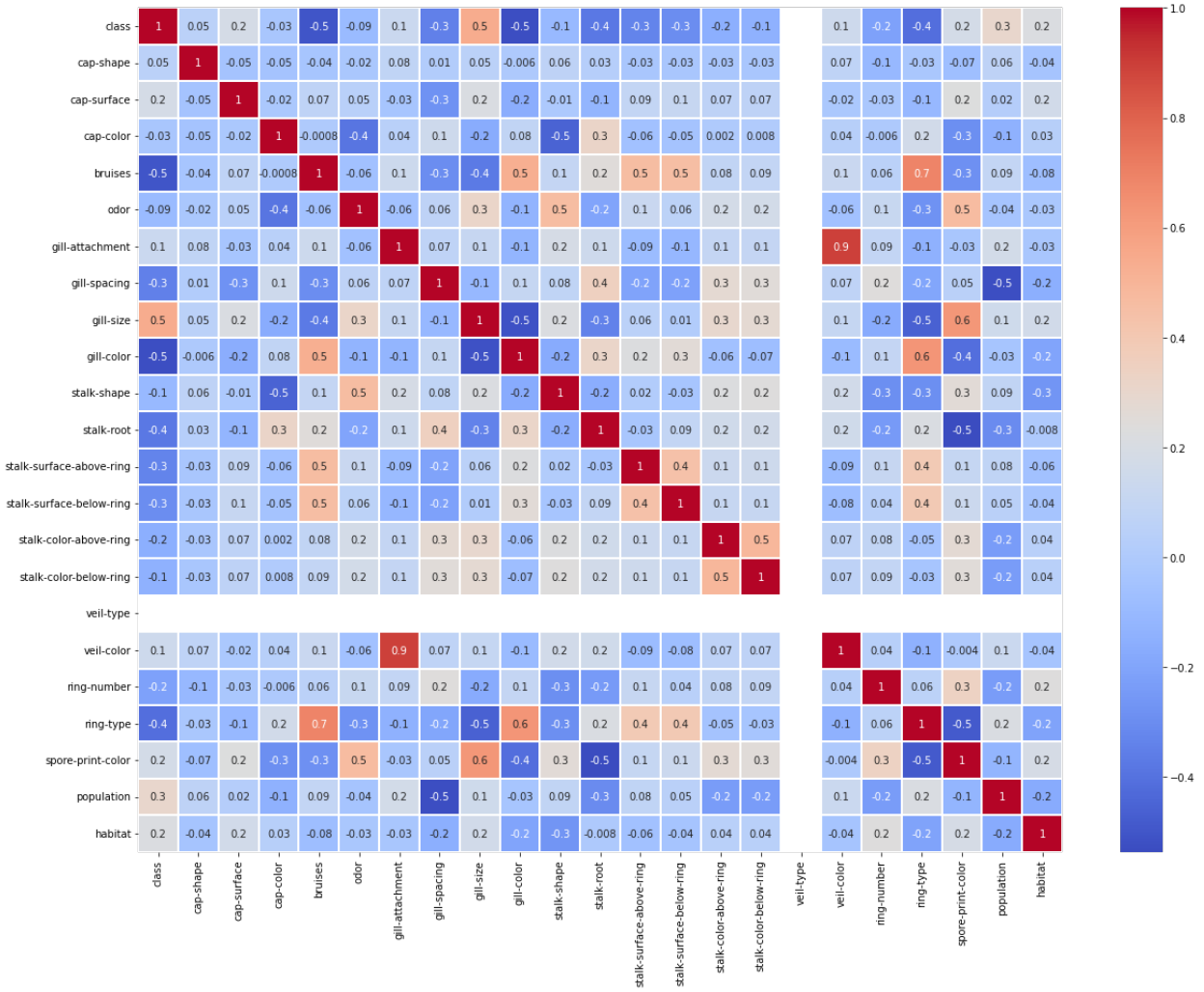

Рис. 2 – корреляция по характеристикам

Таким образом, получим положительно, средне или отрицательно коррелирующие значения. Если в нашем исследование возьмем максимально тесно связанные коррелирующие значения, то получим максимально точные значения и точно обученную модель.

6. Метрика качества

При обучении будут использоваться стандартные метрики:

- **Precision** (*точность*) – количество правильно классифицированных положительных предметов из выбранных для классификации:
$$precision = \frac{TP}{TP + FP}$$
- **Recall** (*полнота*) – количество правильно классифицированных положительных предметов из всех возможных (т.е. множество правильно и неправильно классифицированных предметов):
$$recall = \frac{TP}{TP + FN}$$

TP – количество корректно классифицированных положительных предметов, FP - количество некорректно классифицированных отрицательных предметов, TP - количество некорректно классифицированных положительных предметов. Можно также использовать **accuracy** (количество правильно классифицированных предметов из всех доступных), но её точность малозначима, так как классов неравное количество.

7. Полученные результаты

Sklearn LR

- Precision = 0,96
- Recall = 0,95

LR в собственной реализации

- Precision = 0,95
- Recall = 0,92

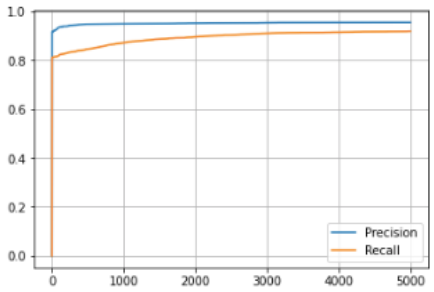

Рис.3 – график метрик для собственного алгоритма

При “ручной” реализации алгоритмов машинного обучения нужно использовать максимально эффективные алгоритмы, чтобы достичь таких же высоких результатов, какие предоставляют нам библиотечные реализации. При выбранных параметрах алгоритмов видно, что модель не переобучалась, поскольку разница в точности классификации на обучающей и тестовой выборках в основном достаточно мала.

8. Источники

[1] П. Домингос Верховный алгоритм. Как машинное обучение изменит наш мир.-М.: Манн, Иванов и Фербер,2016.-336 с.

[2] https://www.kaggle.com/uciml/mushroom-classification