

Федеральное агентство связи
Федеральное государственное бюджетное образовательное учреждение
высшего образования «Сибирский государственный университет
телекоммуникаций и информатики»
(СибГУТИ)

Лабораторная работа №3

Выполнил: студент IV курса ИВТ,
гр. ИП-713

Михеев Н.А.

Проверила: ассистент кафедры ПМиК
Морозова К.И.

Новосибирск, 2020 г.

Цель

Набор данных содержит в себе информацию о вариантах португальского вина "Винью Верде". Входные переменные представляют собой 13 столбцов со значениями, полученными на основе физико-химических тестов, а именно:

0 – цвет вина (“red” / ”white”)

1 - фиксированная кислотность

2 - летучая кислотность

3 - лимонная кислота

4 - остаточный сахар

5 - хлориды

6 - свободный диоксид серы

7 - общий диоксид серы

8 - плотность

9 - pH

10 - сульфаты

11 - спирт

Выходная переменная (на основе сенсорных данных):

12 - качество (оценка от 0 до 10, целое число)

Данные необходимо рассматривать как три набора. Данные для красного вина, данные для белого, общие данные вне зависимости от цвета. Необходимо построить модель для каждого из наборов, обучить её и сравнить полученные при помощи модели результаты с известными. Для обучения использовать 70% выборки, для тестирования 30%. Разбивать необходимо случайным образом, а, следовательно, для корректности тестирования качества модели, эксперимент необходимо провести не менее 10 раз и вычислить среднее значение качества регрессии.

Результат работы

Для выполнения данного задания была разработана программа с использованием библиотеки `sklearn`. Были считаны данные, обнулены все ячейки данных в которых есть неопределенные значения.

Было сделано 3 разделения данных в которые вошли сначала только белые, потом только красные и уже затем все вина.

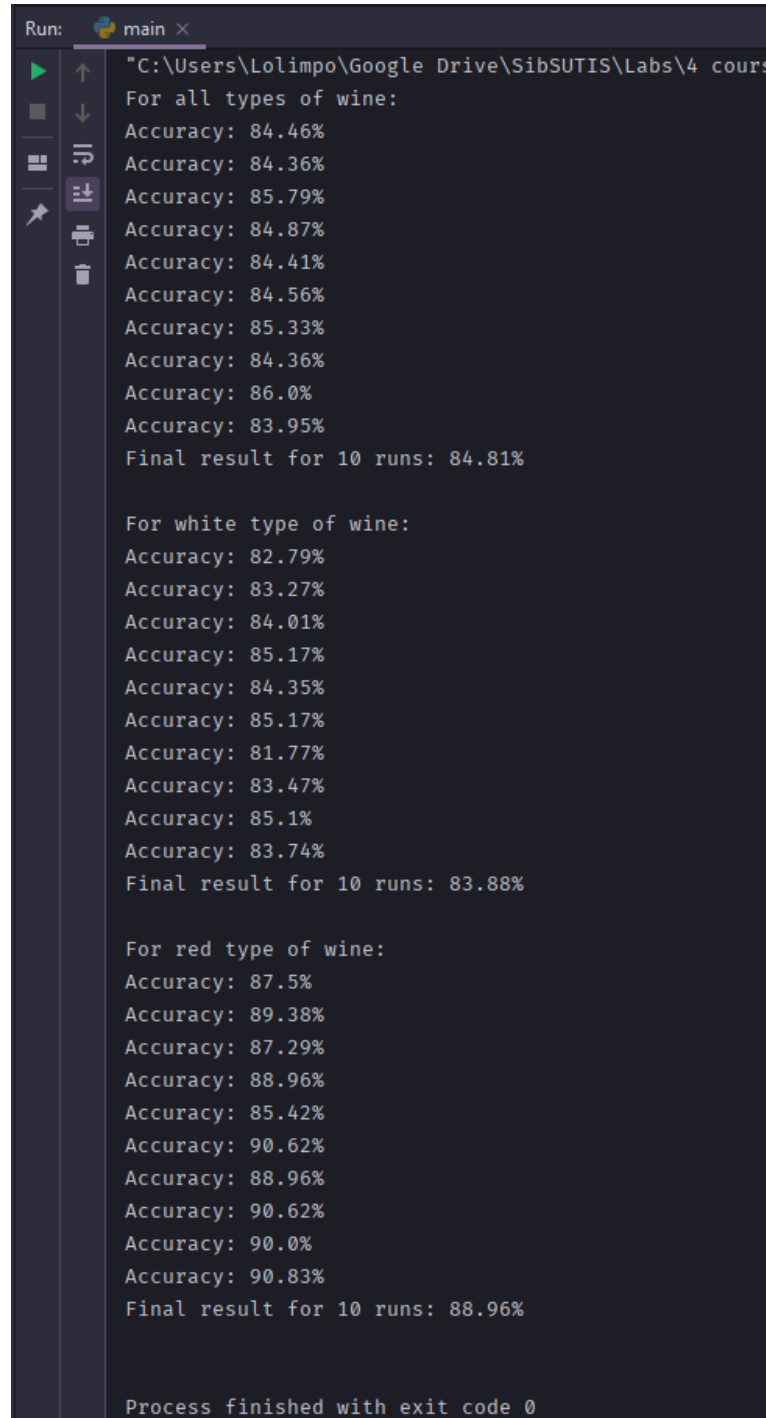
Далее в программе 3 раза был запущен цикл с разными разделениями, в котором идет разбиение на обучающую и тестовую выборки по 0.7 и 0.3 от общего количества данных соответственно. Инициализируется модель линейной регрессии `LASSO`, модель обучается с помощью выборки, идет

предсказание по $x_{\text{тест}}$ и далее подсчитывается точность предсказания. Наша модель Lasso с использованием Кросс-Валидации для автоматического определения коэффициента α .

За 10 прогонов точность модели для всех вин не опускалась ниже 83%, средняя за 10 – 84,81%;

Для белых вин не опускалась ниже 80% и средняя за 10 – 83.88%;

Для красных вин не опускалась ниже 85% и средняя за 10 – 88.96%.



```
Run: main x
"C:\Users\Lolimpo\Google Drive\SibSUTIS\Labs\4 course
For all types of wine:
Accuracy: 84.46%
Accuracy: 84.36%
Accuracy: 85.79%
Accuracy: 84.87%
Accuracy: 84.41%
Accuracy: 84.56%
Accuracy: 85.33%
Accuracy: 84.36%
Accuracy: 86.0%
Accuracy: 83.95%
Final result for 10 runs: 84.81%

For white type of wine:
Accuracy: 82.79%
Accuracy: 83.27%
Accuracy: 84.01%
Accuracy: 85.17%
Accuracy: 84.35%
Accuracy: 85.17%
Accuracy: 81.77%
Accuracy: 83.47%
Accuracy: 85.1%
Accuracy: 83.74%
Final result for 10 runs: 83.88%

For red type of wine:
Accuracy: 87.5%
Accuracy: 89.38%
Accuracy: 87.29%
Accuracy: 88.96%
Accuracy: 85.42%
Accuracy: 90.62%
Accuracy: 88.96%
Accuracy: 90.62%
Accuracy: 90.0%
Accuracy: 90.83%
Final result for 10 runs: 88.96%

Process finished with exit code 0
```

Рис.1 – Демонстрация точности работы программы.

Листинг программы

```
from random import randint

import matplotlib.pyplot as plt
import pandas as pd
from sklearn.linear_model import Lasso, LassoCV
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import normalize

def main() -> int:
    white_size = 0

    data = pd.read_csv('winequalityN.csv', header=0).fillna(0).values
    for i in data:
        if i[0] == 'white':
            i[0] = 0
            white_size += 1
        else:
            i[0] = 1
    x = data[..., 0:12]
    y = data[..., 12]
    for i in range(len(x[0])):
        x[..., i] = normalize([x[..., i]])

    runs = 10
    print(f'For all types of wine:')
    result = 0
    for _ in range(runs):
        x_train, x_test, y_train, y_test = train_test_split(x, y,
test_size=0.3, random_state=randint(0, 10000))
        model = LassoCV()
        model.fit(x_train, y_train)
        prediction = model.predict(x_test)

        success = 0
        for i in range(len(x_test)):
            if abs(y_test[i] - prediction[i]) < 1:
                success += 1
        print(f'Accuracy: {success / len(x_test) * 100:.4}%')
        result += success / len(x_test) * 100

    print(f'Final result for {runs} runs: {result / runs:.4}%\n')

    x_1 = data[0:white_size, 0:12]
    y_1 = data[0:white_size, 12]
    for i in range(len(x_1[0])):
        x_1[..., i] = normalize([x_1[..., i]])
    print(f'For white type of wine:')
    result = 0
    for _ in range(runs):
        x_train, x_test, y_train, y_test = train_test_split(x_1, y_1,
test_size=0.3, random_state=randint(0, 10000))
        model = LassoCV()
        model.fit(x_train, y_train)
        prediction = model.predict(x_test)
```

```

        success = 0
        for i in range(len(x_test)):
            if abs(y_test[i] - prediction[i]) < 1:
                success += 1
        print(f'Accuracy: {success / len(x_test) * 100:.4}%')
        result += success / len(x_test) * 100

    print(f'Final result for {runs} runs: {result / runs:.4}%\n')

    x_2 = data[white_size:, 0:12]
    y_2 = data[white_size:, 12]
    for i in range(len(x_2[0])):
        x_2[:, i] = normalize([x_2[:, i]])
    print(f'For red type of wine:')
    result = 0
    for _ in range(runs):
        x_train, x_test, y_train, y_test = train_test_split(x_2, y_2,
test_size=0.3, random_state=randint(0, 10000))
        model = LassoCV()
        model.fit(x_train, y_train)
        prediction = model.predict(x_test)

        success = 0
        for i in range(len(x_test)):
            if abs(y_test[i] - prediction[i]) < 1:
                success += 1
        print(f'Accuracy: {success / len(x_test) * 100:.4}%')
        result += success / len(x_test) * 100

    print(f'Final result for {runs} runs: {result / runs:.4}%\n')

    return 0

if __name__ == '__main__':
    exit(main())

```