Федеральное государственное бюджетное образовательное учреждение высшего образования

«Сибирский государственный университет телекоммуникаций и информатики»

(СибГУТИ)

Кафедра прикладной математики и кибернетики

Отчёт по лабораторной работе №3

«Linear regression»

Выполнил:

студент группы ИП-113

Шпилев Д. И.

ФИО студента

Работу проверил: Дементьева К. И.

ФИО преподавателя

Новосибирск 2024 г.

# **Задание**

Целью данной лабораторной работы является разработка программы,

реализующей применение метода линейной регрессии к заданному набору

данных.

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)

Классы упорядочены и не сбалансированы (например, нормальных вин

гораздо больше, чем отличных или плохих). В предоставленных данных есть

пропуски и неточности. Задания выполняются согласно варианту. Чтобы

определить номер варианта, воспользуйтесь следующей формулой:

𝑁варианта = (𝑁по списку 𝑚𝑜𝑑 4) + 1

Варианты заданий:

1) Использовать классическую модель LinearRegression

2) Использовать модель LASSO

3) Использовать модель LARS

4) Использовать модель ElasticNet

Задание: 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.

Особенности работы с данными:

1) Данные разнотипные, поэтому необходимо все столбцы привести к

одному типу. Все данные должны быть вещественными числами. В

данных есть пропуски, а это означает, что при считывании они будут

записаны как NaN (либо произойдёт ошибка).

2) Результат работы модели будет тоже вещественным числом. Поэтому

для оценки качества работы модели, необходимо использовать не

прямое сравнение, а учитывать разницу между настоящим значением и

смоделированным.

3) Данные в столбцах имеют разную размерность. Поэтому необходимо их

нормализовать. Можно воспользоваться, например, методом

preprocessing.normalize().

𝑥𝑛𝑜𝑟𝑚 = 𝑥 – 𝑥𝑚𝑖𝑛/𝑥𝑚𝑎𝑥 – 𝑥𝑚𝑖𝑛 ∈ [0,1]

В качестве результата выполненной лабораторной работы должна быть

разработанная программа, решающая поставленную задачу и отчёт с

содержанием текста программы, краткими комментариями и результатами

работы программы.

# **Код программы**

import pandas as pd

from sklearn.model\_selection import train\_test\_split

from sklearn.linear\_model import LinearRegression

from sklearn.preprocessing import normalize

from sklearn.metrics import mean\_squared\_error

from sklearn.impute import SimpleImputer

import numpy as np

def impute(dataset):

    imputer = SimpleImputer(strategy="mean")

    return imputer.fit\_transform(dataset)

data = pd.read\_csv('winequalityN.csv')

for column in data.columns[1:]:

    data[column] = pd.to\_numeric(data[column], errors='coerce')

red\_wine = data[data['type'] == 'red']

white\_wine = data[data['type'] == 'white']

all\_wine = data

def build\_and\_evaluate\_model(wine\_data):

    X = wine\_data.iloc[:, 1:-1]

    y = wine\_data.iloc[:, -1]

    X = impute(X)

    X\_normalized = normalize(X, axis=0)

    mse\_list = []

    for \_ in range(10):

        X\_train, X\_test, y\_train, y\_test = train\_test\_split(X\_normalized, y, test\_size=0.3, random\_state=np.random.randint(1000))

        model = LinearRegression()

        model.fit(X\_train, y\_train)

        y\_pred = model.predict(X\_test)

        mse = mean\_squared\_error(y\_test, y\_pred)

        mse\_list.append(mse)

        print(f"Среднеквадратичная ошибка составляет: {mse}")

    return np.mean(mse\_list)

mse\_red = build\_and\_evaluate\_model(red\_wine)

mse\_white = build\_and\_evaluate\_model(white\_wine)

mse\_all = build\_and\_evaluate\_model(all\_wine)

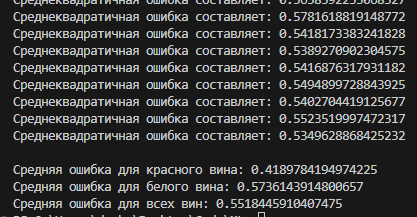
print()

print(f"Средняя ошибка для красного вина: {mse\_red}")

print(f"Средняя ошибка для белого вина: {mse\_white}")

print(f"Средняя ошибка для всех вин: {mse\_all}")

# **Результаты работы**



# **Выводы**

Средняя ошибка для оценки вина в среднем составляет 0.5 балла.