МИНОБРНАУКИ РОССИИ
САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ
ЭЛЕКТРОТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ
«ЛЭТИ» ИМ. В.И. УЛЬЯНОВА (ЛЕНИНА)
Кафедра МО ЭВМ

ОТЧЕТ

по лабораторной работе №7

по дисциплине «Машинное обучение»

Тема: Классификация (Байесовские методы, деревья)

Студент гр. 6304 _____ Иванов Д.В.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами кластеризации модуля Sklearn.

Загрузка данных

1. Датасет скачан и загружен в датафрейм.

```
import pandas as pd
import numpy as np
data = pd.read_csv('iris.data',header=None)
```

2. Выделены данные и их метки, метки преобразованы к числам.

```
X = data.iloc[:,:4].to_numpy()
labels = data.iloc[:,4].to_numpy()
le = preprocessing.LabelEncoder()
Y = le.fit_transform(labels)
```

3. Исходная выборка разбита на обучающую и тестовую.

```
X_train, X_test, y_train, y_test = train_test_split(X, Y,
test_size=0.5)
```

Байесовские методы

1. Проведена классификация наблюдений наивным байесовским методом (GaussianNB). Количество неправильно классифицированных наблюдений и точность классификации:

```
Wrong classified: 1
Score: 0.987
```

2. Построен график зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки.

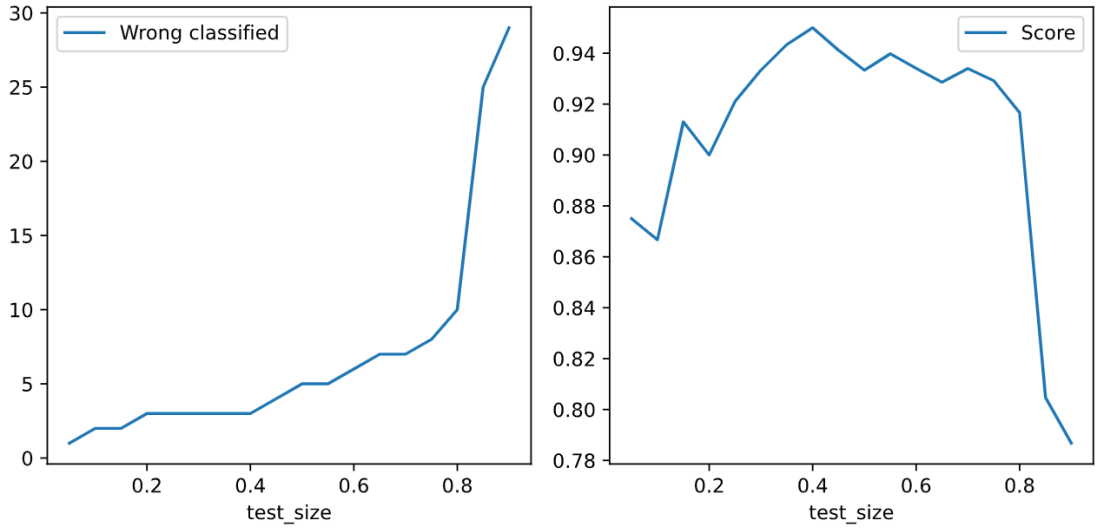

Рис. 1 — Графики зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки для метода GaussianNB

3. Проведена классификация методом MultinomialNB.

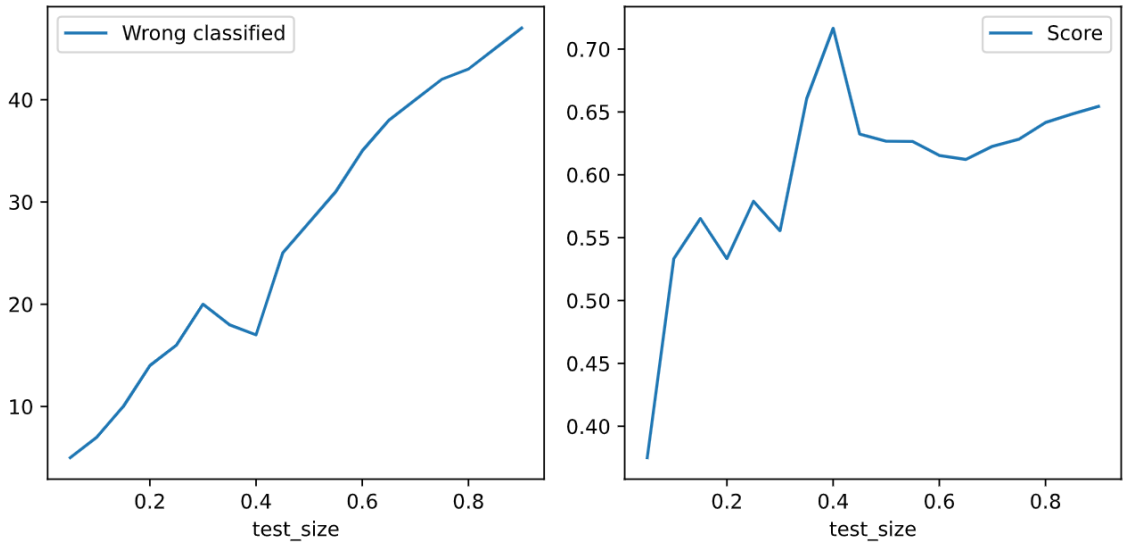

Рис. 2 — Графики зависимости для метода MultinomialNB

4. Проведена классификация методом ComplementNB.

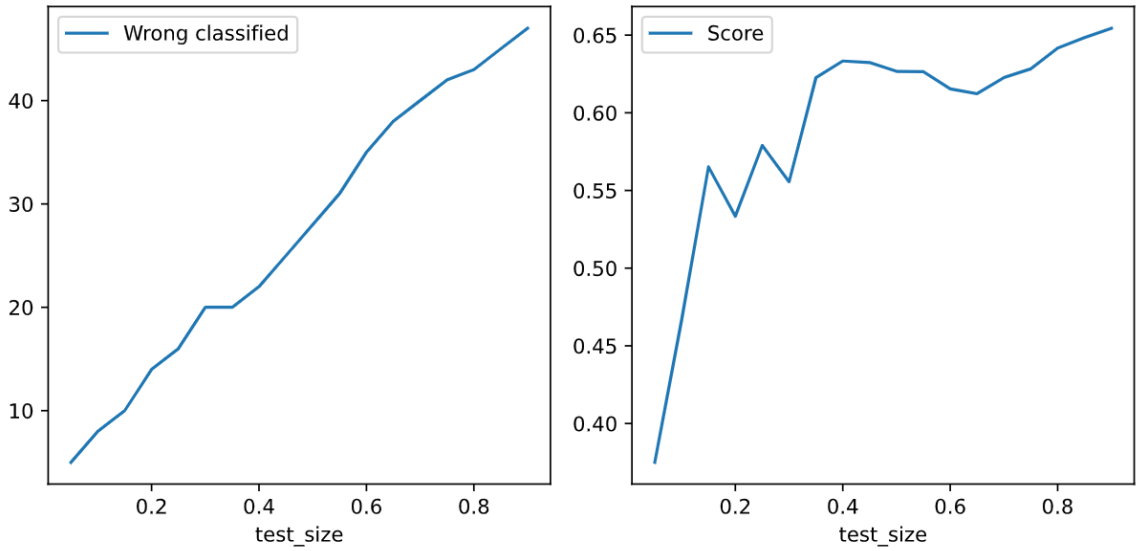

Рис. 3 — Графики зависимости для метода MultinomialNB

5. Проведена классификация методом BernoulliNB.

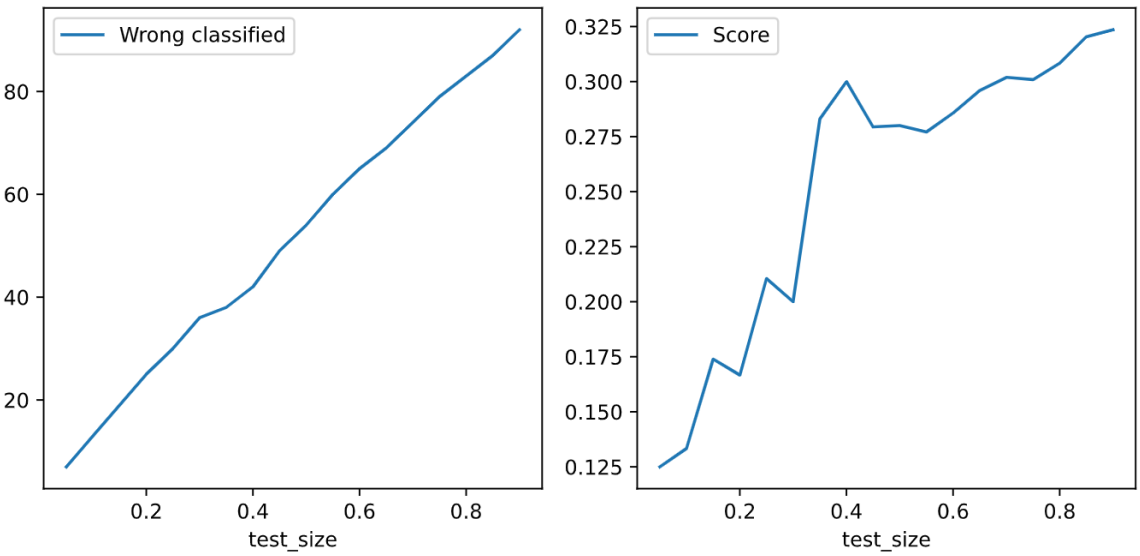

Рис. 4 — Графики зависимости для метода BernoulliNB

Классифицирующие деревья

1. Проведена классификация при помощи деревьев на тех же данных. Количество неверно классифицированных наблюдений, точность классификации, количество листьев дерева и глубина:

```
Wrong classified: 4
Score: 0.947
Num of leaves: 8
Depth: 5
```

2. Изображение построенного дерева:

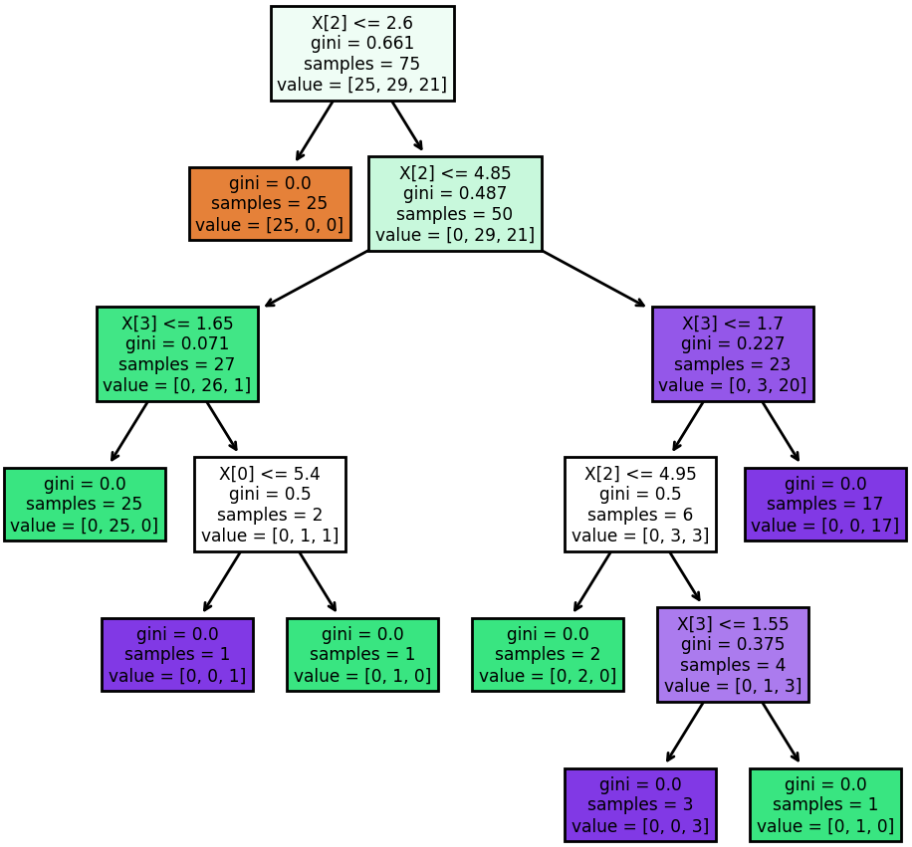

Рис. 5 — Изображение полученного дерева

3. Построен график зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки:

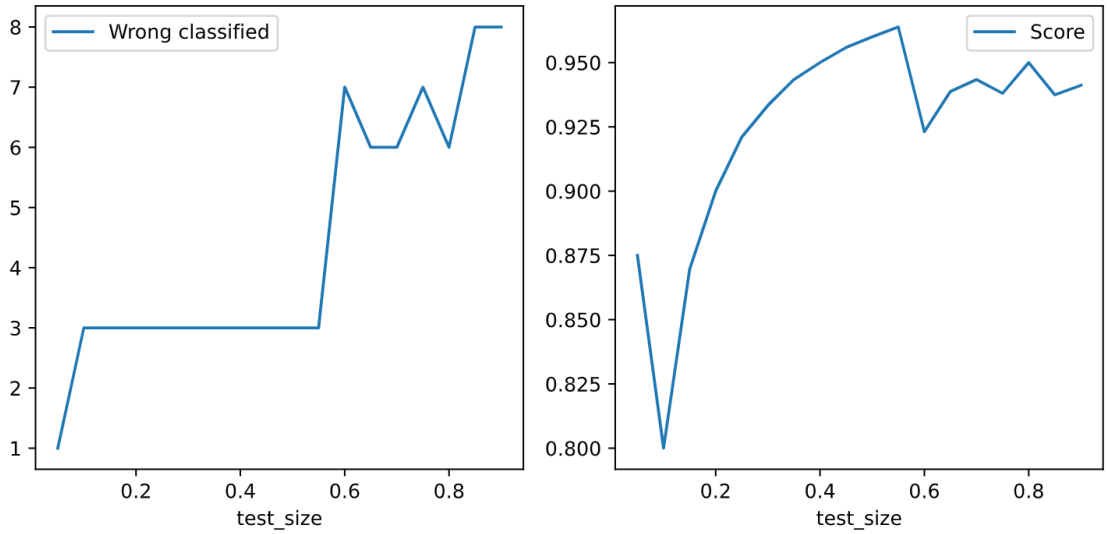

Рис. 6 — Графики зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки

4. Исследуйте работу классифицирующего дерева при различных параметрах DecisionTreeClassifier:

    a. *criterion* – функция измерения качества разбиения (энтропия или *индекс Джини*). На исходных данных для обоих критериев алгоритм даёт схожие результаты.

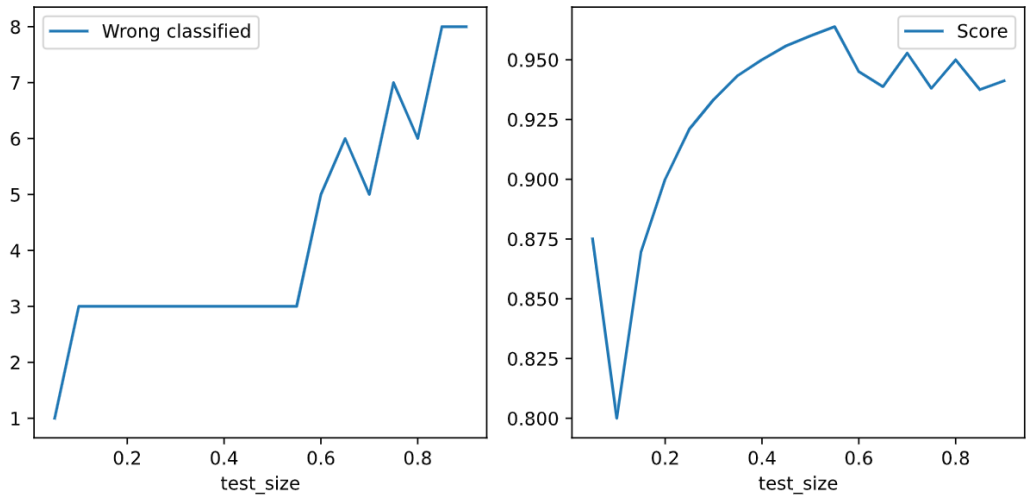

Рис. 7 — Графики зависимости от размера тестовой выборки для criterion = “entropy”

b. *splitter* – стратегия выбора разделения в узле (случайная или *лучшая*). На исходных данных случайная стратегия разделения нестабильна и ведет себя несколько хуже.

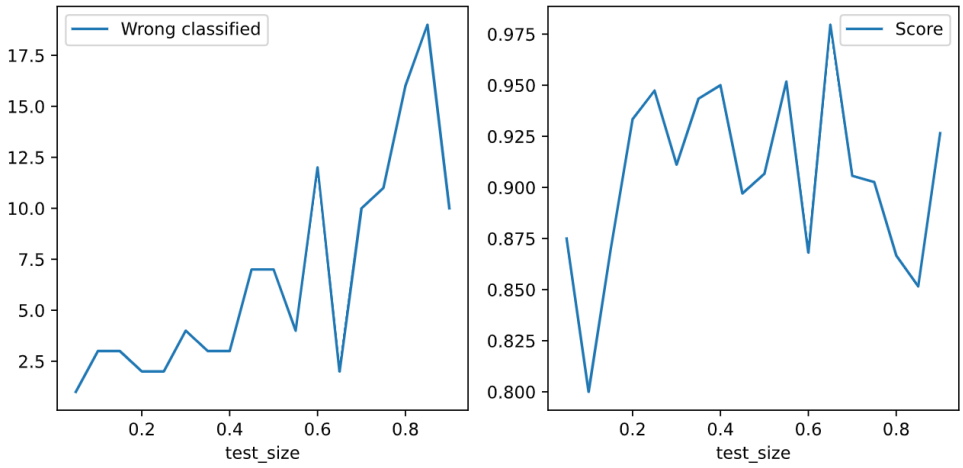

Рис. 8 — Графики зависимости от размера тестовой выборки для b. splitter = “random”

c. *max_depth* – максимальная глубина дерева. Ограничение максимальной глубины сильно снижает качество классификации.

max_depth	Wrong classified	Score
1	47	0.65
2	5	0.95
3	2	0.97
4	3	0.96
5	3	0.96

d. *min_samples_split* – минимальное число наблюдений для разбиения внутреннего узла. При ограничении снижается точность классификации.

min_samples_split	Wrong classified	Score

5	2	0.97
20	2	0.97
35	4	0.95
50	4	0.95
65	4	0.95
80	9	0.87
95	20	0.62

e. *min_samples_leaf* – минимальное число наблюдений для конечного узла. При ограничении снижается точность классификации.

min_samples_leaf	Wrong classified	Score
5	4	0.95
20	4	0.95
35	4	0.89
50	16	0.57
65	30	0.46
80	92	0.32
95	92	0.32

Вывод

В ходе выполнения лабораторной работы произведено знакомство с наивным байесовским классификатором и деревьями решений модуля Sklearn.