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 _____ Антонов С.А.

Преподаватель _____ Жангиров Т.Р.

Санкт-Петербург

2020

Цель работы:

Ознакомиться с методами классификации модуля Sklearn.

Ход работы:

Загрузка данных

1. На данном этапе был скачан и загружен датасет в датафрейм Pandas.

```
data = pd.read_csv('iris.data',header=None)
data.head()
```

	0	1	2	3	4
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa

Рисунок 1 Загруженный датасет

DBSCAN:

1. Так как признаки в выборке соответствуют разным шкалам, была проведена стандартизация данных.

```
data = np.array(data, dtype='float')
min_max_scaler = preprocessing.StandardScaler()
scaled_data = min_max_scaler.fit_transform(data))
```

2. Были выведены данные и их метки, тексты меток были преобразованы к числам с помощью LabelEncoder:

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

3. Выборка была разбита на обучающуюся и тестовую с помощью traint_test_split:

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

Байесовские методы

1. Была проведена классификация данных методом GaussianNB и выведено количество неправильно классифицированных наблюдений:

```
Wrong classified: 3
```

2. С помощью метода score была получена точность классификации, которая составляет 0.906%.
3. Были построены графики зависимости неправильно классифицированных на блюдений и точности классификации от размера тестовой выборки для мето да GaussianNB.

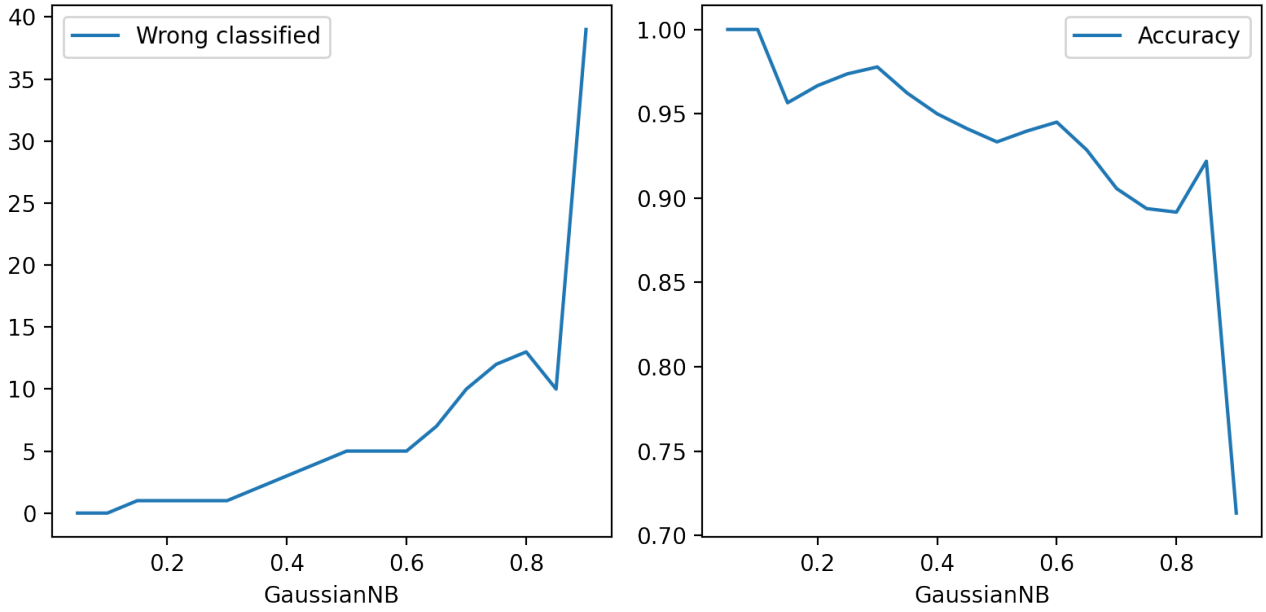

Рисунок 2 Графики для метода GaussianNB

Точность классификации не падает до 90% от всей выборки. Скорее всего такая хорошая классифицируемость связана с x распределением в выборке.

4. Была представлена классификация другими Байесовскими классификаторами, представленными в модуле Sklearn.

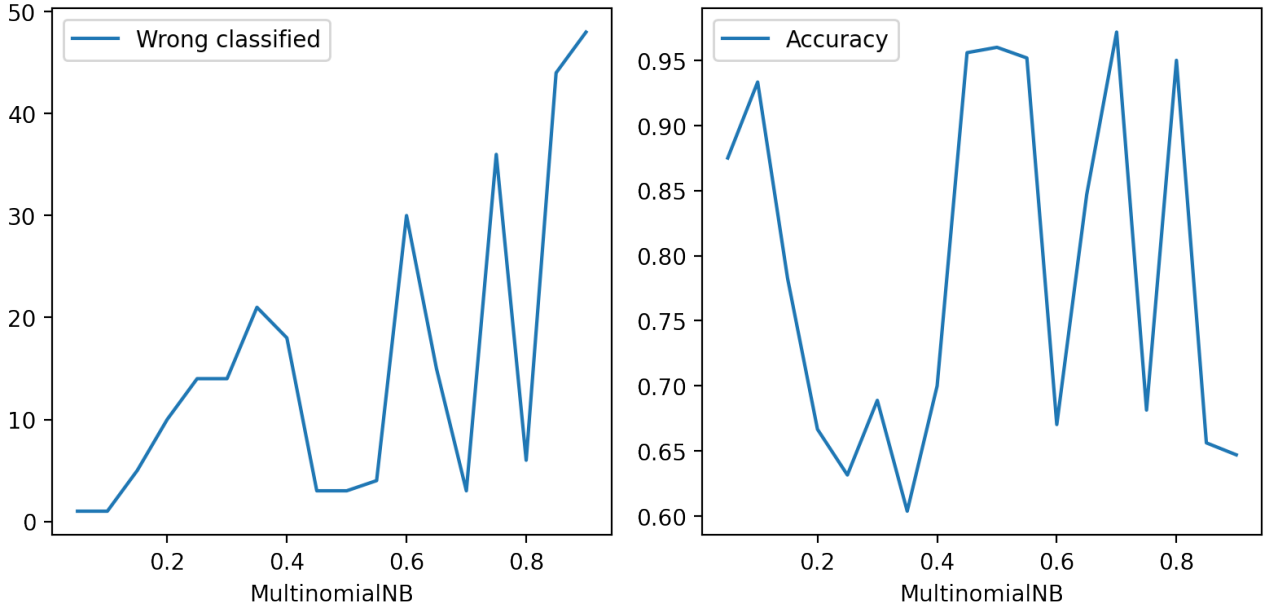

Рисунок 3 Графики для метода MultinominalNB

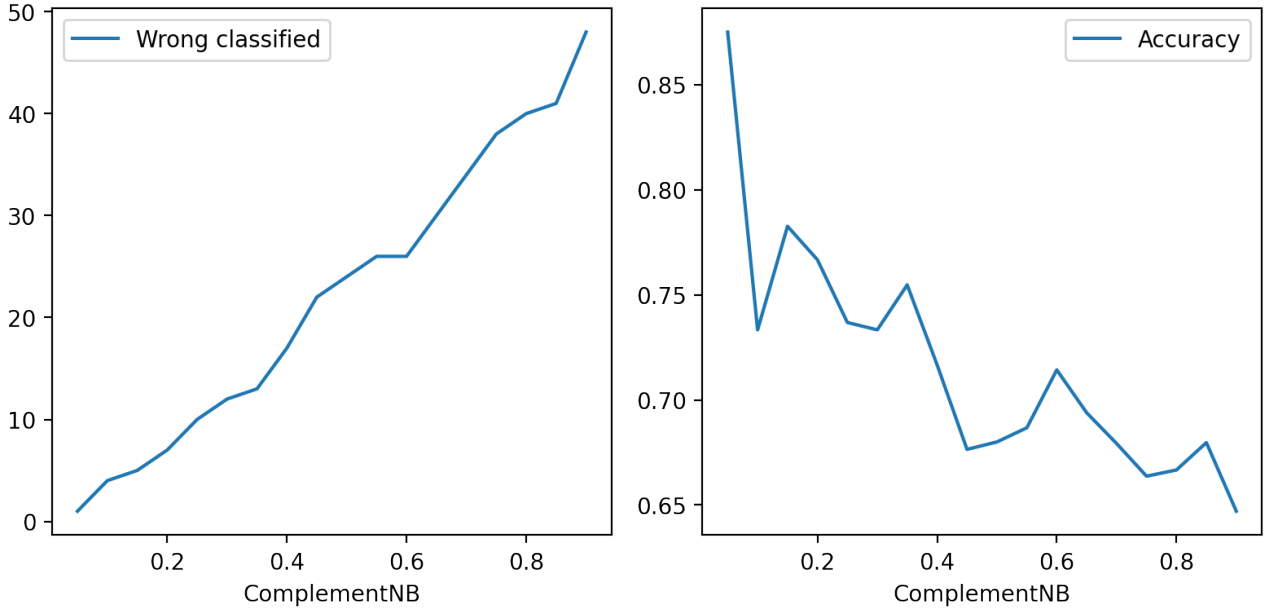

Рисунок 4 Графики для метода ComplementNB

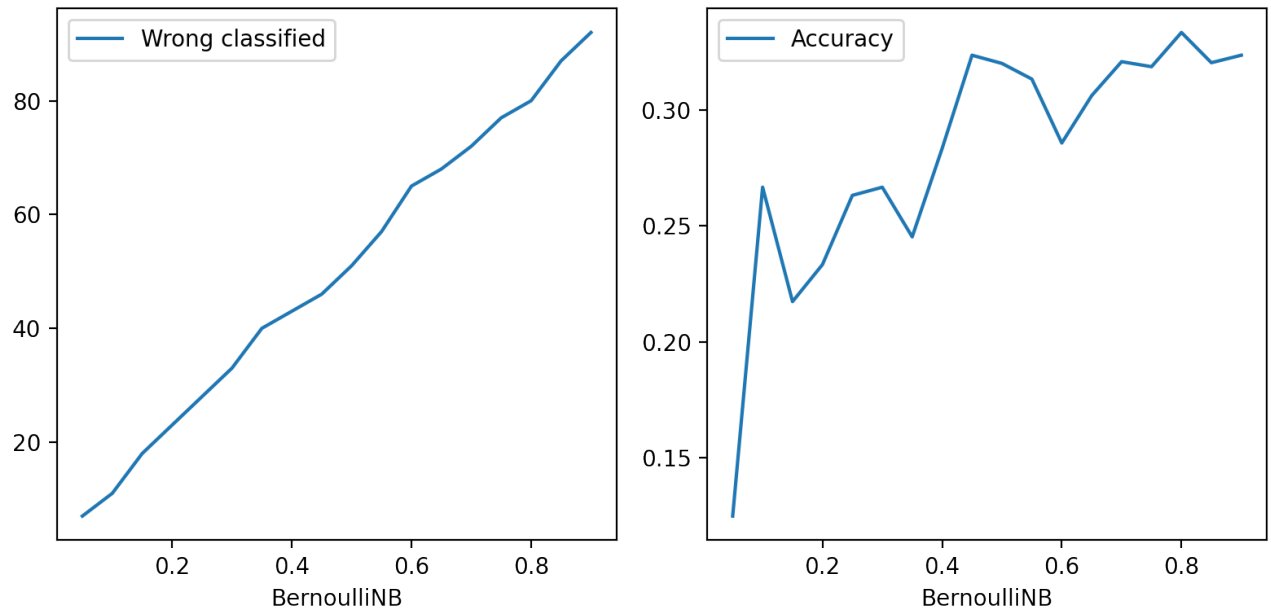

Рисунок 5 Графики для метода BernouliNB

Наилучший результат показал GaussianNB.

В методе MultinominalNB распределение для каждого класса параметризуется векторами, содержащими вероятности вхождения признаков в элемент выборки, соответствующей данному классу.

Метод ComplementNB – это адаптация стандартного полиномиального наивного байесовского алгоритма (MNB), который особенно подходит для несбалансированных наборов данных. В частности, CNB использует статистику з дополнений каждого класса для вычисления весов модели.

BernouliNB реализует наивные байесовские алгоритмы для данных, которые распределяются согласно многомерному распределению Бернулли. Предполагается, что каждый признак является двоичной (логической) переменной.

Классифицирующие деревья

1. Была проведена классификация данных методом DecisionTreeClassifier и выведено количество неправильно классифицированных наблюдений:

```
Wrong classified:  4
```

2. С помощью метода score была получена точность классификации, которая составляет 0.89%.

3. Были выведены количество листьев и глубина с помощью функции get_n_leaves и get_depth соответственно.

```
print('Num of leaves: ', clf.get_n_leaves())
print('Depth: ', clf.get_depth())
```

```
Num of leaves:  7
Depth:  6
```

4. Было выведено изображение полученного дерева (Представлено на рисунке 6).

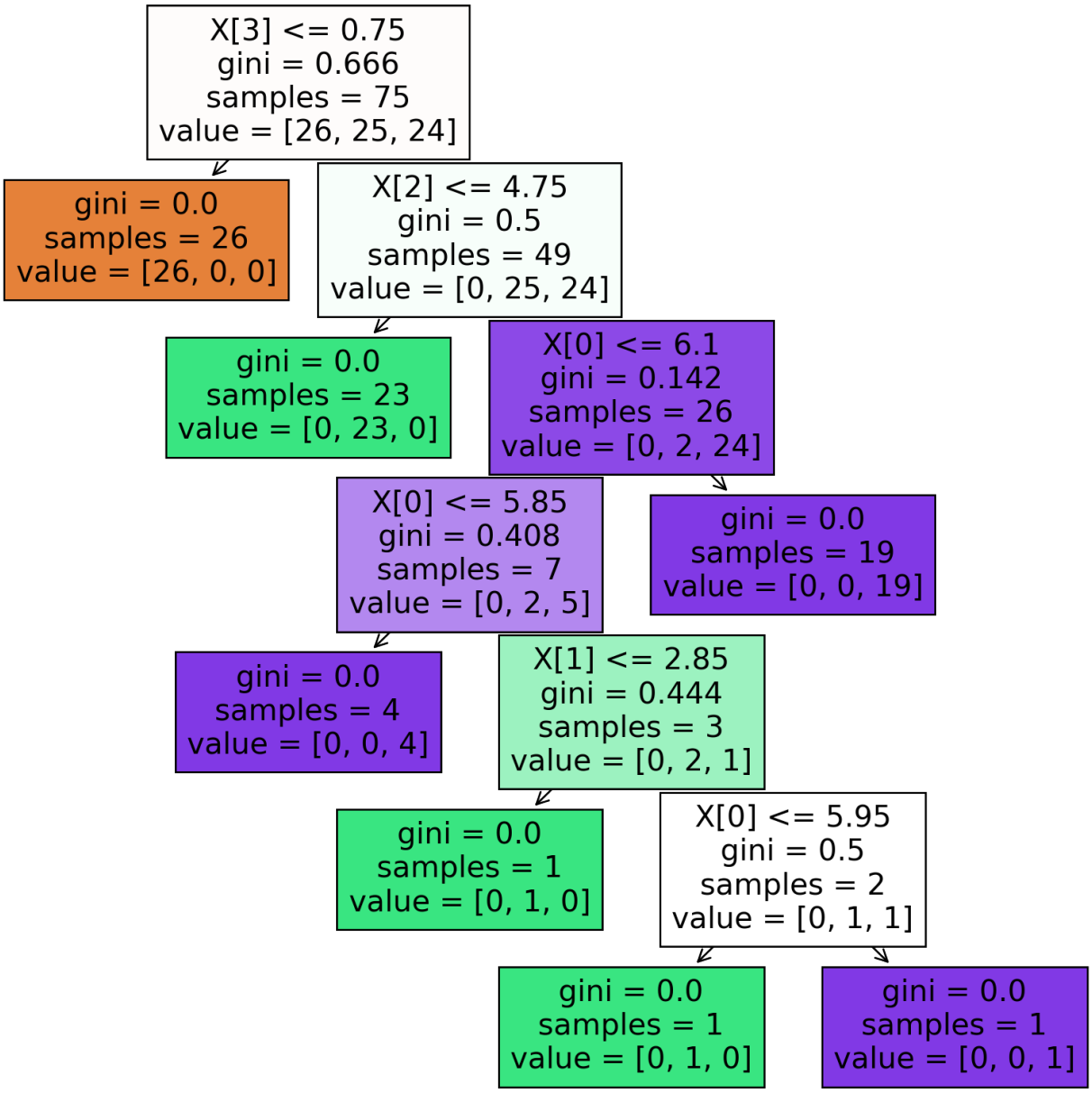

Рисунок 6 Изображение дерева

Для каждого узла на самой верхней строке указывается условие ля разбиения. Далее на каждом листе следует значение примеси Джини, количество наблюдений в узле/листе, а также распределение узлов по классам. Чем больше объектов в узле/листе принадлежит одному классу, тем насыщеннее его цвет.

5. Были построены графики зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки ля метода DecisionTreeClassifier. Графики представлены на рисунке 7.

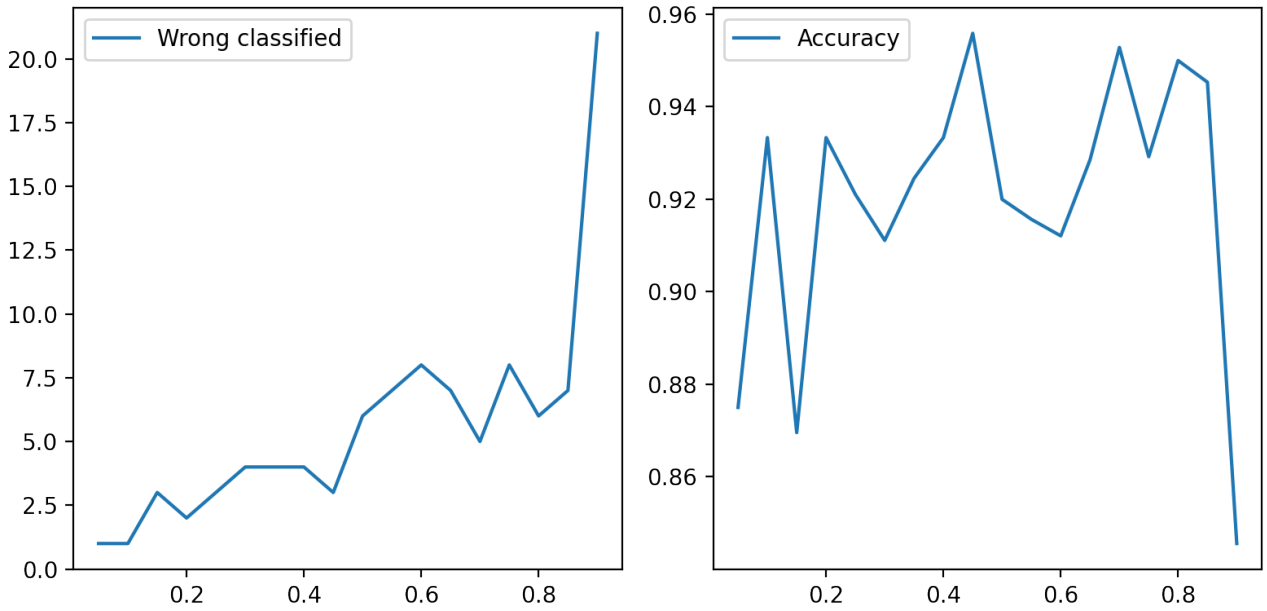

Рисунок 7 Графики для метода DecisionTreeClassifier

Слабая зависимость результатов классификации от размера тестовой выборки, как и в случае с байесовским классификатором подтверждает хорошую классифицируемость данных выборки.

6. Была исследована работа классифицирующего алгоритма при различных значениях параметров criterion, splitter, max_depth, min_samples_split, min_samples_leaf.
    a) Criterion – отвечает за функцию изменения качества разбиения. Критерием может быть или примесь Джини, или энтропия. Для обоих значений получились идентичные результаты классификации.

b) Splitter – отвечает за стратегию, используемую для выбора разделения в каждом узле. Можно выбрать и наилучшее разбиение, или наилучшее случайное разбиение. Результаты классификации при обоих параметрах равны.
c) Max_depth – Отвечает за максимальную глубину дерева. При значении 1 результат классификации заметно ухудшился, так как такой глубины недостаточно для классификации выборки. При значении 2 или выше были показаны идентичные результаты классификации.
d) Min_samples_split – Отвечает за минимальное число наблюдений необходимых для разбиения внутреннего узла. С увеличением значения наблюдается ухудшение классификации, однако оно не значительно виду того, что данные выборки хорошо классифицируемы. Также для классификации достаточно небольшого количества ровней дерева.
e) Min_samples_leaf – отвечает за минимальное число наблюдений для конечного узла. Рост значения сильно сказывается на результате классификации, так как параметр начинает сильно влиять на процесс разделения, заставляя оставлять в конечных узлах большее количество наблюдений.

Выводы:

В ходе выполнения лабораторной работы было произведено знакомство с классификацией методами GaussianNB, MultinominalNB, ComplementNB, BernoulliNB и DecisionTreeClassifier модуля Sklearn.