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 _____ Ястребков А. С.

Преподаватель _____ Жангиров Т. Р.

Санкт-Петербург
2020

Цель работы:

Ознакомиться с методами кластеризации модуля Sklearn.

Ход работы

Загрузка данных.

Был загружен датасет iris.data (фрагмент исходного датасета показан на рис. 1).

	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

Рис. 1. Фрагмент исходного датасета.

1. Датасет был разделён на данные и заголовки кластеров, причём заголовки преобразованы к числовому виду. Далее датасет был разбит на тестовую и обучающую выборки функцией train_test_split пакета sklearn. Код подготовки датасета показан в листинге 1.

Листинг 1 — Подготовка датасета к обучению

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

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

2. Была проведена классификация наивным байесовсем методом (классификатор GaussianNB) с параметрами по умолчанию. При этом были

неверно классифицированы 3 наблюдения. Метрика score показала точность классификации 96%.

Классификатор имеет следующий набор атрибутов:

- class_count_ — количество наблюдений в обучающих выборках для каждого класса;
- class_prior_ — вероятность встретить наблюдение каждого класса;
- classes_ — метки классов, полученных при классификации;
- epsilon_ — величина аддитивной дисперсии;
- sigma_ — дисперсия каждого признака по классу;
- theta_ — среднее каждого признака по классу.

3. Для байесовского классификатора были построены графики зависимости точности классификации и неверно классифицированных точек от размера тестовой выборки. График показан на рис. 2. Видно, что точность не падает вплоть до размера выборки 80%, что можно объяснить удачным распределением данных в выборке.

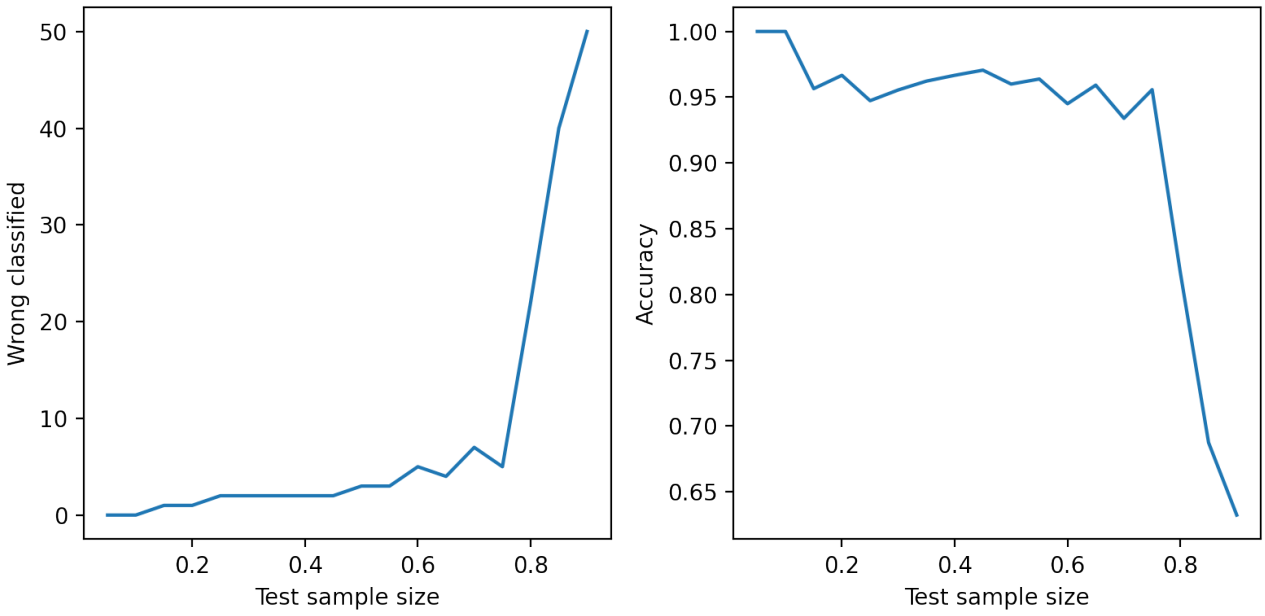

Рис. 2. График зависимости количества неверно классифицированных данных и точности от размера тестовой выборки.

4. Были изучены другие методы классификации:

- MultinomialNB — классификатор, в котором распределение для каждого класса параметризуется векторами, содержащими вероятности вхождения признаков в элемент выборки, соответствующий данному классу.
- ComplementNB — адаптация полиномиального наивного байесовского метода для несбалансированных наборов данных. Использует статистику из дополнения каждого класса для вычисления весов модели.
- BernoulliNB — оптимизация наивного байесовского метода для данных, распределённых по многомерному распределению Бернулли, применяется для случаев, когда каждый признак является логическим значением.

Для всех методов (включая GaussianNB) было проведено исследование поведения, результаты которого сведены в таблицу 1. Наилучшие результаты показал метод GaussianNB, однако его результаты слишком хороши. Вероятно, это обсуловлено неудачным значением random_state для разделения на тестовую и обучающую выборку.

Таблица 1 — Лучшие показатели различных классификаторов

Метод	**Размер тестовой выборки**	**Неверно классифицированных точек**	**Точность**
GaussianNB	0.1	0	1
MultinomialNB	0.4	1	0.984
ComplementNB	0.9	26	0.809
BernoulliNB	0.6	61	0.329

5. Была проведена классификация тех же данных с помощью алгоритма дерева решений. Были получены следующие результаты (размер тестовой выборки — 0.5):

- неверно классифицировано: 4 точки;
- score: 0.947;
- количество листьев: 8;

- глубина: 5.

6. На рис. 3 показано изображение полученного дерева решений. Каждый узел и лист содержат следующие данные:

- первая строка узла — условие для разбиения;
- значение примеси Джини (вероятность того, что случайно выбранная метрика будет неверно классифицирована, произведение вероятности выбора элемента на вероятность его ошибочной классификации);
- число наблюдений в узле/листе;
- распределение узла/листа по классам, чем выше значения, тем ярче цвет.

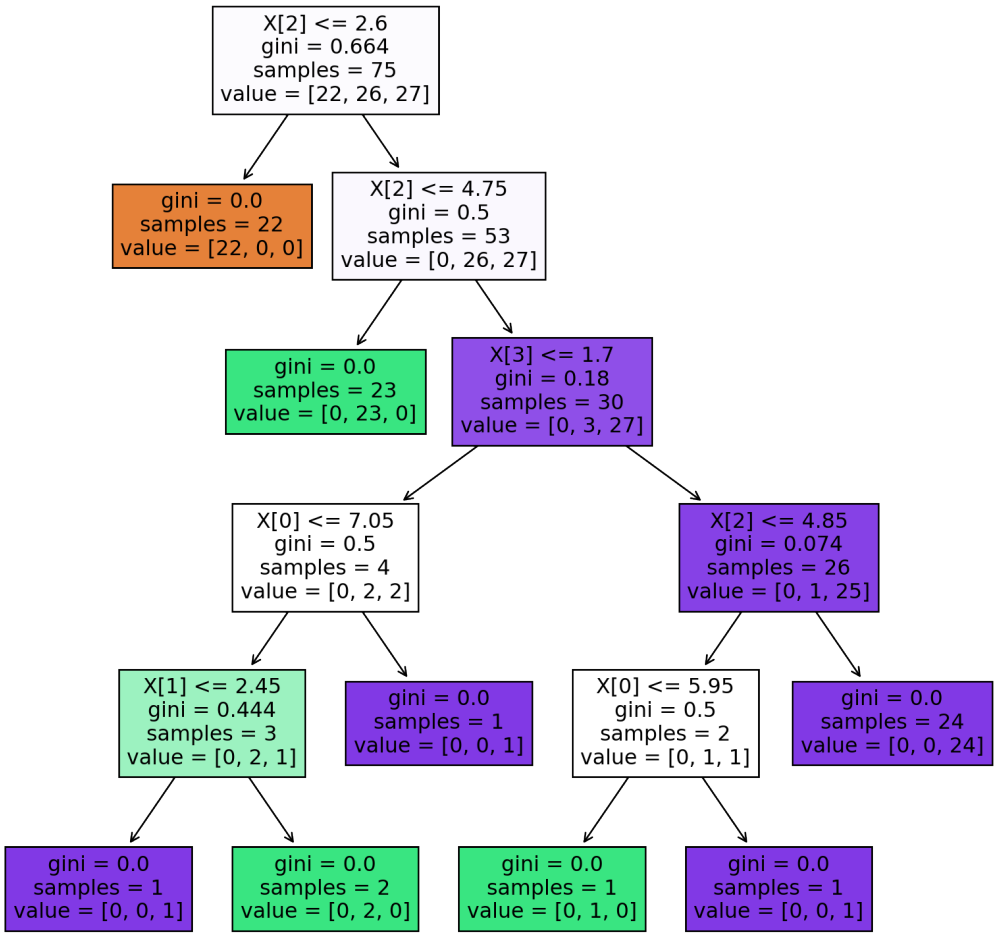

Рис. 3. Визуализация построенного дерева решений.

7. Был построен график зависимости числа неверно классифицированных точек и точности от размера тестовой выборки. График показан на рис. 4. Как и

в случае байесовского классификатора, при малых размерах тестовой выборки с заданным random_state имеет место слишком высокая точность. На рис. 5 показан график при случайном значении random_state, по нему можно судить о хорошей классифицируемости данных, поскольку нет зависимости между размером выборки и точностью классификации.

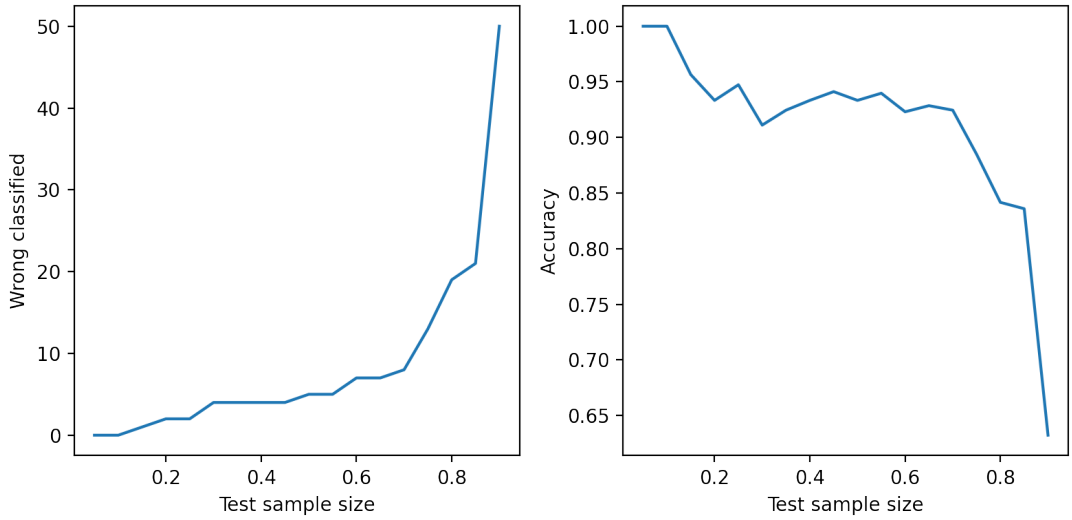

Рис. 4. График зависимости количества неверно классифицированных данных и точности от размера тестовой выборки дерева решений при заданном random_state.

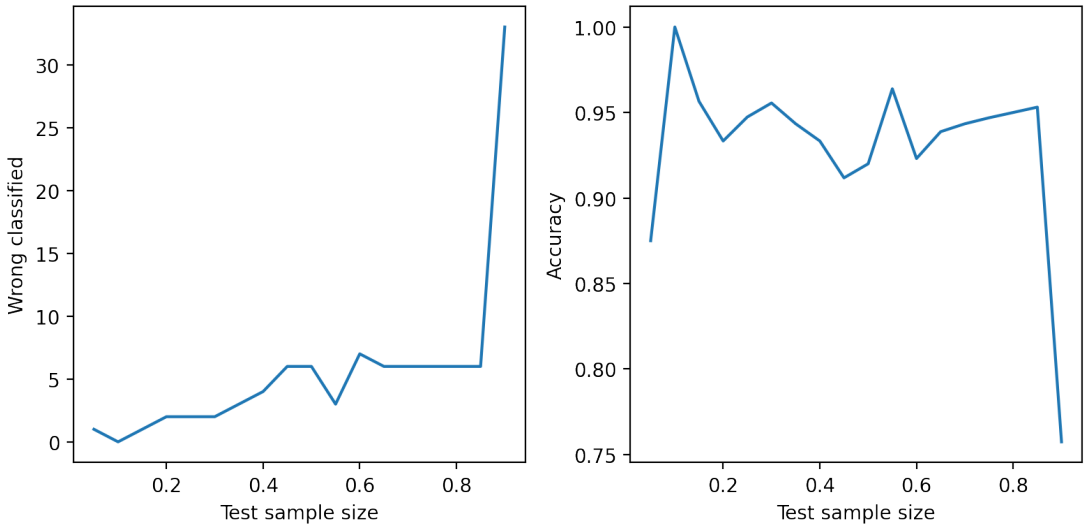

Рис. 5. График зависимости количества неверно классифицированных данных и точности от размера тестовой выборки дерева решений при случайном random_state.

8. Было исследовано влияние параметров дерева решений на результат классификации:

- criterion — функция измерения качества разбиения, примесь Джини или энтропия; не оказала влияния на качество классификации;
- splitter — стратегия выбора разделения в каждом узле, наилучшее разбиение или наилучшее случайное разбиение; не оказала влияния на качество классификации;
- max_depth — максимальная глубина дерева; при малых значениях (1-2) значительно ухудшается качество классификации, при значениях выше разница пропадает, график точности показан на рис. 6;

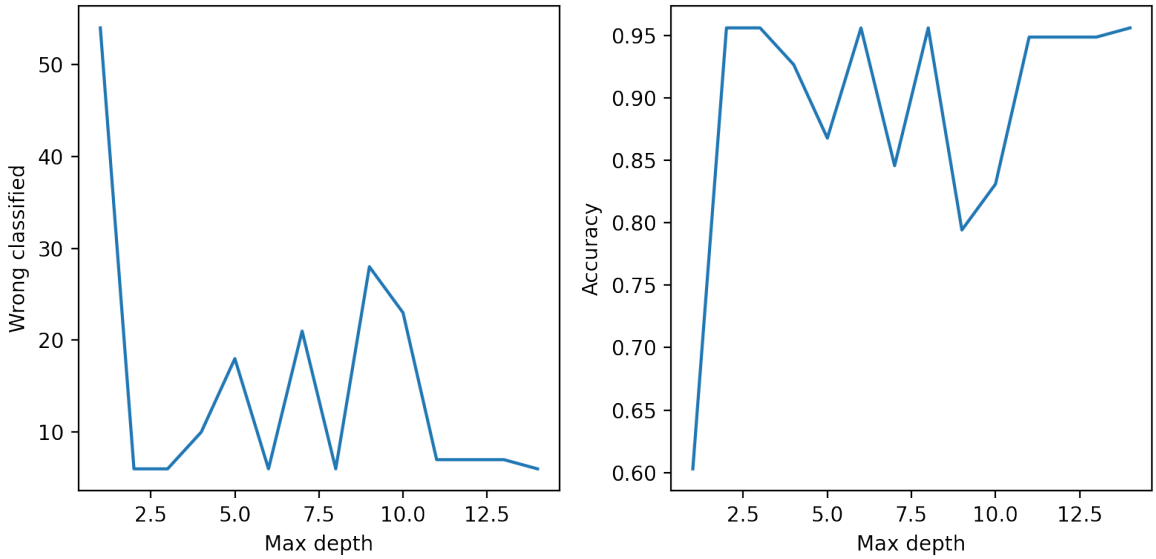

Рис. 6. График зависимости количества неверно классифицированных данных и точности от максимальной глубины дерева решений.

- min_samples_split — минимальное число наблюдений, необходимое для разбиения внутреннего узла; увеличение значения с определённого предела приводит к практически нулевой точности классификации, график показан на рис. 7.

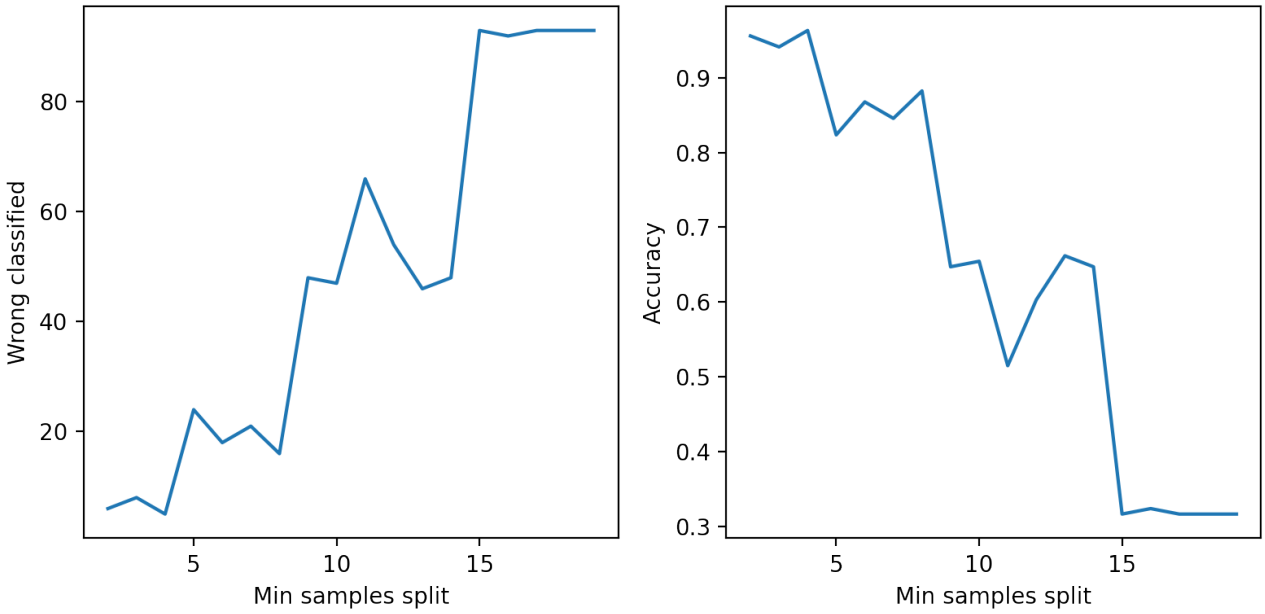

Рис. 7. График зависимости количества неверно классифицированных данных и точности от параметра min_samples_split.

- min_samples_leaf — минимальное число наблюдений для конечного узла; увеличение значения сильно уменьшает точность классификации, заставляя дерево оставлять слишком много данных в листьях, график точности показан на рис. 8.

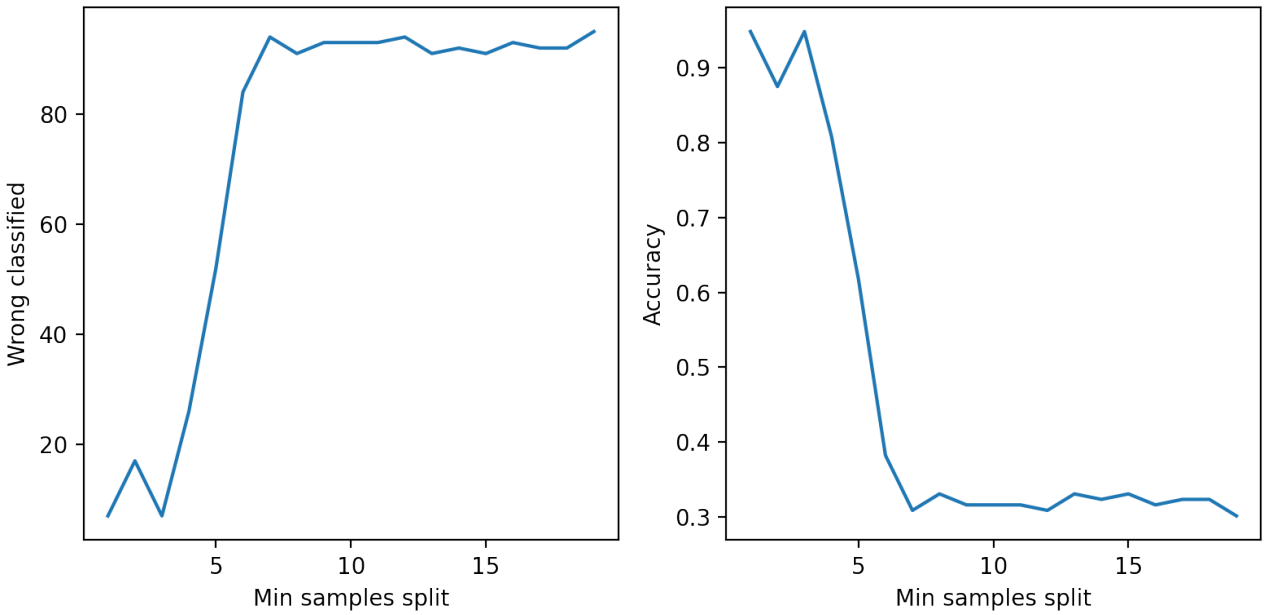

Рис. 8. График зависимости количества неверно классифицированных данных и точности от параметра min_samples_leaf.

Вывод:

В результате выполнения данной работы были изучены байесовские методы классификации и классификация деревом решений. Было проведено сравнение байесовских классификаторов и влияние размера тестовой выборки на точность классификации. Аналогичное исследование, а также влияние параметров самого классификатора было проведено для дерева решений.