

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

ОТЧЕТ
по лабораторной работе №7
по дисциплине «Машинное обучение»
Тема: Классификация (Байесовские методы, деревья)

Студент гр. 6304

Ковынев М.В.

Преподаватель

Жангиров Т.Р.

Санкт-Петербург

2020

Цель работы

Ознакомиться с методами классификации модуля *Sklearn*.

Ход работы

Загрузка данных

Датасет загружен в датафрейм. Вид данных представлен 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
...
145	6.7	3.0	5.2	2.3	Iris-virginica
146	6.3	2.5	5.0	1.9	Iris-virginica
147	6.5	3.0	5.2	2.0	Iris-virginica
148	6.2	3.4	5.4	2.3	Iris-virginica
149	5.9	3.0	5.1	1.8	Iris-virginica

150 rows × 5 columns

Рисунок 1 – Исходные данные

Выделены данные и их метки, тексты меток преобразованы к числам. Выборка разбита на обучающую и тестовую *train_test_split*.

Байесовские методы

1. Проведена классификация наблюдений наивным байесовским методом. Выявлено 4 неправильно классифицированных наблюдения. Атрибуты классификатора представлены в табл. 1.

Таблица 1 – Атрибуты *GaussianNB*

Атрибут	Описание
class_count_	Количество обучающих выборок, наблюдаемых в каждом классе
class_prior_	Вероятность каждого класса
classes_	Метки классов, известные классификатору
epsilon_	Абсолютная аддитивная величина дисперсий
sigma_	Дисперсия каждого признака по классу
theta_	Среднее каждого признака по классу

2. Точность классификации получена с помощью функции `score()` и составляет 97%.
3. Построен график зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки. График представлен на рис. 2.

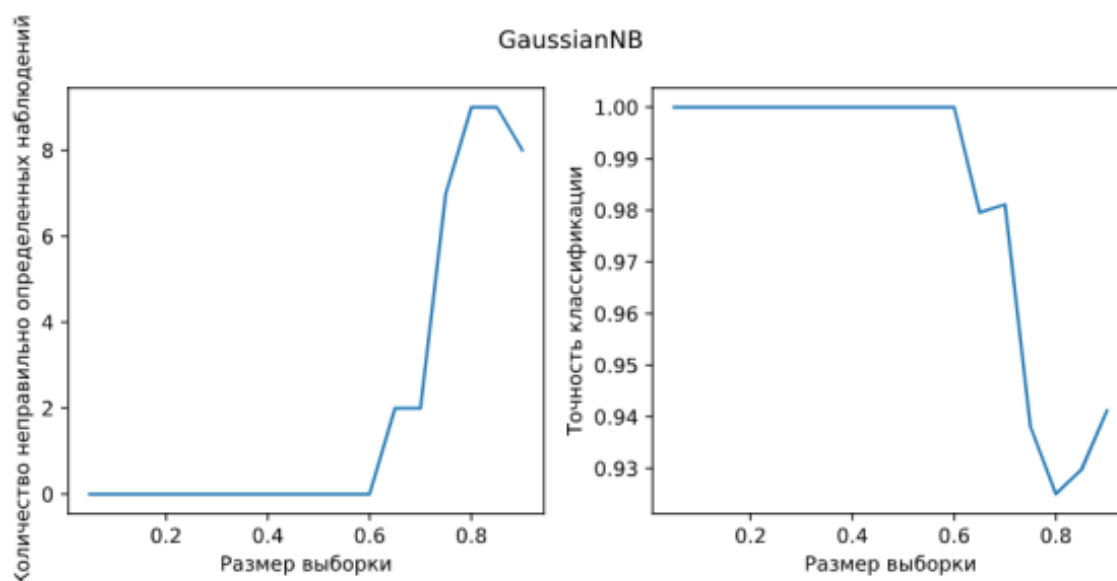


Рисунок 2 – Классификация *GaussianNB*

4. Классификация проведена с помощью *MultinomialNB*, *ComplementNB*, *BernoulliNB*. Результат представлен на рис. 3-5.

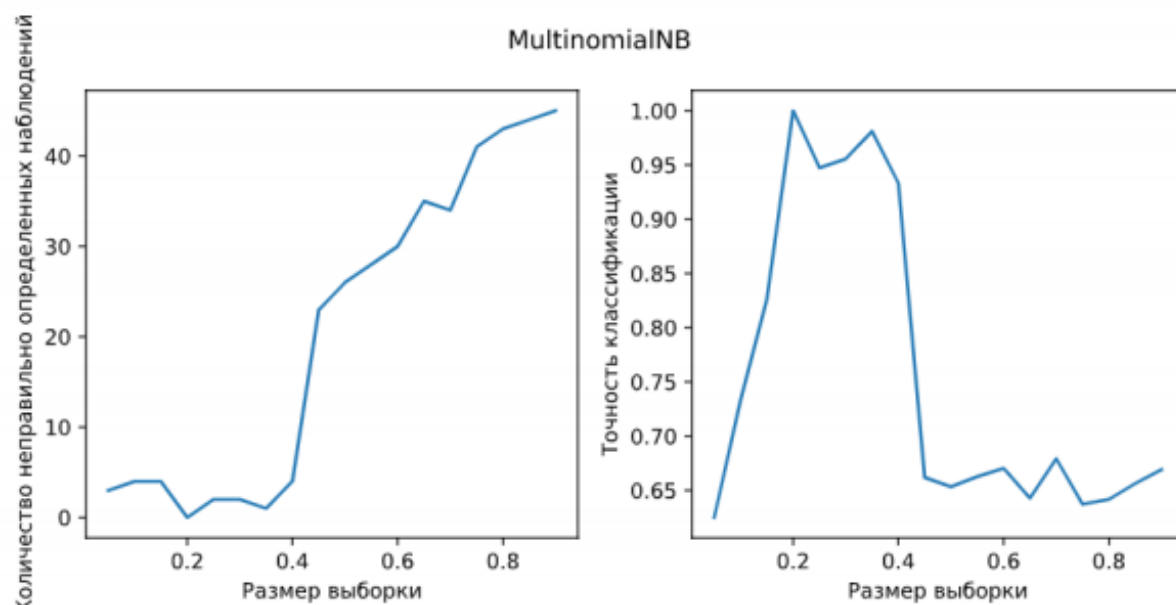


Рисунок 3 – Классификация *MultinomialNB*

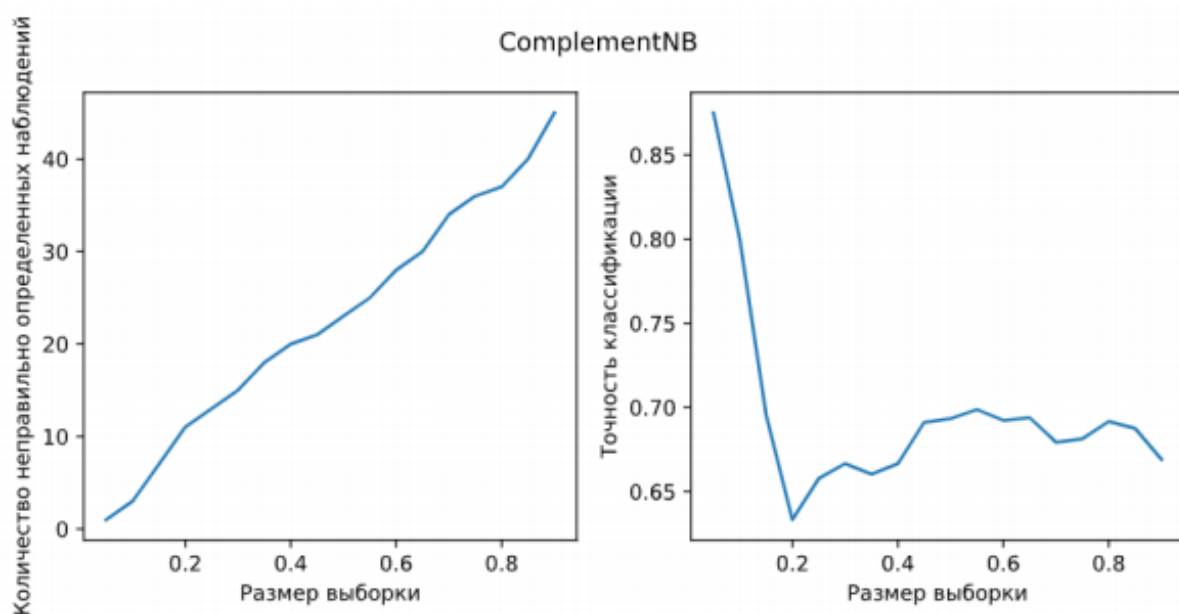


Рисунок 4 – Классификация *ComplementNB*

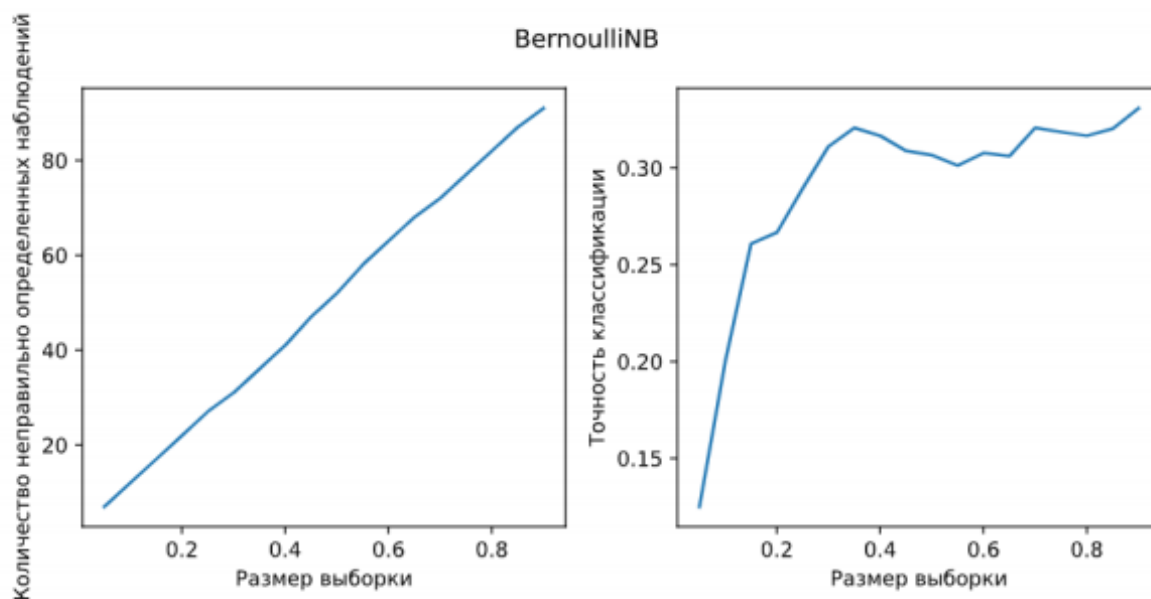


Рисунок 5 – Классификация *BernoulliNB*

MultinomialNB – полиномиальный наивный байесовский классификатор, подходит для классификации с дискретными признаками (например, подсчет слов для классификации текста). *MultinomialNB* реализует наивный алгоритм Байеса для полиномиально распределенных данных. Распределение для каждого класса параметризуется векторами, содержащими вероятности вхождения признаков в элемент выборки, соответствующий данному классу.

ComplementNB – адаптация *MultinomialNB*, подходит для несбалансированных наборов данных. В частности, CNB использует статистику из дополнения каждого класса для вычисления весов модели. *ComplementNB* часто превосходит *MultinomialNB* в задачах классификации текста.

BernoulliNB – как и *MultinomialNB*, этот классификатор подходит для дискретных данных. Разница в том, что в то время, как *MultinomialNB* работает с подсчетом вхождений, *BernoulliNB* предназначен для двоичных/логических признаков.

Классифицирующие деревья

1. Проведена классификация наблюдений с помощью деревьев решений на тех же данных. Выявлено 4 неправильно классифицированных наблюдения.
2. Точность классификации получена с помощью функции `score()` и составляет 100%.
3. Получившееся дерево имеет глубину, равную 3, и 4 листа.
4. Дерево продемонстрировано на рис. 6.

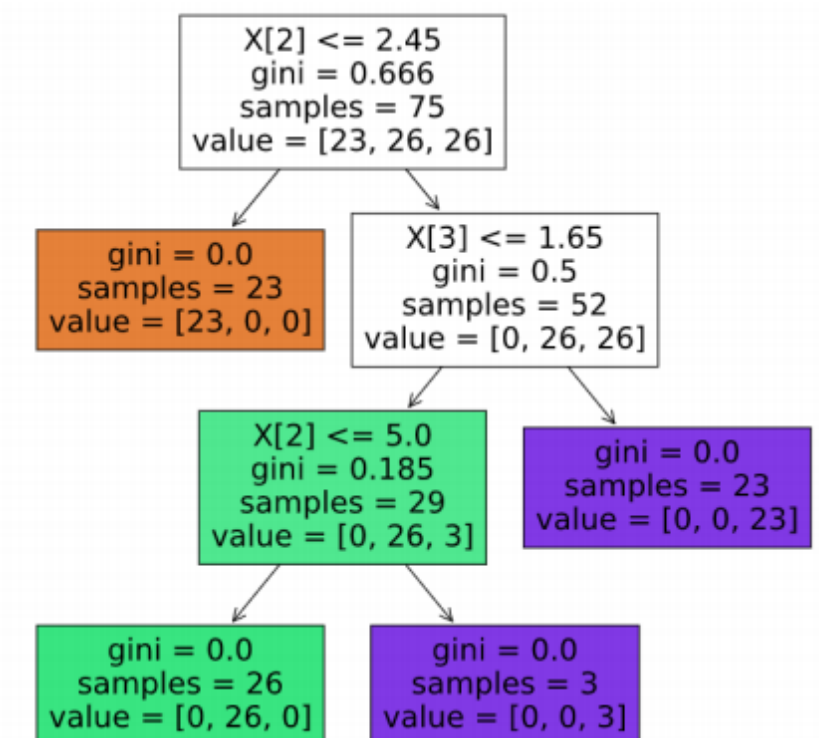


Рисунок 6 – Дерево решений для классификации

5. Построен график зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки. График представлен на рис. 7.

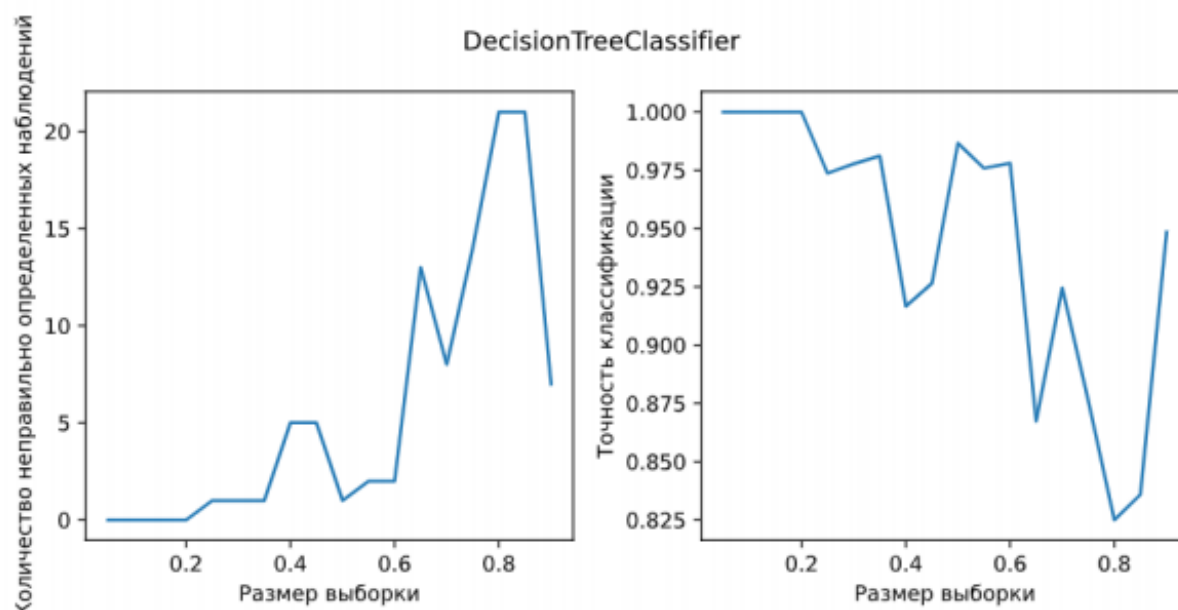


Рисунок 7 – Классификация *DecisionTreeClassifier*

6. Исследованы параметры *DecisionTreeClassifier*, результаты представлены в табл. 2. и на рис. 8.

Таблица 2 – Параметры *DecisionTreeClassifier*

Параметр	Описание
criterion	Функция измерения качества разбиения. Поддерживается индекс Джини и энтропия.
splitter	Стратегия, используемая для выбора разбиения на каждом узле. Поддерживается выбор наилучшего разбиения и случайный выбор.
max_depth	Максимальная глубина дерева. Если None, то узлы расширяются до тех пор, пока все листья не станут чистыми или пока все листья не будут содержать менее min_samples_split выборок.
min_samples_split	Минимальное количество выборок, необходимых для деления внутреннего узла.

min_samples_leaf	Минимальное количество выборок, которое требуется для конечного узла. Точка разделения на любой глубине будет учитываться только в том случае, если она оставляет не менее min_samples_leaf обучающих выборок в каждой из левой и правой ветвей.
------------------	--

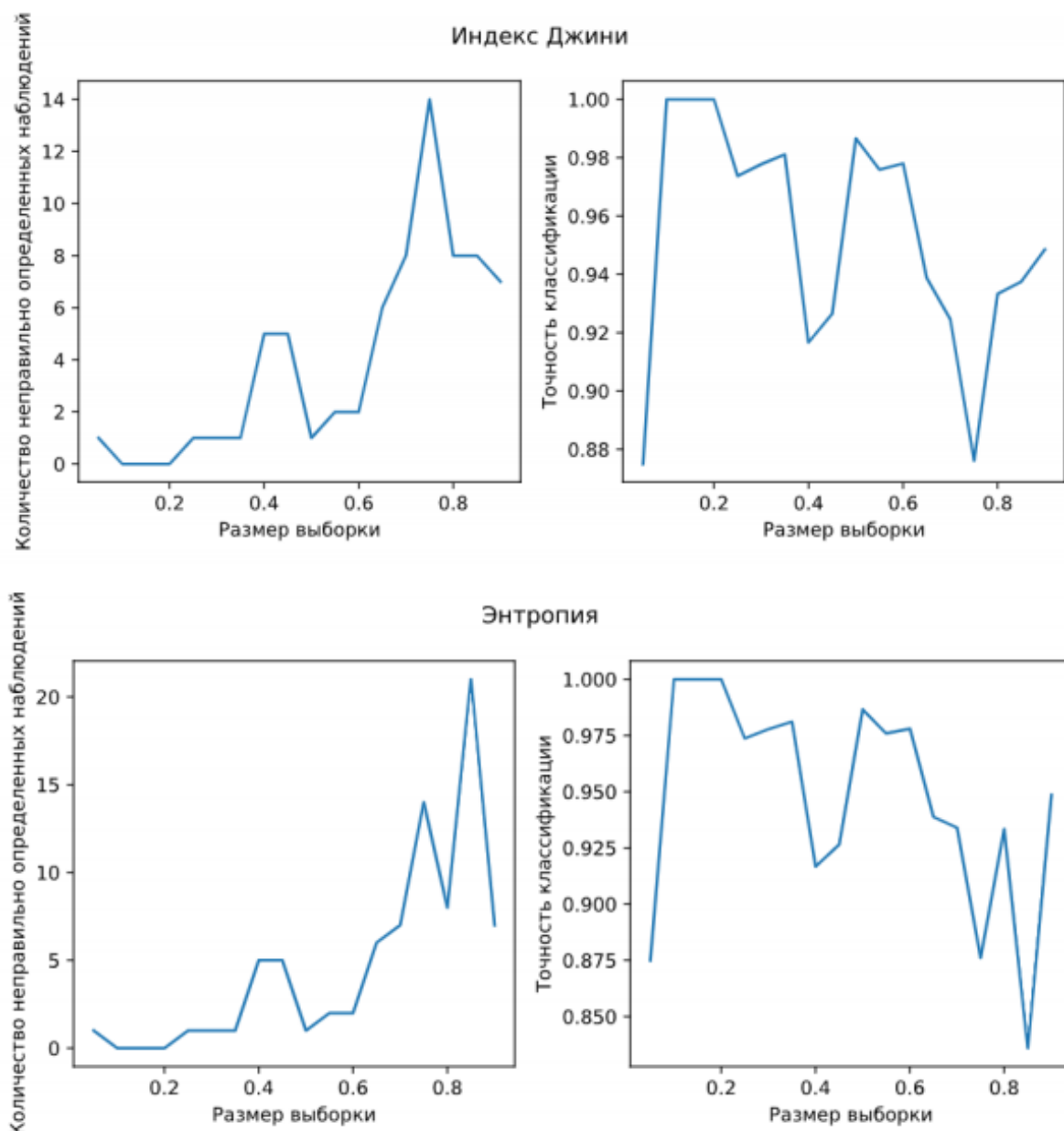


Рисунок 8 – Разные функции измерения качества разбиения

Выводы

В ходе лабораторной работы рассмотрены такие методы классификации модуля *Sklearn*, как *GaussianNB*, *MultinomialNB*, *ComplementNB*, *BernoulliNB* и *DecisionTreeClassifier*.