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. 8304 |  | Сергеев А.Д. |
| Преподаватель |  | Жангиров Т.Р. |

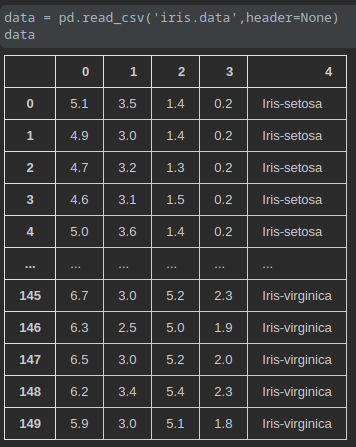
Санкт-Петербург 2021

# Цель работы

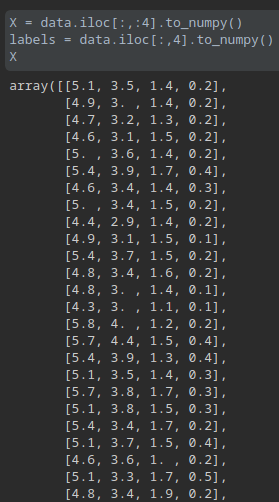
Ознакомиться с методами классификации модуля *Sklearn*.

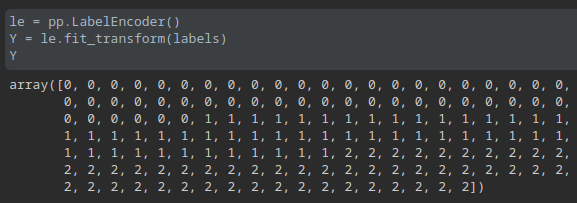
**Загрузка данных**

1. Данные загружены в датафрейм.

Рисунок 1 – Исходные данные

1. Выделены данные и их метки, тексты меток преобразованы к числам.

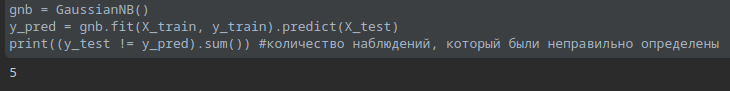
Рисунок 2 – Выделенные данные

Рисунок 3 — Метки, преобразованные к числам

1. Выборка разбита на обучающую и тестовую *train\_test\_split*.

**Байесовские методы**

1. Проведена классификация наблюдений наивным байесовским методом.

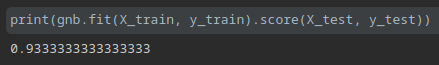
Рисунок 4 — Классификация наивным байесовским методом

Выявлено 5 неправильно классифицированных наблюдения.

Таблица 1 – Атрибуты *GaussianNB*

|  |  |
| --- | --- |
| **Атрибут** | **Описание** |
| class\_count\_ | Количество выборок каждого класса, участвующих в обучении |
| class\_prior\_ | Вероятность каждого класса в данных для обучения |
| classes\_ | Метки классов |
| n\_features\_in\_ | Количество признаков в данных обучения |
| theta\_ndarray of shape (n\_classes, n\_features) | Среднее значение каждого признака в классах |

1. Точность классификации получена с помощью функции score() и составляет 93.3%.

Рисунок 5 — Точность классификации

1. Построен график зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки (0.05 до 0.95 с шагом 0.05), random\_state = 830434.

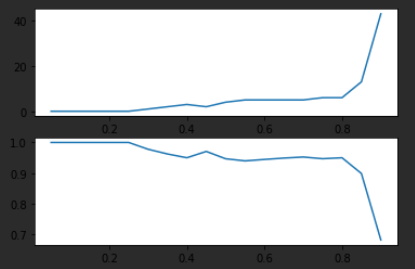


Рисунок 6 – График зависимости точности классификации и неправильно классифицированных наблюдений от размера тестовой выборки

Такой метод используется для данных, которые имеют непрерывную описательную функцию.

1. Классификация проведена с помощью *GaussianNB, MultinominalNB*, *ComplementNB*, *BernoulluNB*.

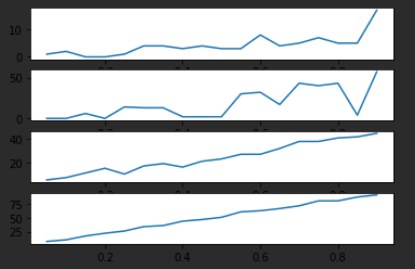


Рисунок 7 – Точность классификации различными методами

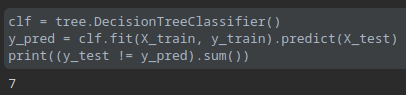
*MultinominalNB* – полиномиальный наивный байесовский классификатор, подходит для классификации с дискретными признаками (например, подсчет слов для классификации текста). *MultinominalNB* реализует наивный алгоритм Байеса для полиномиально распределенных данных. Распределение для каждого класса параметризируется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

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

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

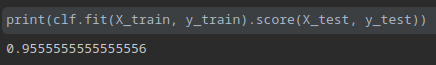
**Классифицирующие деревья**

1. Проведена классификация наблюдений с помощью деревьев решений на тех же данных.

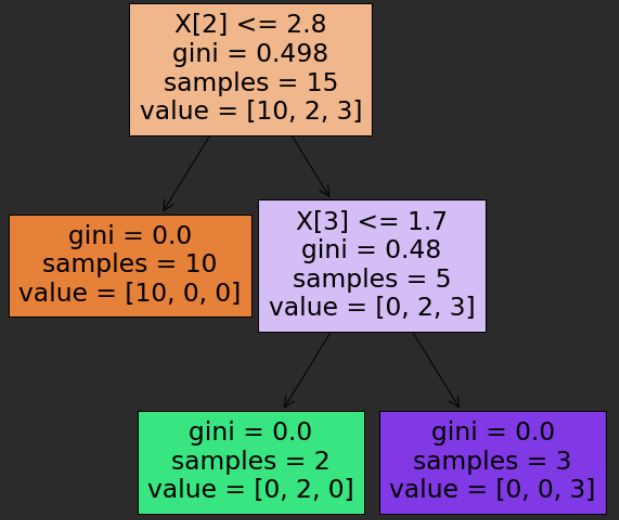
Рисунок 8 — Классификация с помощью деревьев решений

Выявлено 7 неправильно классифицированных наблюдения.

1. Точность классификации получена с помощью функции score() и составляет 95.6%.

Рисунок 9 — Точность классификации

1. Получившееся дерево имеет глубину 2, и 3 листа.

Рисунок 10 – Дерево решений для классификации

Узлом дерева является критерий, разделяющий данные на два подмножества. В узле содержится:

1) вопрос, на основании которого происходит разделение

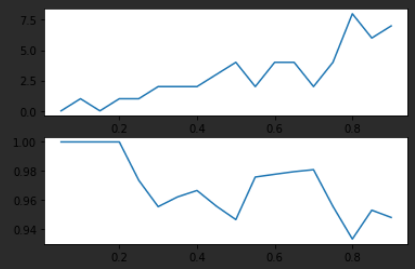
2) коэффициент Джини – показатель, определяющий чистоту узла

3) количество элементов в разделяемом множестве

4) список, показывающий, сколько образцов на данном узле попадают в каждую категорию

В листьях не содержится вопрос. В листьях содержатся экземпляры только одного класса.

1. Построен график зависимости неправильно классифицированных наблюдений и точности классификации от размера тестовой выборки. Размер тестовой выборки изменяется от 0.05 до 0.95 с шагом 0.05. Параметр random\_state выбран равным номеру зачетной книжки (830434).

Рисунок 11 – График зависимости точности классификации и неправильно классифицированных наблюдений от размера тестовой выборки

1. Исследованы параметры *DecisionTreeClassifier.*

|  |  |
| --- | --- |
| Параметр | Описание |
| criterion | Функция измерения качества разбиения.  Поддерживается индекс Джини и энтропия.  На исходных данных для обоих параметров алгоритм даёт похожие результаты. |
| splitter | Стратегия, используемая для выбора разбиения на каждом узле.  Поддерживается выбор наилучшего разбиения и случайный выбор.  На исходных данных случайная стратегия разделения дает больший результат ошибочных измерений. |
| 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.  При увеличении максимально возможной глубины число ошибочно классифицированных значений уменьшается. |
| min\_samples\_split | Минимальное количество выборок, необходимых для разделения внутреннего узла.  При увеличении минимального числа наблюдений для разбиения узла значительно возрастает количество неправильно классифицированных наблюдений. |
| 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.  При увеличении минимально возможного числа наблюдений в конечном узле увеличивается число неправильно классифицированных наблюдений. |

# Выводы

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