МИНОБРНАУКИ РОССИИ

Федеральное государственное автономное образовательное

учреждение высшего образования

«Южный федеральный университет»

Институт высоких технологий и пьезотехники

|  |  |
| --- | --- |
| Описание: https://pp.vk.me/c623122/v623122412/1acb7/yTJRs3eyFiE.jpg | Кафедра прикладной информатики и инноватики  Направление подготовки: 09.03.03 Прикладная информатика |

**Отчет на тему «Лабораторная работа №1–3»**

**По дисциплине «Основы машинного обучения»**

**Выполнил:**

студент 3 курса бакалавриата 6 группы

Кузин М.А.

**Проверил:**

доцент, к. ф-м. н.

Ефимов А.И.

Ростов-на-Дону

2022

**Цели и задачи работы**

Изучить принцип работы методов машинной инженерии Наивного Байеса, Деревьев решений для задач классификации. Реализовать на практике данные методы с помощью стандартных методов, библиотеки scikit-learn для Toy’s datasets классификации, создать с помощью pandas и numpy свою реализацию Наивного Байеса, Деревьев решений для заданных датасетов. Также для Toy’s datasets, написать программу, которая реализует ROC-кривую по данным обучения. Изучить принцип метода кластеризации К-средних. Реализовать данный метод как с помощью scikit-learn, так и с помощью numpy для Toy’s datasets.

**Использованные инструменты**

Python 3.9

Библиотеки: Sklearn(datasets, KNeighborsClassifier и т.д.), numpy, matplotlib(pyplot)

**Краткая теория**

**Наи́вный ба́йесовский классифика́тор** — простой вероятностный классификатор, основанный на применении теоремы Байеса со строгими (наивными) предположениями о независимости.

Все параметры модели могут быть аппроксимированы относительными частотами из набора данных обучения. Это оценки максимального правдоподобия вероятностей. Непрерывные свойства, как правило, оцениваются через нормальное распределение. В качестве математического ожидания и дисперсии вычисляются статистики — среднее арифметическое и среднеквадратическое отклонение соответственно.

Если данный класс и значение свойства никогда не встречаются вместе в наборе обучения, тогда оценка, основанная на вероятностях, будет равна нулю. Это проблема, так как при перемножении нулевая оценка приведет к потере информации о других вероятностях. Поэтому предпочтительно проводить небольшие поправки во все оценки вероятностей так, чтобы никакая вероятность не была строго равна нулю.

*ROC-кривая*

Кривые ROC обычно показывают истинную положительную скорость по оси Y и ложноположительную скорость по оси X. Это означает, что верхний левый угол графика является “идеальной” точкой - ложноположительная скорость равна нулю, а истинная положительная скорость равна единице. Это не очень реалистично, но это означает, что большая площадь под кривой (AUC) обычно лучше.

**Деревья решений (DT)** — это непараметрический контролируемый метод обучения, используемый для классификации и регрессии. Цель состоит в том, чтобы создать модель, которая предсказывает значение целевой переменной, изучая простые правила принятия решений, выведенные из характеристик данных. Дерево можно рассматривать как кусочно-постоянное приближение.

**Метод k-средних (k-Means Clustering)** – это очень известный и мощный алгоритм Обучения без учителя (Unsupervised Learning), который группирует похожие элементы в k кластеров. Он используется для решения многих сложных задач Машинного обучения (ML).

*Метод k-средних* пытается сгруппировать похожие элементы в три этапа:

1. Выберем значение k
2. Инициализируем центроиды (разделительные линии)
3. Выберем группу и найдем среднее значение расстояния между точками.

**Ход работы**

*Часть 1. Реализация метода Наивного Байеса*

Изображение выглядит как текст

Автоматически созданное описание

Рисунок - 1. Импорт нужных библиотек.

Для реализации данного метода классификации удобно использовать ООП. (Рис. 2)

Изображение выглядит как текст

Автоматически созданное описание

Рисунок - 2. Реализация метода Наивного Байеса (Метод обучения).

Изображение выглядит как текст

Автоматически созданное описание

Рисунок - 3. Реализация метода Наивного Байеса (Методы подсчета процентов правильности).

Изображение выглядит как текст

Автоматически созданное описание

Рисунок - 4. Реализация метода Наивного Байеса (Кроссвалидация и дальнейшее обучение).

Изображение выглядит как текст

Автоматически созданное описание

Рисунок - 5. Реализация метода Наивного Байеса с помощью библиотеки scikit-learn.

В библиотеке scikit-learn уже реализованы все необходимые методы.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок - 6. Импорт библиотек и реализация функции для вывода ROC-кривой.

Для вывода графика ROC-кривой используем библиотеку matplotlib(pyplot). (рис.5)

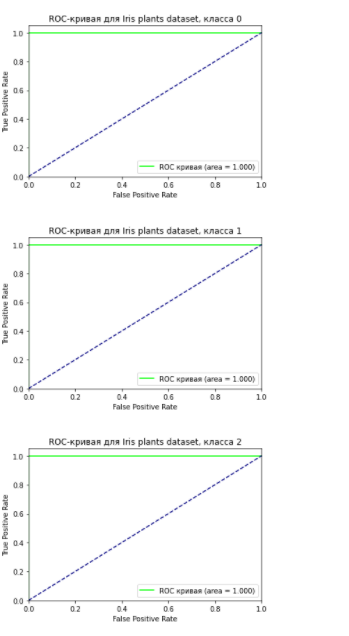


Рисунок - 7. Вывод ROC-кривых.

*Часть 2. Реализация деревьев решений*

Изображение выглядит как текст

Автоматически созданное описание

Рисунок - 8. Методы энтропии и gini.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок - 9. Методы построения дерева решений (с помощью рекурсии).

Изображение выглядит как текст

Автоматически созданное описание

Рисунок - 10. Методы предсказания.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок - 11. Построение графика.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок - 12. Применение кросс-валидации и вывод рез-тов.

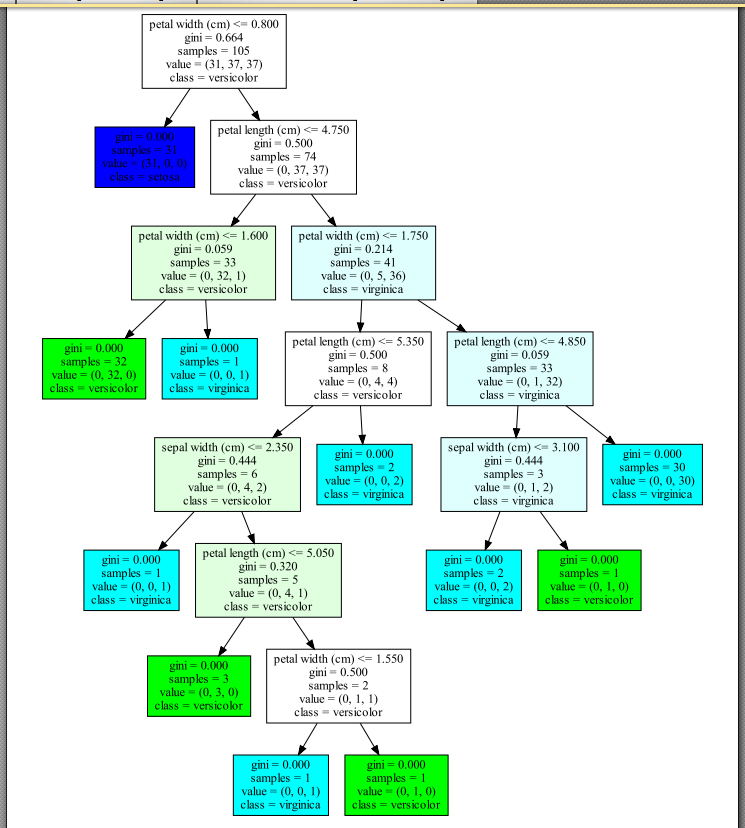


Рисунок - 13. Бинарный графа для Iris.

Изображение выглядит как текст, снимок экрана

Автоматически созданное описание

Рисунок - 14. Бинарный графа графа для Digits.

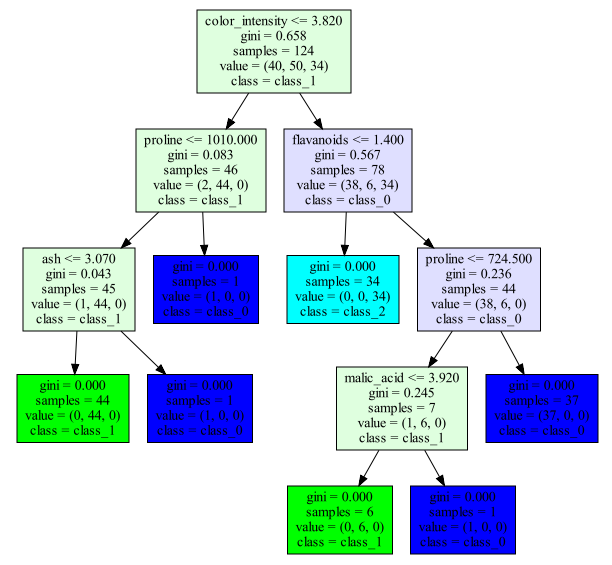


Рисунок - 15. Бинарный графа для Wine.

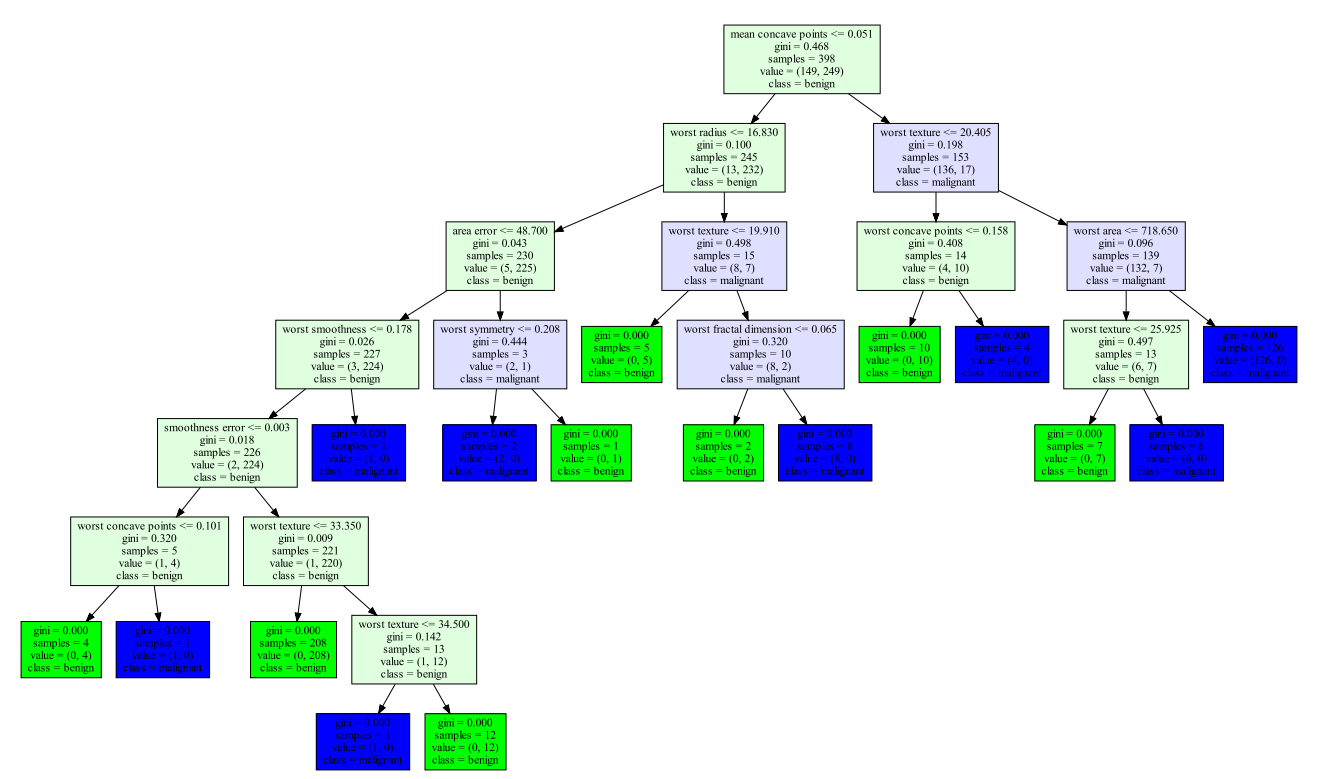


Рисунок - 16. Бинарный графа для Breast cancer wisconsin.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок - 17. Результаты обучения для Toy’s datasets.

*Часть 2. Реализация K-средних*

Изображение выглядит как текст

Автоматически созданное описание

Рисунок - 18. Определение кол-ва кластеров.

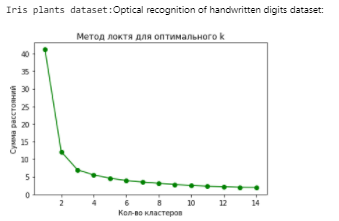


Рисунок – 19. Метод локтя.



Рисунок – 20. Реализация метода k-средних библиотеки scikit-learn.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок – 21. Результаты обучения.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок – 22. Реализация метода k-средних c помощью numpy.

Изображение выглядит как текст

Автоматически созданное описание

Рисунок – 23. Результаты обучения.

**Вывод**

Изучили принцип работы методов машинной инженерии Наивного Байеса, Деревьев решений для задач классификации. Реализовали на практике данные методы с помощью стандартных методов, библиотеки scikit-learn для Toy’s datasets классификации, создали с помощью pandas и numpy свою реализацию Наивного Байеса, Деревьев решений для заданных датасетов. Также для Toy’s datasets, написали программу, которая реализует ROC-кривую по данным обучения. Изучили принцип метода кластеризации К-средних. Реализовали данный метод как с помощью scikit-learn, так и с помощью numpy для Toy’s datasets.