

**Московский авиационный институт  
(Национальный исследовательский университет)**

Факультет: «Информационные технологии и прикладная математика»

Кафедра: 806 «Вычислительная математика и программирование»

Дисциплина: «Машинное обучение»

**Лабораторная работа № 1**

Студент: Паленов Павел

Группа: 80-301Б

Преподаватель: Ахмед Самир Халид

Дата:

Оценка:

Москва, 2021

### 1. Постановка задачи

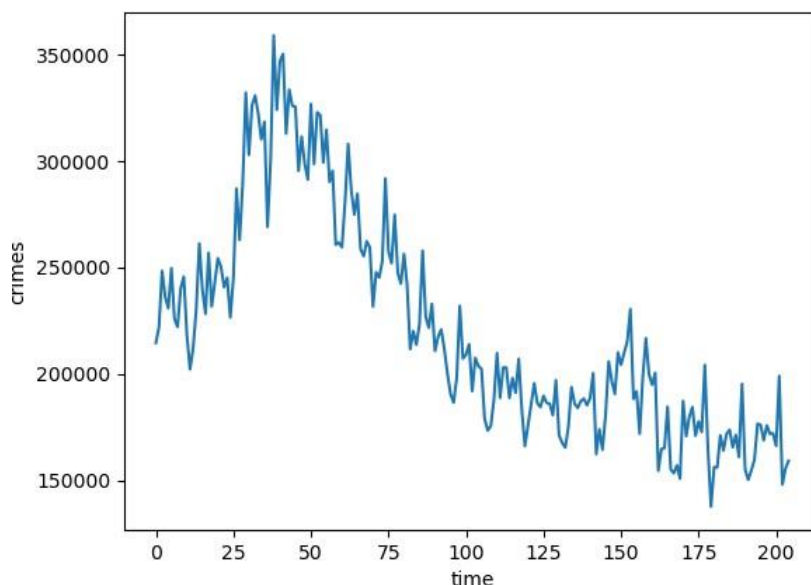
Найти себе набор данных (датасет) для следующей лабораторной работы и проанализировать его. Выявить проблемы набора данных, устранить их. Визуализировать зависимости, показать распределение некоторых признаков. Реализовать алгоритмы К ближайших соседа с использованием весов и Наивный Байесовский классификатор и сравнить с реализацией библиотеки `sklearn`.

### 2. Датасет

На Dataset Search я нашел интересный мне датасет, представленный исследовательской группой ВШЭ, в котором собрана статистика преступлений на территории РФ за последние годы.

### 3. Подготовка датасета

Датасет представляет количество совершенных преступлений по различным категориям: угон, мошенничество, с применением оружия и т.д. Ниже представлено как менялось общее число преступлений за последнее время.

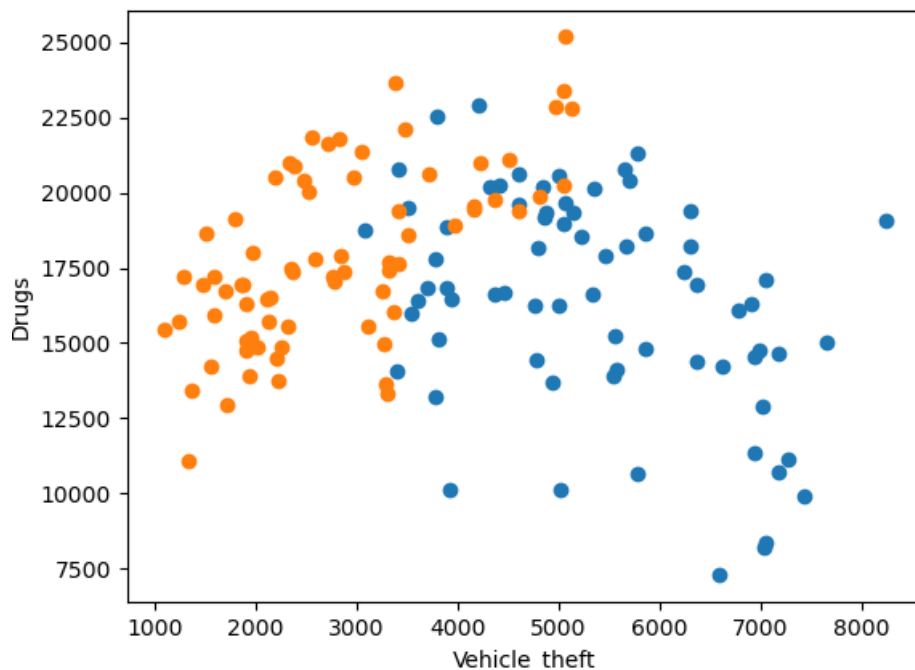


Для подготовки датасета необходимо обработать пропущенные значения и сделать парсинг.

В ходе проверки было выявлено, что пропущенных значений в датасете нет, что говорит о его качестве.

Мною было принято решение выделить два значимых параметра: преступления, связанные с угонами автомобилей и наркотиками. Данные можно разделить по временному параметру, преступления, совершенные до 2010 года (класс 0) и после 2012 года (класс 1).

Графически представляем эти данные.



Также разбиваем данные на тренировочные и тестовые случайным образом.

#### 4. KNN

Для классификации каждого из объектов тестовой выборки необходимо последовательно выполнить следующие операции:

- Вычислить расстояние до каждого из объектов обучающей выборки
- Отобрать  $k$  объектов обучающей выборки, расстояние до которых минимально
- Класс классифицируемого объекта — это класс, наиболее часто встречающийся среди  $k$  ближайших соседей

При взвешенном способе во внимание принимается не только количество попавших в область определённых классов, но и их удалённость от нового значения. Для каждого класса  $j$  определяется оценка близости:

$$Q_j = \sum_{i=1}^n \frac{1}{d(x, a_i)^2},$$

где  $d(x, a_i)$  — расстояние от нового значения  $x$  до объекта  $a_i$ .

У какого класса выше значение близости, тот класс и присваивается новому объекту.

#### 5. Наивный Байесовский классификатор

Наивный Байесовский классификатор основан на применении теоремы Байеса и предполагает независимость параметров.

Для оценки вероятности будем использовать гауссовскую функцию. Для этого нужно искать мат. ожидание и дисперсию для каждого атрибута.

$$P(x_i | y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

## 6. Результаты и сравнение с sklearn

```
(first) D:\MAI\ML\lab1>python lab1.py
my knn:      [0, 0, 0, 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
my bayes:    [0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1]
true val:    [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]
knn accuracy = 0.8260869565217391  bayes accuracy = 0.8260869565217391

sklearnKnn:  [0 0 0 0 1 0 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1]
accuracy = 0.8260869565217391

sklearnNB:   [0 0 0 0 0 0 1 0 1 1 0 1 0 1 1 1 1 1 1 1 1 1]
accuracy = 0.8260869565217391
```

Как видно из данного примера, точности реализованных алгоритмов и sklearn совпадают. При повторении генерации тестовых данных можно заметить, что алгоритм knn допускает на 1-2 ошибки больше, чем аналог sklearn. А алгоритмы NB работают с одинаковой точностью.

## СПИСОК ЛИТЕРАТУРЫ

### 1. KNN классификатор

<https://habr.com/ru/post/149693/>

### 2. Наивный байесовский классификатор в Python

<https://coderlessons.com/articles/bazy-dannykh-articles/naivnyi-baiesovskii-uchebnik-naivnyi-baiesovskii-klassifikator-v-python>