Алгоритъм за k-най-близки съседи в Python

В темата ще бъдат разгледани следните основни въпроси:

- Постановка на проблема
- Импортиране на множество от данни
- Статистика и корелация на множеството от данни

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Алгоритъм за k-най-близки съседи в Python

За краткост алгоритъмът за k-най-близки съседи ще бъде съкращаван с kNN (k-Nearest Neighbors). Същността и програмирането на kNN алгоритъма ще бъде разгледан с пример за предсказване възрастта на морските охлюви.

В темата се използва наборът от данни Abalone, който съдържа измервания на възрастта на голям брой охлюви. На фиг. 9.1 е показано как изглежда едно морско ухо.

Фиг. 9.1. Морско ухо

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Морските уши (абалон) са малки морски охлюви, които приличат малко на миди.

1. Постановка на проблема с охлювите

Възрастта на морското ухо може да се установи, като се разреже черупката му и се преброят пръстените на черупката. В набора от данни за охлювите може да се намерят измервания на възрастта на голям брой охлюви заедно с много други физически измервания.

Целта на проекта е да се разработи модел, който може да предскаже възрастта на охлюв въз основа единствено на други физически измервания. Това би позволило на изследователите да оценят възрастта на морското ухо, без да се налага да режат черупката му и да броят пръстените.

В проекта се прилага kNN, за да се намери възможно най-близкия прогнозен резултат.

2. Импортиране на множество от данни за Abalone

В тази тема ще се работи с набора от данни Abalone. Данните могат да се изтеглят и да се използва pandas, за да се импортират в Python. По-бързият вариант за реализация е pandas да импортира данните директно.

Импортирането на данните с помощта на pandas е показано на фиг. 9.2.

```
import numpy as np
import statsmodels.api as sm
```

Фиг. 8.1. Импортиране на пакети

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

```
import pandas as pd
url = ("http://archive.ics.uci.edu/ml/machine-learning-
databases/abalone/abalone.data")
# we use pandas to read the dataset and specify its feature
names
abalone = pd.read_csv(url, header=None)
abalone.columns = ['Sex', 'Length', 'Diameter', 'Height',
                   'Whole weight', 'Shucked weight',
                   'Viscera weight', 'Shell weight', 'Rings']
# this line displays the shape of the dataset
print(abalone.shape)
# this line displays the first five rows of the dataset
print(abalone.head())
```

а)

```
(4177, 9)
Sex Length Diameter Height Whole weight Shucked weight Viscera weight Shell weight Rings
0   M  0.455  0.365  0.095  0.5140       0.2245         0.1010         0.150      15
1   M  0.350  0.265  0.090  0.2255       0.0995         0.0485         0.070       7
2   F  0.530  0.420  0.135  0.6770       0.2565         0.1415         0.210       9
3   M  0.440  0.365  0.125  0.5160       0.2155         0.1140         0.155      10
4   I  0.330  0.255  0.080  0.2050       0.0895         0.0395         0.055       7
```

б)

Фиг. 9.2. Импортиране на данните с помощта на pandas

а. Код на Python **б.** Изход от програмата

В този код първо се импортира pandas, след което се използва, за да се прочетат данните, като се посочва пътя да бъде URL, така че файлът да бъде извлечен директно през Интернет. За да се провери дали, данните са импортирани правилно, може да се отпечата техния размер и първите 5 реда от тях. Наборът от данни се импортира в Python като pandas DataFrame. Имената на колоните могат да се намерят във файла abalone.names в

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

хранилището за машинно обучение на UCI. Те се добавят към DataFrame, както следва (фиг. 9.3).

```
abalone.columns = ['Sex', 'Length', 'Diameter', 'Height',
                   'Whole weight', 'Shucked weight',
                   'Viscera weight', 'Shell weight', 'Rings']
```

Фиг. 9.3. Добавяне на имената на колоните към DataFrame

Импортираните данни вече са по-разбираеми, трябва да се премахне колоната Sex. Целта на проекта е да се използват физически измервания, за да се предскаже възрастта на морското ухо. Тъй като полът не е чисто физическа мярка, трябва да се премахне от множеството от данни. Колоната Sex може да се изтрие с помощта на метода .drop () (фиг. 9.4).

```
abalone = abalone.drop("Sex", axis=1)
```

Фиг. 9.4. Изтриване на колоната Sex

С този код се изтрива колоната Sex, тъй като тя няма да има значение за модела.

3. Статистика от множеството от данни за морското ухо

Когато се работи с алгоритми за машинно обучение, е задължително са се познават данните, с които се работи. Това изследване е от голямо значение и е обект на една нова дисциплина „Наука за данните“ (Data

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Science). В темата, без да се навлиза в дълбочина на анализа на данните, може да се изчислят някои проучвателни статистики и графики.

Целевата променлива на този проект е пръстените Rings, така че можете да се започне с това. Хистограмата на разпределение на пръстените Rings може да даде бърз и полезен преглед на възрастовите диапазони, които могат да се очакват (фиг. 9.5).

```
import matplotlib.pyplot as plt
abalone["Rings"].hist(bins=15)
plt.show()
```

Фиг. 9.5. Изтриване на колоната Sex

Този код използва функцията .hist() на pandas, за да генерира хистограма с петнадесет контейнера. Частта matplotlib.pyplot от библиотеката matplotlib е съвкупност от функции, които визуализират matplotlib като функциите в MATLAB. За целта се използва метода .схов(). Преди да се импортира matplotlib.pyplot в сорс кода, трябва да се инсталира библиотеката matplotlib (фиг. 9.6).

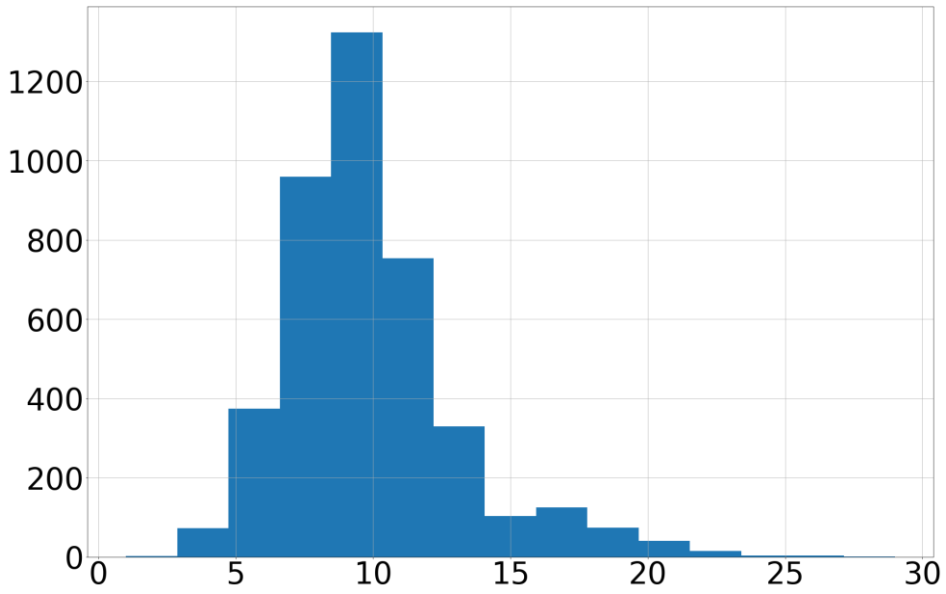
```
pip install matplotlib
```

Фиг. 9.5. Инсталиране на библиотеката matplotlib

Решението да се използват петнадесет контейнера при извеждане на хистограмата се основава на няколко опита. Когато се определя броя на

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

контейнерите, обикновено се спазва правилото да няма нито твърде много наблюдения на контейнер, нито твърде малко. Твърде малко контейнери могат да скрият определени модели, докато твърде много контейнери могат да накарат хистограмата да няма гладкост. Видът на хистограмата може да се видите на следната графика (фиг. 9.7).



Фиг. 9.7. Хистограма на разпределение на пръстените Rings

Хистограмата показва, че повечето охлюви в множеството от данни имат между пет и петнадесет пръстена, но е възможно да се получат до двадесет и пет пръстена. По-старите охлюви са недостатъчно представени в това множество от данни. Това изглежда интуитивно, тъй като възрастовите разпределения обикновено са изкривени по този начин поради естествени процеси.

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

Второ подходящо изследване е да се установи коя от другите променливи, ако има такива, има силна корелация с възрастта. Силна корелация между независима променлива и вашата целева променлива би била добър знак, тъй като това ще потвърди, че физическите измервания и възрастта са свързани.

Може да се изчисли пълната корелационна матрица в correlation_matrix. Най-важните корелации са тези с целевата променлива Rings (фиг. 9.8).

```
correlation_matrix = abalone.corr()
print(correlation_matrix["Rings"])
```

а)

```
Length            0.556720
Diameter          0.574660
Height            0.557467
Whole weight      0.540390
Shucked weight    0.420884
Viscera weight    0.503819
Shell weight      0.627574
Rings             1.000000
Name: Rings, dtype: float64
```

б)

Фиг. 9.2. Изчисляване на пълната корелационна матрица и извеждане на корелациите с целевата променлива Rings

а. Код на Python **б.** Изход от програмата

Разглеждайки коефициентите на корелация за пръстените с другите променливи, може да се заключи, че съществува някаква връзка между

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.

физическите измервания на охлювите и тяхната възраст, но тя също не е много висока. Колкото по-близо са коефициентите до 1, толкова по-голяма е корелацията.

Много високите корелации означават, че може да се очаква ясен процес на моделиране. В този случай ще трябва да се опита и да се види какви резултати могат да получат с помощта на алгоритъма kNN.

Съществуват много допълнителни възможности за изследване на данни с помощта на pandas. За да се научи повече за изследването на данни с pandas, може да се види официалната страница на pandas [https://pandas.pydata.org/].

Проект BG05M2OP001-2.016-0003„Модернизация на Национален военен университет "В. Левски" - гр. Велико Търново и Софийски университет "Св. Климент Охридски" - гр. София, в професионално направление 5.3 Компютърна и комуникационна техника“, финансиран от Оперативна програма „Наука и образование за интелигентен растеж“, съфинансирана от Европейския съюз чрез Европейските структурни и инвестиционни фондове.