Софийски университет „Св. Климент Охридски“
Факултет по математика и информатика

ДОМАШНА РАБОТА №2
по
Системи основани на знания
зимен семестър 2021/2022

Задача k-NN

Изготвил: Никола Петров Кирилов,
специалност ИС, ф.н 71986

Януари 2022
София

1.Описание на използвания метод за решаване на задачата

Задачата, която трябва да се реши е задача за класификация, посредством метод на машинно обучение. Нека имаме списък от пациенти, които страдат от едно и също заболяване и са лекувани с различни по вид медикаменти. Програмната система трябва да приеме списък с информация, в който се включват следните колони: **лекарството, с което се лекува даден пациент, години, пол, кръвно налягане, холестерол и съотношението натрий към калий в организма.**
Различните лекарства са: drugA, drugB, drugC, drugX и drugY. В задачата се изисква да се напише програма, предвиждаща кое е най-подходящото лекарство за нововъведен пациент, като използва знанието от вече направените опити с предишните пациенти. Първо ще трансформираме данните от таблицата и след това ще приложим **k-NN** алгоритъма.
За целта ще разбъркаме по индекс и ще разделим данните от файла(“**drug200.csv”**) на два вида. Данни за трениране “**Training”** и данни за тестване **“Testing”**, за да може, когато тестваме да не измисляме ние нови примерни данни, а да използваме готовите и така да разберем дали модела наистина работи. Идеята е данните от таблицата да ги представим като вектори, като направим т. нар “**preprocessing**” и данните ще ги представим като числа, като за целта използваме два вида **encoding** на данните, представящи категория**, Binary и One-hot**. Когато имаме само две категории, използваме **Binary**(съпоставяме едната категория да е **0**, а другата категория да е **1**), а когато имаме повече от 2, искаме да създадем **n-брой** нови колони и тези колони ще са **One-hot encoded**, и ще имат **1** само там където искаме да обозначим въпросната категория. Използваме к-NN като при появата на нов индивид се изчисляват разстоянията до всички останали класифицирани, чрез Евклидово разстояние, като се вземат най-близките K съседа и се прави проверка кое от тях се среща най-често. Нека зададем **k = 1**, за да можем да вземем най-близкият индивид и спрямо него ще се взима решение за новия. Също така ще използваме метриката **accuracy**, която варира в интервала **[0-1]** и ще показва колко добър е нашият модел, ако получим **0** означава много зле, а **1** означава много добре.

2. Описание на реализацията с псевдокод

За реализирането на решението съм използвал езика **Java(версия 15.0.2)**.
(с “//” и син цвят по псевдокода ще бележа коментари)

```
//Функция, която взема нашата матрица от вектори с данни прилага
//predictHelperFunction функцията по метода на функционалното програмиране
//и връща като резултат вектор от предложените лекарства
function  predict(List<List<Double>> X) {
        return map(predictHelperFunction);
```

```
    }


//Помощна функция, която прилага функцията за Евклидово разстояние, като
//се вземат най-близките K съседа и се прави проверка кое от тях се среща
//най-често
function predictHelper(list<double> x) {
    list<string> nearestN =
        .max(X, s ->euclideanDistance(s, x)
        .map(s -> y.get(X.indexOf(s)))

    list<string> distinct = nearestN
        .distinct()
        .sort(s -> frequency(nearestN, s)))
    return distinct[distinctSize - 1];
  }


//Функцията за намиране на Евклидово разстояние
function findEuclideanDistance(List<Double> s1, List<Double> s2) {
        int size = s1.size();

//За всеки елемент i докато стигнем размера събираме разликата на квадратите
        for (int i = 0; i < size; i++) {
            s += pow(s1[i] - s2[i]);
        }

//Връщаме корен квадратен от сбора
        return sqrt(s);
    }
```

3. Инструкции за компилиране на програмата

За да се компилира кодът е нужно потребителят да има инсталирана Java virtual machine версия 1.8, или по-висока. Въпросната програма може да бъде стартирана от команден прозорец:

1.Влезте в терминала през src папката на проекта
2.Въвеждате: javac Main.java KNNAlgorithm.java
3. .cd ../../out/production/Homework_KNN_2/KNN
4. java Main.class

Също така кодът може да се компилира на всяко IDE, което поддържа Java(Например IntelliJ, Visual Studio Code, Eclipse и тн.), като това е по-лесният и удобен начин. Влизате в програмата, която сте избрали и зареждате проекта, след което стартирате **“Main.java”**.

Програмата работи със **.csv** файл, като той трябва да е във формат като този даден в условието на задачата(**“drug200.csv”**).
Примерно съдържание на файла:

Age	Sex	BP	Cholesterol	Na_to_K	Drug	
23	F	HIGH	HIGH	25.355	drugY	
47	M	LOW	HIGH	13.093	drugC	
47	M	LOW	HIGH	10.114	drugC	
28	F	NORMAL	HIGH	7.798	drugX	
61	F	LOW	HIGH	18.043	drugY	
22	F	NORMAL	HIGH	8.607	drugX	
49	F	NORMAL	HIGH	16.275	drugY	
41	M	LOW	HIGH	11.037	drugC	
60	M	NORMAL	HIGH	15.171	drugY	
43	M	LOW	NORMAL	19.368	drugY	
47	F	LOW	HIGH	11.767	drugC	
34	F	HIGH	NORMAL	19.199	drugY	
43	M	LOW	HIGH	15.376	drugY	
74	F	LOW	HIGH	20.942	drugY	
50	F	NORMAL	HIGH	12.703	drugX	
16	F	HIGH	NORMAL	15.516	drugY	
69	M	LOW	NORMAL	11.455	drugX	

4. Примерни резултати

Пример 1:

За първия тест нека вкараме **200 записа**, които ще вкараме от файла **“drug200.csv”**

Вход:

	A	B	C	D	E	F
1	Age	Sex	BP	Cholesterol	Na_to_K	Drug
2	23	F	HIGH	HIGH	25.355	drugY
3	47	M	LOW	HIGH	13.093	drugC
4	47	M	LOW	HIGH	10.114	drugC
5	28	F	NORMAL	HIGH	7.798	drugX
6	61	F	LOW	HIGH	18.043	drugY
7	22	F	NORMAL	HIGH	8.607	drugX
8	49	F	NORMAL	HIGH	16.275	drugY
9	41	M	LOW	HIGH	11.037	drugC
10	60	M	NORMAL	HIGH	15.171	drugY
11	43	M	LOW	NORMAL	19.368	drugY
12	47	F	LOW	HIGH	11.767	drugC
13	34	F	HIGH	NORMAL	19.199	drugY
14	43	M	LOW	HIGH	15.376	drugY
15	74	F	LOW	HIGH	20.942	drugY
16	50	F	NORMAL	HIGH	12.703	drugX
17	16	F	HIGH	NORMAL	15.516	drugY
18	69	M	LOW	NORMAL	11.455	drugX
19	43	M	HIGH	HIGH	13.972	drugA
20	23	M	LOW	HIGH	7.298	drugC
21	32	F	HIGH	NORMAL	25.974	drugY
22	57	M	LOW	NORMAL	19.128	drugY

...

193	23	M	HIGH	HIGH	8.011	drugA
194	72	M	LOW	HIGH	16.31	drugY
195	72	M	LOW	HIGH	6.769	drugC
196	46	F	HIGH	HIGH	34.686	drugY
197	56	F	LOW	HIGH	11.567	drugC
198	16	M	LOW	HIGH	12.006	drugC
199	52	M	NORMAL	HIGH	9.894	drugX
200	23	M	NORMAL	NORMAL	14.02	drugX

Резултат на конзолата при K = 1:

K е: 1

Тествани екземпляри: [[0.0, 1.0, 0.0, 69.0, 1.0, 1.0, 10.065], [1.0, 0.0, 0.0, 58.0, 1.0, 1.0, 38.247], [0.0, 1.0, 0.0, 73.0, 1.0, 1.0, 19.221], [1.0, 0.0, 0.0, 65.0, 1.0, 0.0, 13.769], [0.0, 0.0, 1.0, 74.0, 0.0, 0.0, 15.436], [1.0, 0.0, 0.0, 47.0, 0.0, 1.0, 10.114], [0.0, 1.0, 0.0, 64.0, 0.0, 1.0, 7.761], [0.0, 0.0, 1.0, 70.0, 0.0, 1.0, 13.967], [0.0, 0.0, 1.0, 49.0, 0.0, 0.0, 8.7], [0.0, 0.0, 1.0, 53.0, 1.0, 0.0, 12.495], [0.0, 0.0, 1.0, 41.0, 0.0, 0.0, 15.156], [1.0, 0.0, 0.0, 74.0, 1.0, 1.0, 20.942], [0.0, 0.0, 1.0, 73.0, 1.0, 1.0, 18.348], [0.0, 0.0, 1.0, 19.0, 1.0, 0.0, 25.969], [0.0, 1.0, 0.0, 60.0, 0.0, 0.0, 10.091], [0.0, 1.0, 0.0, 39.0, 1.0, 0.0, 17.225], [0.0, 0.0, 1.0, 31.0, 0.0, 0.0, 17.069], [0.0, 0.0, 1.0, 50.0, 0.0, 1.0, 7.49], [1.0, 0.0, 0.0, 34.0, 1.0, 0.0, 12.923], [1.0, 0.0, 0.0, 38.0, 0.0, 1.0, 18.295], [0.0, 1.0, 0.0, 20.0, 1.0, 0.0, 9.281], [1.0, 0.0, 0.0, 58.0, 1.0, 1.0, 26.645], [0.0, 0.0, 1.0, 31.0, 0.0, 1.0, 30.366], [0.0, 0.0, 1.0, 51.0, 0.0, 1.0, 18.295], [1.0, 0.0, 0.0, 45.0, 0.0, 1.0, 17.951], [0.0, 0.0, 1.0, 18.0, 1.0, 1.0, 37.188], [0.0, 0.0, 1.0, 21.0, 1.0, 0.0, 28.632], [0.0, 1.0, 0.0, 68.0, 1.0, 0.0, 27.05], [1.0, 0.0, 0.0, 52.0, 0.0, 0.0, 32.922], [0.0, 0.0, 1.0, 68.0, 1.0, 0.0, 10.189], [0.0, 0.0, 1.0, 24.0, 0.0, 0.0, 9.475], [0.0, 1.0, 0.0, 18.0, 1.0, 0.0, 8.75], [0.0, 1.0, 0.0, 46.0, 0.0, 0.0, 7.285], [1.0, 0.0, 0.0, 61.0, 1.0, 1.0, 18.043], [0.0, 1.0, 0.0, 51.0, 1.0, 1.0, 13.597], [0.0, 0.0, 1.0,

58.0, 1.0, 0.0, 14.239], [1.0, 0.0, 0.0, 16.0, 0.0, 1.0, 12.006], [1.0, 0.0, 0.0, 42.0, 1.0, 0.0, 29.271], [0.0, 1.0, 0.0, 23.0, 0.0, 1.0, 16.85]]
Предсказани: [drugB, drugY, drugY, drugB, drugX, drugC, drugX, drugY, drugX, drugB, drugA, drugY, drugY, drugY, drugX, drugY, drugY, drugA, drugA, drugY, drugX, drugY, drugY, drugY, drugY, drugY, drugY, drugY, drugY, drugB, drugA, drugX, drugX, drugY, drugX, drugB, drugX, drugY, drugY]
Взети за тестване:[drugX, drugY, drugY, drugX, drugY, drugC, drugX, drugB, drugA, drugB, drugY, drugY, drugY, drugY, drugX, drugY, drugY, drugA, drugX, drugY, drugX, drugY, drugY, drugY, drugY, drugY, drugY, drugY, drugY, drugB, drugA, drugX, drugX, drugY, drugX, drugB, drugC, drugY, drugY]

Познат брой: 31

Тестван брой 39

Accuracy(Познат брой/Общ брой тествани): 0.794872

Получаваме, че нашата метрика е приблизително 79.5%, което е доста добре за модела

Резултат на конзолата при K = 3:

K е: 3

Тествани екземпляри: [[1.0, 0.0, 0.0, 34.0, 1.0, 0.0, 12.923], [0.0, 0.0, 1.0, 39.0, 0.0, 1.0, 9.664], [1.0, 0.0, 0.0, 28.0, 1.0, 1.0, 13.127], [0.0, 0.0, 1.0, 58.0, 1.0, 1.0, 19.416], [0.0, 0.0, 1.0, 15.0, 1.0, 0.0, 16.725], [0.0, 1.0, 0.0, 23.0, 0.0, 0.0, 14.02], [0.0, 0.0, 1.0, 19.0, 1.0, 0.0, 25.969], [0.0, 1.0, 0.0, 67.0, 0.0, 0.0, 9.514], [1.0, 0.0, 0.0, 49.0, 0.0, 0.0, 11.014], [1.0, 0.0, 0.0, 47.0, 0.0, 1.0, 13.093], [1.0, 0.0, 0.0, 33.0, 1.0, 1.0, 33.486], [0.0, 0.0, 1.0, 22.0, 0.0, 0.0, 28.294], [1.0, 0.0, 0.0, 26.0, 1.0, 1.0, 14.16], [0.0, 0.0, 1.0, 35.0, 1.0, 1.0, 12.894], [0.0, 1.0, 0.0, 18.0, 1.0, 0.0, 8.75], [0.0, 0.0, 1.0, 73.0, 1.0, 1.0, 18.348], [0.0, 0.0, 1.0, 32.0, 1.0, 0.0, 10.292], [0.0, 1.0, 0.0, 39.0, 0.0, 1.0, 15.969], [0.0, 0.0, 1.0, 65.0, 0.0, 0.0, 11.34], [0.0, 1.0, 0.0, 32.0, 1.0, 1.0, 7.477], [0.0, 1.0, 0.0, 50.0, 0.0, 0.0, 15.79], [0.0, 1.0, 0.0, 34.0, 0.0, 1.0, 22.456], [0.0, 0.0, 1.0, 15.0, 0.0, 0.0, 17.206], [0.0, 1.0, 0.0, 43.0, 0.0, 0.0, 12.859], [0.0, 1.0, 0.0, 46.0, 0.0, 0.0, 7.285], [1.0, 0.0, 0.0, 23.0, 0.0, 1.0, 7.298], [0.0, 0.0, 1.0, 16.0, 0.0, 0.0, 19.007], [0.0, 0.0, 1.0, 66.0, 0.0, 1.0, 16.347], [0.0, 0.0, 1.0, 59.0, 0.0, 1.0, 13.935], [0.0, 0.0, 1.0, 37.0, 1.0, 0.0, 23.091], [0.0, 1.0, 0.0, 15.0, 0.0, 1.0, 9.084], [0.0, 0.0, 1.0, 64.0, 0.0, 0.0, 20.932], [1.0, 0.0, 0.0, 57.0, 0.0, 0.0, 19.128], [0.0, 0.0, 1.0, 58.0, 1.0, 0.0, 14.239], [1.0, 0.0, 0.0, 58.0, 1.0, 1.0, 26.645], [1.0, 0.0, 0.0, 47.0, 1.0, 1.0, 10.067], [0.0, 0.0, 1.0, 31.0, 0.0, 0.0, 11.871], [0.0, 1.0, 0.0, 73.0, 1.0, 1.0, 19.221], [1.0, 0.0, 0.0, 32.0, 1.0, 1.0, 9.712]]
Предсказани: [drugX, drugX, drugA, drugY, drugA, drugX, drugY, drugX, drugX, drugY, drugY, drugY, drugA, drugA, drugX, drugY, drugA, drugY, drugX,

drugX, drugY, drugY, drugA, drugA, drugX, drugX, drugY, drugY, drugB, drugY, drugX, drugY, drugY, drugX, drugY, drugC, drugA, drugY, drugX]
Взети за тестване:[drugX, drugA, drugC, drugY, drugY, drugX, drugY, drugX, drugX, drugC, drugY, drugY, drugC, drugA, drugX, drugY, drugA, drugY, drugB, drugX, drugY, drugY, drugY, drugX, drugX, drugC, drugY, drugY, drugB, drugY, drugX, drugY, drugY, drugB, drugY, drugC, drugA, drugY, drugC]
Познат брой: 28
Тестван брой 39
Accuracy(Познат брой/Общ брой тествани): 0.717949

Получаваме точност, когато K = 3, от 71.8%, което е добре за модела.

Пример 2:

За първия тест нека вкараме **45 записа**, които ще вкараме с файла **“drug45.csv”**

Вход:

A	B	C	D	E	F	G	H
54	M	NORMAL	HIGH	24.658	drugY		
18	F	HIGH	NORMAL	24.276	drugY		
70	M	HIGH	HIGH	13.967	drugB		
28	F	NORMAL	HIGH	19.675	drugY		
24	F	NORMAL	HIGH	10.605	drugX		
41	F	NORMAL	NORMAL	22.905	drugY		
31	M	HIGH	NORMAL	17.069	drugY		
26	M	LOW	NORMAL	20.909	drugY		
36	F	HIGH	HIGH	11.198	drugA		
26	F	HIGH	NORMAL	19.161	drugY		
19	F	HIGH	HIGH	13.313	drugA		
32	F	LOW	NORMAL	10.84	drugX		
60	M	HIGH	HIGH	13.934	drugB		
64	M	NORMAL	HIGH	7.761	drugX		
32	F	LOW	HIGH	9.712	drugC		
38	F	HIGH	NORMAL	11.326	drugA		
47	F	LOW	HIGH	10.067	drugC		
59	M	HIGH	HIGH	13.935	drugB		
51	F	NORMAL	HIGH	13.597	drugX		
69	M	LOW	HIGH	15.478	drugY		
37	F	HIGH	NORMAL	23.091	drugY		
50	F	NORMAL	NORMAL	17.211	drugY		
62	M	NORMAL	HIGH	16.594	drugY		
41	M	HIGH	NORMAL	15.156	drugY		
29	F	HIGH	HIGH	29.45	drugY		
42	F	LOW	NORMAL	29.271	drugY		
56	M	LOW	HIGH	15.015	drugY		
36	M	LOW	NORMAL	11.424	drugX		
58	F	LOW	HIGH	38.247	drugY		
56	F	HIGH	HIGH	25.395	drugY		
20	M	HIGH	NORMAL	35.639	drugY		
15	F	HIGH	NORMAL	16.725	drugY		
31	M	HIGH	NORMAL	11.871	drugA		
45	F	HIGH	HIGH	12.854	drugA		
28	F	LOW	HIGH	13.127	drugC		
56	M	NORMAL	HIGH	8.966	drugX		
22	M	HIGH	NORMAL	28.294	drugY		
37	M	LOW	NORMAL	8.968	drugX		
22	M	NORMAL	HIGH	11.953	drugX		
42	M	LOW	HIGH	20.013	drugY		
72	M	HIGH	NORMAL	9.677	drugB		
23	M	NORMAL	HIGH	16.85	drugY		
50	M	HIGH	HIGH	7.49	drugA		
47	F	NORMAL	NORMAL	6.683	drugX		
35	M	LOW	NORMAL	9.17	drugX		
65	F	LOW	NORMAL	13.769	drugX		

Резултат на конзолата при K = 1:

***В този случай програмата е заделила за трениране 38 записа от файла, останалите 8 ползваме за тестване**

K е: 1

Тествани екземпляри(Показани са данни след трансормацията): [[1.0, 0.0, 0.0, 42.0, 1.0, 0.0, 29.271], [0.0, 1.0, 0.0, 64.0, 0.0, 1.0, 7.761], [0.0, 0.0, 1.0, 26.0, 1.0, 0.0, 19.161], [0.0, 1.0, 0.0, 47.0, 1.0, 0.0, 6.683], [0.0, 0.0, 1.0, 31.0, 0.0, 0.0, 17.069], [0.0, 0.0, 1.0, 38.0, 1.0, 0.0, 11.326], [1.0, 0.0, 0.0, 36.0, 0.0, 0.0, 11.424], [0.0, 0.0, 1.0, 59.0, 0.0, 1.0, 13.935]]

Предсказани :[drugY, drugX, drugY, drugA, drugY, drugA, drugA, drugB]

Взети за тестване:[drugY, drugX, drugY, drugX, drugY, drugA, drugX, drugB]

Познат брой: 6

Тестван брой: 8

Accuracy(Познат брой/Тестван брой): 0.750000

Получаваме точност в този тест от 75%, което е доста добре за нашия модел.

Резултат на конзолата при K = 3:

K е: 3

Тествани екземпляри: [[1.0, 0.0, 0.0, 58.0, 1.0, 1.0, 38.247], [0.0, 0.0, 1.0, 59.0, 0.0, 1.0, 13.935], [0.0, 0.0, 1.0, 20.0, 0.0, 0.0, 35.639], [0.0, 0.0, 1.0, 41.0, 0.0, 0.0, 15.156], [0.0, 0.0, 1.0, 22.0, 0.0, 0.0, 28.294], [1.0, 0.0, 0.0, 47.0, 1.0, 1.0, 10.067], [0.0, 0.0, 1.0, 56.0, 1.0, 1.0, 25.395], [0.0, 0.0, 1.0, 18.0, 1.0, 0.0, 24.276]]

Предсказани: [drugY, drugY, drugY, drugA, drugY, drugA, drugY, drugY]

Взети за тестване:[drugY, drugB, drugY, drugY, drugY, drugC, drugY, drugY]

Познат брой: 5

Тестван брой 8

Accuracy(Познат брой/Общ брой тествани): 0.625000

Получаваме точност в този тест от 62.5%.