Machine Learning Abgabe von

Prof. Dr. Dominik Heider Taha El Amine Kassabi (takas100)

Wintersemester 2023/24 Emre Gökcek (emgoe104)

Heinrich-Heine-Universität Düsseldorf Lutfi Orabi (luora100)

Ibrahim Shinahov (ibshi100)

# Übungsblatt 6 10. Dezember 2023

# Task 1. K-nearest neighbor (10 points)

**Note:** It is recommended to do this exercise by hand.

## a) Given the following data set: (7 P.)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | Temperature | Humidity | Wind | Play Tennis |
| D1 | Sunny | 26 | High |  | No |
| D2 | Sunny | 28 | High | Strong | No |
| D3 | Overcast | 29 | High | Weak | Yes |
| D4 | Rain | 23 | High | Weak | Yes |
| D5 | Rain |  | Normal | Weak | Yes |
| D6 | Rain | 12 | Normal | Strong | No |
| D7 | Overcast | 8 |  | Strong | Yes |
| D8 | Sunny | 25 | High | Weak | No |
| D9 | Sunny | 18 | Normal | Weak | Yes |
| D10 | Rain | 20 | Normal | Weak | Yes |
| D11 | Sunny | 20 | Normal | Strong |  |
| D12 | Overcast | 21 | High | Strong | Yes |
| D13 |  | 26 | Normal | Weak | Yes |
| D14 | Rain | 24 | High | Strong | No |
| D15 | Sunny | 23 | Normal | Weak | No |
| D16 | Sunny | 21 | Normal | Weak | Yes |

Use the Nearest Neighbors method to determine missing values. Choose k = 3. Normalize the attributes to [0, 1]. Use the Manhattan metric for distance or the 0/1 distance for nominal attributes.

I decided to go for 0/1 distance.

Normalized:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | Temperature | Humidity | Wind | Play Tennis |
| D1 | Sunny |  | 1 |  | 0 |
| D2 | Sunny |  | 1 | 1 | 0 |
| D3 | Overcast |  | 1 | 0 | 1 |
| D4 | Rain |  | 1 | 0 | 1 |
| D5 | Rain |  | 0 | 0 | 1 |
| D6 | Rain |  | 0 | 1 | 0 |
| D7 | Overcast |  |  | 1 | 1 |
| D8 | Sunny | 17/21 | 1 | 0 | 0 |
| D9 | Sunny | 10/21 | 0 | 0 | 1 |
| D10 | Rain | 12/21 | 0 | 0 | 1 |
| D11 | Sunny |  | 0 | 1 |  |
| D12 | Overcast |  | 1 | 1 | 1 |
| D13 |  |  | 0 | 0 | 1 |
| D14 | Rain |  | 1 | 1 | 0 |
| D15 | Sunny |  | 0 | 0 | 0 |
| D16 | Sunny |  | 0 | 0 | 1 |

### D1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| d | Outlook | Temperature | Humidity | Play Tennis | Distance |
| D2 | 0 |  | 0 | 0 |  |
| D3 | 1 |  | 0 | 1 |  |
| D4 | 1 |  | 0 | 1 |  |
| D6 | 1 |  | 1 | 0 |  |
| D8 | 0 |  | 0 | 0 |  |
| D9 | 0 |  | 1 | 1 |  |
| D10 | 1 |  | 1 | 1 |  |
| D12 | 1 |  | 0 | 1 |  |
| D14 | 1 |  | 0 | 0 |  |
| D15 | 0 |  | 1 | 0 |  |
| D16 | 0 |  | 1 | 1 |  |

Closest: D8, D2, D14

D1.Wind = 1 (Strong)

### D5

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| d | Outlook | Humidity | Wind | Play Tennis | Distance |
| D2 | 1 | 1 | 1 | 1 | 4 |
| D3 | 1 | 1 | 0 | 0 | 2 |
| D4 | 0 | 1 | 0 | 0 | 1 |
| D6 | 0 | 0 | 1 | 1 | 2 |
| D8 | 1 | 1 | 0 | 1 | 3 |
| D9 | 1 | 0 | 0 | 0 | 1 |
| D10 | 0 | 0 | 0 | 0 | 0 |
| D12 | 1 | 1 | 1 | 0 | 3 |
| D14 | 0 | 1 | 1 | 1 | 3 |
| D15 | 1 | 0 | 0 | 1 | 2 |
| D16 | 1 | 0 | 0 | 0 | 1 |

Closest: D10, D4, D9 (and D16)

D5.Temp =

### D7

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| d | Outlook | Temperature | Wind | Play Tennis | Distance |
| D2 | 1 |  | 0 | 1 |  |
| D3 | 0 |  | 1 | 0 |  |
| D4 | 1 |  | 1 | 0 |  |
| D6 | 1 |  | 0 | 1 |  |
| D8 | 1 |  | 1 | 1 |  |
| D9 | 1 |  | 1 | 0 |  |
| D10 | 1 |  | 1 | 0 |  |
| D12 | 0 |  | 0 | 0 |  |
| D14 | 1 |  | 0 | 1 |  |
| D15 | 1 |  | 1 | 1 |  |
| D16 | 1 |  | 1 | 0 |  |

Closest: D12, D3, D6

D7.Humidity = 1 (High)

### D11

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| d | Outlook | Temperature | Humidity | Wind | Distance |
| D2 | 0 |  | 1 | 0 |  |
| D3 | 0 |  | 1 | 1 |  |
| D4 | 1 |  | 1 | 1 |  |
| D6 | 1 |  | 0 | 0 |  |
| D8 | 0 |  | 1 | 1 |  |
| D9 | 0 |  | 0 | 1 |  |
| D10 | 1 |  | 0 | 1 |  |
| D12 | 1 |  | 1 | 0 |  |
| D14 | 1 |  | 1 | 0 |  |
| D15 | 0 |  | 0 | 1 |  |
| D16 | 0 |  | 0 | 1 |  |

Closest: D16, D9, D15

D11.PlayTennis = 1 (Yes)

### D13

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| d | Temperature | Humidity | Wind | Play Tennis | Distance |
| D2 |  | 1 | 1 | 1 |  |
| D3 |  | 1 | 0 | 0 |  |
| D4 |  | 1 | 0 | 0 |  |
| D6 |  | 0 | 1 | 1 |  |
| D8 |  | 1 | 0 | 1 |  |
| D9 |  | 0 | 0 | 0 |  |
| D10 |  | 0 | 0 | 0 |  |
| D12 |  | 1 | 1 | 0 |  |
| D14 |  | 1 | 1 | 1 |  |
| D15 |  | 0 | 0 | 1 |  |
| D16 |  | 0 | 0 | 0 |  |

Closest: D16, D9, D10

D13.Outlook = Sunny

Finally we have the following dataset

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | Temperature | Humidity | Wind | Play Tennis |
| D1 | Sunny | 26 | High | Yes | No |
| D2 | Sunny | 28 | High | Strong | No |
| D3 | Overcast | 29 | High | Weak | Yes |
| D4 | Rain | 23 | High | Weak | Yes |
| D5 | Rain | 20 1/3 | Normal | Weak | Yes |
| D6 | Rain | 12 | Normal | Strong | No |
| D7 | Overcast | 8 | High | Strong | Yes |
| D8 | Sunny | 25 | High | Weak | No |
| D9 | Sunny | 18 | Normal | Weak | Yes |
| D10 | Rain | 20 | Normal | Weak | Yes |
| D11 | Sunny | 20 | Normal | Strong | Yes |
| D12 | Overcast | 21 | High | Strong | Yes |
| D13 | Sunny | 26 | Normal | Weak | Yes |
| D14 | Rain | 24 | High | Strong | No |
| D15 | Sunny | 23 | Normal | Weak | No |
| D16 | Sunny | 21 | Normal | Weak | Yes |

## b) Do the classification labels (PlayTennis) have to be included? Why or why not? (1 P.)

I don’t think so. It depends on personal preferences and doesn’t necessarily have a high correlation with the data. I think we should separate them, estimate the weather data with KNN and finally try to predict PlayTennis.

An example of weird correlation is the coldest and the hottest days tennis is played. Both have string wind, which you usually avoid while playing tennis. Additionally they are both overcast, which is the only acceptable part.

There are also days where it rains or there are weak winds (which maybe would have been their weird preference), but they still play.

## c) Classify the new sample D17 for k=1. (1 P.)

D17: Outlook=Sunny, Temperature=23, Humidity=High, Wind=Strong

--> D17: Sunny, 15/21, 1, 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | Temperature | Humidity | Wind | Distance |
| D1 | 0 | 3/21 | 0 | 0 | 3/21 |
| D2 | 0 | 5/21 | 0 | 0 | 5/21 |
| D3 | 1 | 4/21 | 0 | 1 | 25/21 |
| D4 | 1 | 0 | 0 | 1 | 42/21 |
| D5 | 1 | (2 2/3)/21 | 1 | 1 | > 65/21 |
| D6 | 1 | 11/21 | 1 | 0 | 53/21 |
| D7 | 1 | 15/21 | 0 | 0 | 36/21 |
| D8 | 0 | 2/21 | 0 | 1 | 23/21 |
| D9 | 0 | 5/21 | 1 | 1 | 47/21 |
| D10 | 1 | 3/21 | 1 | 1 | 66/21 |
| D11 | 0 | 3/21 | 1 | 0 | 24/21 |
| D12 | 1 | 2/21 | 0 | 0 | 23/21 |
| D13 | 1 | 3/21 | 1 | 1 | 66/21 |
| D14 | 1 | 1/21 | 0 | 0 | 22/21 |
| D15 | 0 | 0/21 | 1 | 1 | 42/21 |
| D16 | 0 | 2/21 | 1 | 1 | 44/21 |

D1 is closest, therefore D17.PlayTennis = 0 (No)

## d) Test different values of k. At what value of k does the assignment change compared to k=1? (1 P.)

The rows will only include the value that is the k-closest (arbitrary for ties). When the PlayTennis = 1 count is more (or equal) than PlayTennis = 0 then it changes.

|  |  |  |
| --- | --- | --- |
|  | k-Closest | PlayTennis |
| 1 | D1 | 0 |
| 2 | D2 | 0 |
| 3 | D14 | 0 |
| 4 | D8 | 0 |
| 5 | D12 | 1 |
| 6 | D11 | 1 |
| 7 | D3 | 1 |
| 8 | D7 | 1 |

For we have , which rounded . Of course you could go all the way to .