

MDS – K-Nearest-Neighbors (KNN)

Aufgabe 1

Quelle: Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2021). *An Introduction to Statistical Learning: with Applications in R*. New York: Springer. 2nd Edition.

Wir wollen den *K-Nearest-Neighbors* Algorithmus für das Klassifikationsproblem anwenden. Wir haben folgenden kleinen Datensatz:

Beobachtung	x_{i1}	x_{i2}	x_{i3}	y_i
1	0	3	0	0
2	2	0	0	0
3	0	1	3	0
4	0	1	2	1
5	-1	0	1	1
6	1	1	1	0

y_i bezeichnet unsere Zielvariable und wir haben drei Prädiktoren bzw. Input-Variablen x_{i1} , x_{i2} , und x_{i3} . Im *K-Nearest-Neighbors* Algorithmus bestimmen wir die Vorhersagen für einen Datenpunkt basierend auf den Beobachtungen im Trainingsdatensatz, die am nächsten beim Datenpunkt liegen. Eine Möglichkeit die Nähe zwischen einem Datenpunkt und einer Beobachtung im Trainingsdatensatz zu berechnen ist die **Euklidische Distanz**. Um die Euklidische Distanz zwischen einem Punkt $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3})$ und einem Punkt $\mathbf{z}_i = (z_{i1}, z_{i2}, z_{i3})$ zu berechnen, verwenden wir folgende Formel:

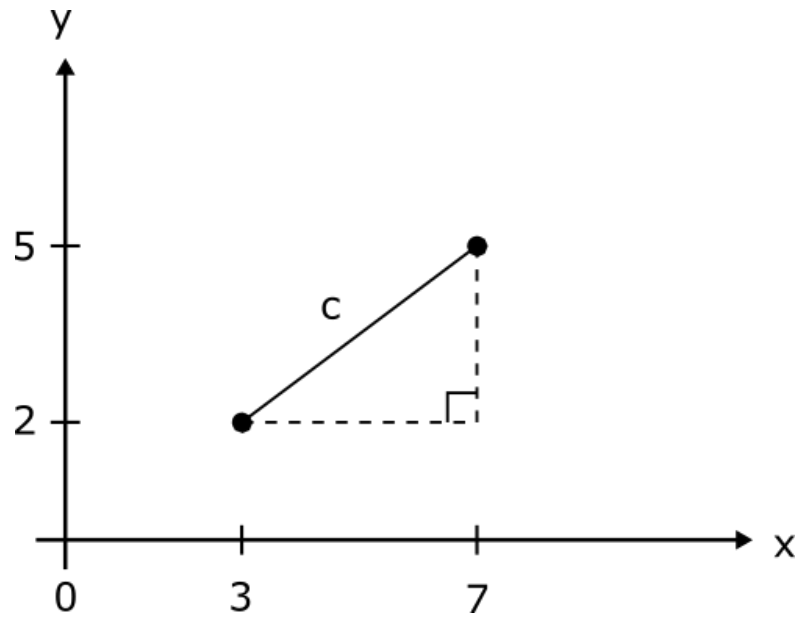
$$d(\mathbf{x}_i, \mathbf{z}_i) = \sqrt{(x_{i1} - z_{i1})^2 + (x_{i2} - z_{i2})^2 + (x_{i3} - z_{i3})^2}$$

Wir wollen nun eine Vorhersage für den Datenpunkt $\mathbf{x}^* = (0,0,0)$ machen, d.h. für einen Datenpunkt mit $x_1^* = x_2^* = x_3^* = 0$.

- Berechnen Sie die Euklidische Distanz zwischen jedem Datenpunkt im Trainingsdatensatz (die 6 Beobachtungen in der Tabelle) und der neuen Beobachtung $\mathbf{x}^* = (0,0,0)$.
- Was wäre unsere Vorhersage für \mathbf{x}^* , wenn $K = 1$? Warum?
- Was wäre unsere Vorhersage für \mathbf{x}^* , wenn $K = 3$? Warum?

Aufgabe 2

Verwenden Sie den Satz des Pythagoras, um die Länge der Seite c zu berechnen.



Was erkennen Sie, wenn Sie Ihre Lösung mit der Formel für die Euklidische Distanz in der ersten Aufgabe vergleichen?