Themen

k-nächste-Nachbarn-Algorithmus

Ein wichtiger Einsatzbereich des maschinellen Lernens ist die Klassifikation von Daten. Dabei lernt ein Algorithmus anhand von Trainingsdaten, neue Daten in Kategorien oder Klassen einzuordnen, beispielsweise beim Sortieren von E-Mails in Spam und Nicht-Spam.

Der k-nächste-Nachbarn-Algorithmus ist einer der gebräuchlichsten Algorithmen für die Klassifizierung von Daten und zugleich einer der einfachsten Algorithmen des überwachten maschinellen Lernens. Er geht davon aus, dass die Ähnlichkeit zweier Datenpunkte umso größer ist, je näher sie beieinander liegen. Das Lernen erfolgt, indem beispielhafte Datenpunkte gespeichert werden, die bereits den jeweiligen Klassen zugeordnet sind.

Man könnte den k-nächste-Nachbarn-Algorithmus beispielsweise anwenden, um für ein Heizgebläse vorherzusagen, ob es bei einer bestimmten Kombination von Heizleistung und Einschaltdauer ausfällt oder nicht. Als Ausgangsbasis dienen Daten von anderen Heizgebläsen, von denen bereits bekannt ist, ob sie ausgefallen sind ○ oder nicht ●.

Diese Trainingsdaten kann man zusammen mit den Betriebsdaten des neuen Heizgebläses ● (Testdaten) in ein Diagramm eintragen. Nun werden die den Testdaten ● am nächsten platzierten Datenpunkte betrachtet und gezählt, wie viele dieser nahen Datenpunkte zur Klasse „ausgefallen“ ○ gehören und wie viele zur Klasse „nicht ausgefallen“ ●. Die Klasse des neuen Datenpunkts wird von der Mehrheit der ihn umgebenden nächsten Nachbarn bestimmt.

Das k im Namen des Algorithmus bestimmt dabei die Anzahl der einbezogenen Nachbarpunkte.

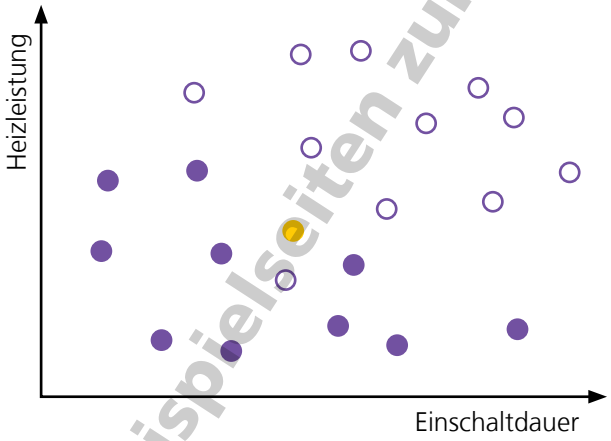
Für k = 1 betrachtet man nur den nächstgelegenen Datenpunkt. Da dieser zur Klasse „ausgefallen“ gehört, liegt die Vermutung nahe, dass auch das neue Heizgebläse aufgrund seiner Kombination aus Heizleistung und Betriebsdauer ausfallen wird.

Wählt man hingegen k = 3, werden zwei weitere Punkte in die Betrachtung einbezogen, die beide zur Klasse „nicht ausgefallen“ gehören. Demnach bestünde beim neuen Heizgebläse eine gute Chance, dass es nicht allzu bald ausfällt.

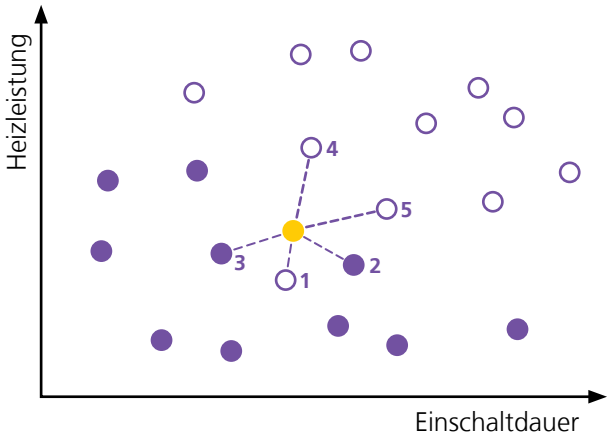
Bei k = 5 sind hingegen wieder die Datenpunkte der Klasse „ausgefallen“ in der Mehrzahl, das neue Heizgebläse müsste folglich in die Klasse „ausgefallen“ eingruppiert werden.

Die Wahl eines geeigneten Wertes für k ist folglich entscheidend für das Ergebnis, das der k-nächste-Nachbarn-Algorithmus liefert. Bei einem kleinen k können „Ausreißer“ in den Trainingsdaten die Klassenzuordnung verfälschen.

Wird k zu groß gewählt, besteht die Gefahr, Datenpunkte mit einem großen Abstand (und folglich geringer Ähnlichkeit) zum Testpunkt in die Klassifikationsentscheidung einzubeziehen. Diese Gefahr besteht insbesondere dann, wenn nur wenige Trainingsdaten verfügbar sind oder diese nicht gleichverteilt vorliegen.



Trainingsdaten:
○ Heizgebläse, die ausgefallen sind
● Heizgebläse, die nicht ausgefallen sind



Testdaten:
● neues Heizgebläse
Die fünf nächsten Nachbarn sind gekennzeichnet.

k-nächste-Nachbarn-Algorithmus

Aufgabe 1

Erläutere das Prinzip des k-nächste-Nachbarn-Algorithmus.

Welche Rolle spielt dabei der Parameter k?

Der k-nächste-Nachbarn-Algorithmus geht davon aus, dass die Ähnlichkeit zweier Datenpunkte umso größer ist, je näher sie beieinander liegen. Das Lernen erfolgt, indem beispielhafte Datenpunkte gespeichert werden, die bereits den jeweiligen Klassen zugeordnet sind. Die Klasse des neuen Datenpunkts wird von der Mehrheit der ihn umgebenden nächsten Nachbarn bestimmt.

Das k im Namen des Algorithmus bestimmt die Anzahl der einbezogenen Nachbarpunkte.

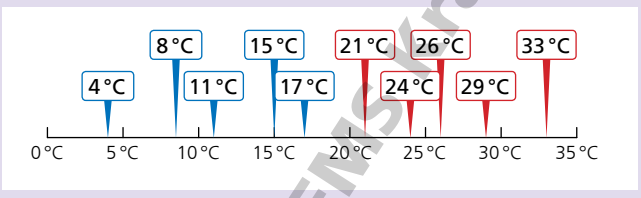
Aufgabe 2

Welche Auswirkungen kann es haben, wenn für k ein gerader Wert (2, 4, 6, ...) gewählt wird?

Wird für k ein gerader Wert gewählt, wird eine gerade Anzahl an Nachbarpunkten einbezogen. Dadurch kann es sein, dass sich die Nachbarpunkte zu gleichen Teilen auf zwei mögliche Klassen verteilen und keine Mehrheit vorliegt.

Aufgabe 3

In der Zeichnung sind als Trainingsdaten zehn Temperaturen eingezeichnet, die den beiden Klassen „kalt“ ▭ und „warm“ ▭ zugeordnet sind.



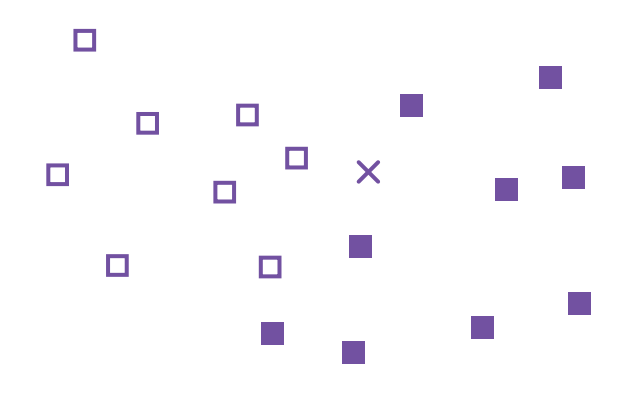
Ordne diese Testwerte mit Hilfe des k-nächste-Nachbarn-Algorithmus den Klassen „kalt“ ▭ und „warm“ ▭ zu.

	k	Testwert	Klasse
a)	3	19 °C	kalt
b)	5	19 °C	warm
c)	5	18 °C	kalt
d)	3	20 °C	warm
e)	7	20 °C	warm
f)	4	16 °C	kalt

k-nächste-Nachbarn-Algorithmus

Aufgabe 4
Markiere in der Grafik, welche Punkte rund um den Testwert ✕ bei einem k-Wert von 3 durch den k-nächste-Nachbarn-Algorithmus berücksichtigt werden.

Zu welcher Klasse gehört der Testwert ✕?

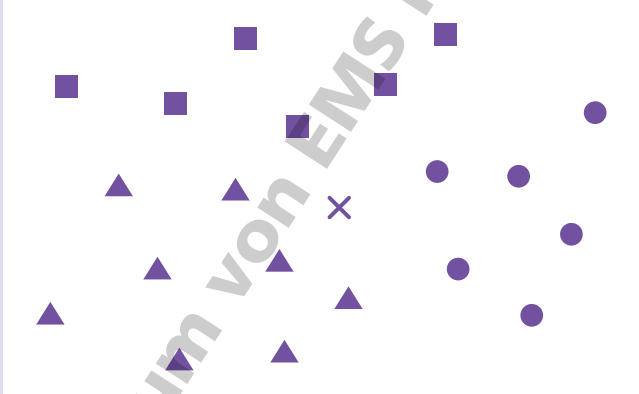


Beispiellösung

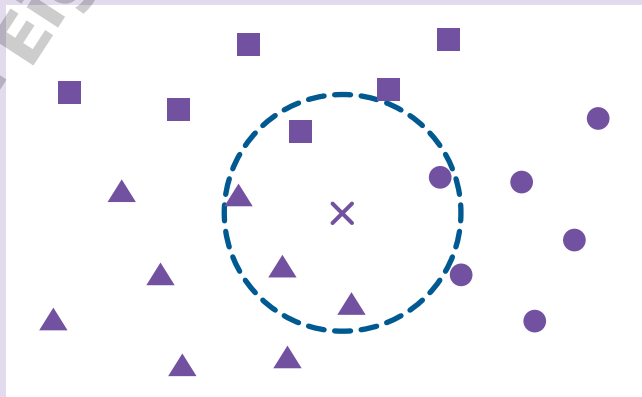
Der Testwert X gehört zur Klasse ■.

Aufgabe 5
Markiere in der Grafik, welche Punkte rund um den Testwert ✕ bei einem k-Wert von 5 durch den k-nächste-Nachbarn-Algorithmus berücksichtigt werden.

Zu welcher Klasse gehört der Testwert ✕?



Beispiellösung



Der Testwert X gehört zur Klasse ▲.