Woche 07 Theorie: k-nearest neighbours Verfahren

# Skript

Erarbeitet von
Katja Theune

Lernziele ..... 1
Inhalt ..... 2
Einstieg ..... 2
Was ist eine Klassifikation? ..... 2
Klassifikation – Beispiel ..... 2
K-nearest neighbours – Idee ..... 3
K-nearest neighbours – Finden der nächsten Nachbarn ..... 4
Abschluss ..... 6
Weiterführendes Material ..... 7
Disclaimer ..... 8

## Lernziele

- Definieren, was eine Klassifikation ist
- Erläutern der Idee und Vorgehensweise des k-nearest neighbours Verfahren
- Anwenden der Vorgehensweise des Verfahrens auf ein neues Beispiel
- Beispiele nennen, wozu man das k-nearest neighbours Verfahren verwendet



Projekt „KI für alle"
HeiCAD –Heine Center for Artificial Intelligence and Data Science

# Inhalt

## Einstieg

Ihr habt euch doch bestimmt auch schon mal mit anderen Personen verglichen und überlegt, wem ihr vielleicht vom Typ oder Verhalten her am ähnlichsten seid? Dieses Prinzip nutzt auch der k-nearest neighbours oder auf Deutsch der k-nächste Nachbarn Algorithmus. Dieses Verfahren gehört zum supervised learning und wird vorzugsweise für eine Klassifikation verwendet, die wir hier auch besprechen wollen. Wir können den k-nearest neighbours Algorithmus aber auch für ein Regressionsproblem verwenden. Die Idee dieses Verfahrens lässt sich ganz einfach übertragen.

## Was ist eine Klassifikation?

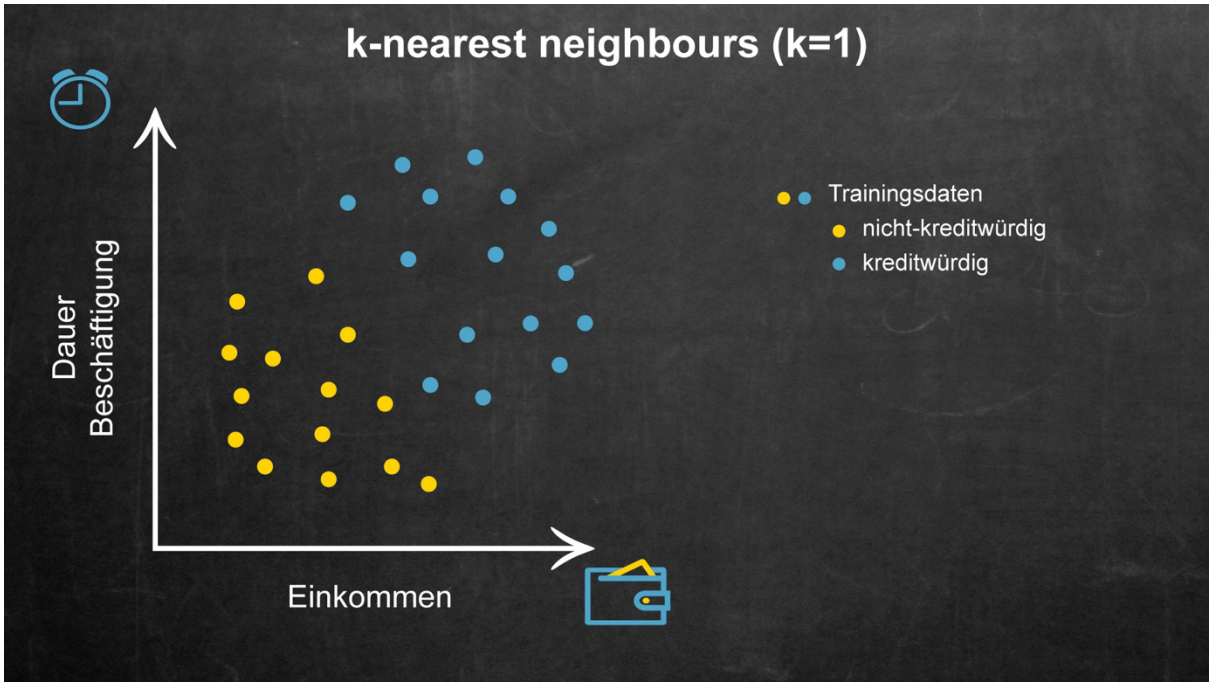
Aber was bedeutet denn eigentlich genau Klassifikation? Wenn wir uns nochmal an die lineare Regression erinnern, hilft sie uns, metrische Zielgrößen, z. B. das Einkommen, vorherzusagen. Im Gegensatz dazu hilft uns eine Klassifikation, eine Beobachtung anhand ihrer features einer bestimmten Kategorie, oder auch Klasse genannt, zuzuordnen. Wir haben hier also eine kategoriale Zielgröße. Aus unserem Alltag können wir uns z. B. die Einteilung von E-Mails in die beiden Klassen „Spam“ und „kein Spam“ vorstellen. In der Medizin wäre die Zuteilung von Patientinnen und Patienten zu verschiedenen Therapiegruppen ein Beispiel für eine Klassifikation. Je nach Untersuchungsproblem werden unterschiedlich viele Klassen berücksichtigt.

## Klassifikation – Beispiel

Um das noch besser zu verstehen, schauen wir uns mal ein typisches Beispiel für eine Klassifikation näher an. Stellen wir uns vor, ein Kreditinstitut möchte eine Entscheidung darüber treffen, ob ein aktueller Kreditantrag von einem Kunden oder einer Kundin genehmigt wird oder nicht. Unsere Beobachtungen sind jetzt also Antragstellende, die wir entweder der Klasse „kreditwürdig“ oder der Klasse „nicht kreditwürdig“ zuordnen wollen. Wir haben hier also ein Klassifikationsproblem mit zwei Klassen.

Unsere Trainingsdaten sind in unserem Beispiel Daten über bereits abgewickelte Kreditanträge bzw. ihre Antragstellenden. Darin sind zum einen die Klassen und zum anderen die features, die wir für die Klassifizierung nutzen wollen, enthalten. Hier nehmen wir jetzt als Beispiel das Einkommen der Antragstellenden und die Dauer des bestehenden Beschäftigungsverhältnisses. Zur besseren Veranschaulichung haben wir hier nur zwei features gewählt. In der Praxis haben wir natürlich meistens einige mehr.

Wir können jetzt zunächst unsere Trainingsdaten in einem Koordinatensystem als Datenpunkte eintragen. Da wir zwei features verwenden, haben wir hier ein zweidimensionales Koordinatensystem.




Wir tragen das Einkommen auf der horizontalen Achse und die Dauer des bestehenden Beschäftigungsverhältnisses auf der vertikalen Achse ab. Die gelben und blauen Datenpunkte repräsentieren unsere Beobachtungen, also die bisherigen Antragstellenden und ihre Eigenschaften. Die gelben Beobachtungen haben in unserem Beispiel die Klasse „nicht-kreditwürdig“ und die blauen Beobachtungen die Klasse „kreditwürdig“. Je weiter rechts die Beobachtungen liegen, desto höher ist ihr Einkommen. Je weiter oben die Beobachtungen liegen, desto länger ist die Dauer des bestehenden Beschäftigungsverhältnisses. Wir sehen hier, dass die blauen – also die „kreditwürdigen“ – Beobachtungen tendenziell bei beiden features höhere Werte haben als die gelben – also „nicht kreditwürdigen“ – Beobachtungen.

## K-nearest neighbours – Idee

Aber was genau ist denn jetzt eigentlich unser Ziel?
Im Prinzip wollen wir nun neue Beobachtungen – hier neue Antragstellende – einer dieser beiden Klassen zuordnen. Von diesen neuen Beobachtungen kennen wir nur ihre features, aber nicht ihre Klasse. Es gibt viele verschiedene Algorithmen zur Klassifikation. Ich möchte hier aber zunächst den k-nearest neighbours Algorithmus erklären, da er auch zu den einfachsten Klassifikationsverfahren im machine learning gehört.

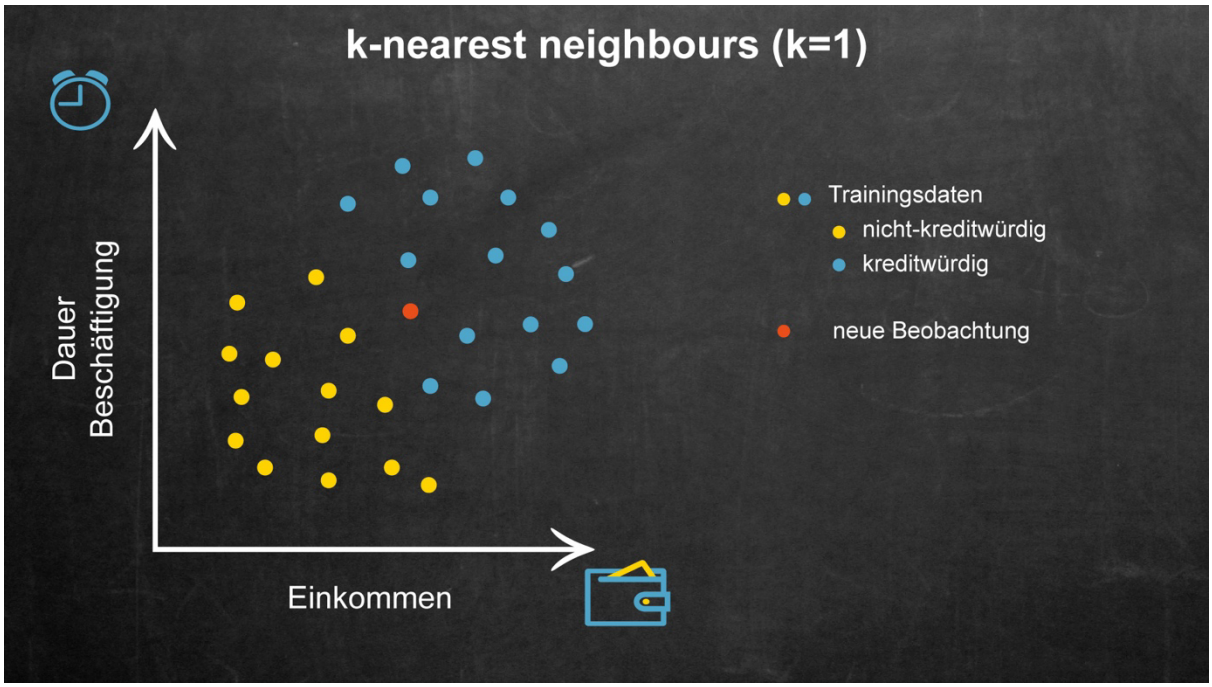
Der Name lässt es schon vermuten: das k-nearest neighbours Verfahren verwendet für die Klassifizierung das Konzept der Nähe zwischen Datenpunkten. Wir können diese Nähe z. B. auch mit Ähnlichkeit zwischen Beobachtungen beschreiben. Hier liegt die Annahme zugrunde, dass ähnliche Dinge oder Personen eben auch ähnliche Eigenschaften besitzen. Eine neue Beobachtung klassifizieren wir also dadurch, dass wir nach ihren k nächsten Nachbarn suchen. Das sind diejenigen Beobachtungen, die ihr am ähnlichsten bzw. am nächsten sind. D. h. sie haben ähnliche Eigenschaften, also hier ein ähnliches Einkommen

und eine ähnliche Dauer des Beschäftigungsverhältnisses. k steht für eine von uns zu wählende Anzahl an Nachbarn, die wir in die Klassifikation mit einbeziehen möchten. Der neuen Beobachtung ordnen wir dann diejenige Klasse zu, die am häufigsten bei ihren k nächsten Nachbarn vorkommt.

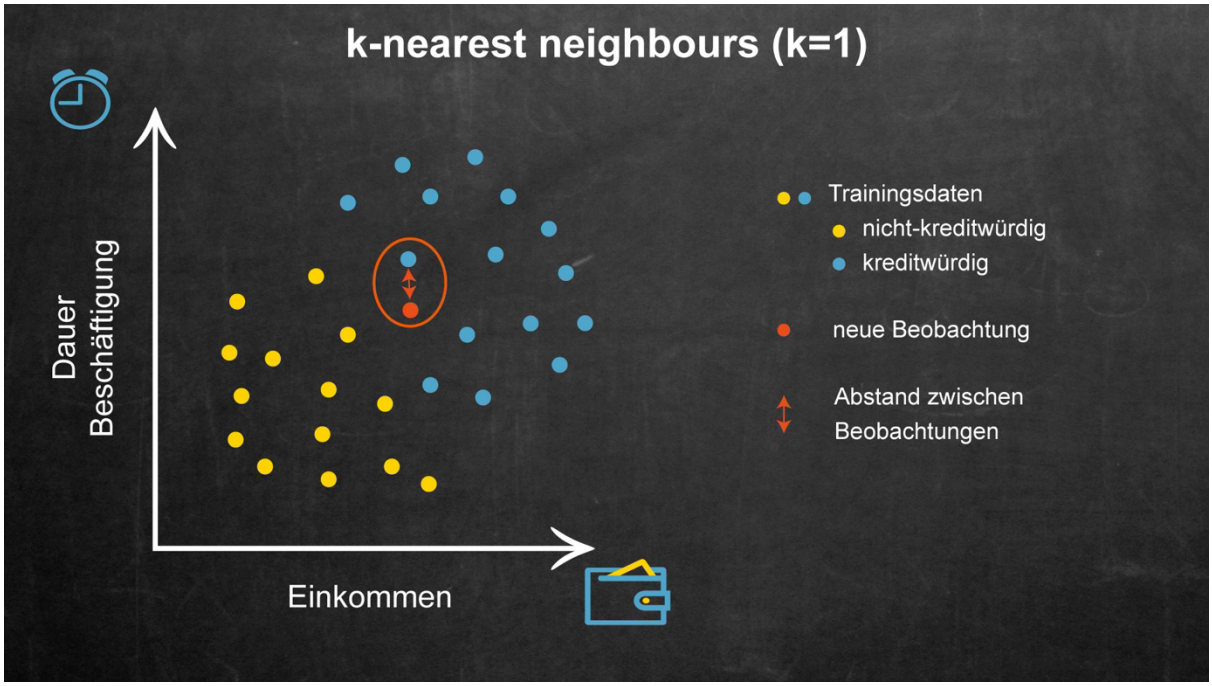
Das k-nearest neighbours Verfahren wird auch als lazy oder träges Lernverfahren bezeichnet, da es im Prinzip nur die vorliegenden Trainingsdaten speichert und neue Beobachtungen mit ihnen vergleicht. Es durchläuft also keine Trainingsphase und lernt nicht im eigentlichen Sinne.

## K-nearest neighbours – Finden der nächsten Nachbarn

Lasst uns jetzt noch mal unsere Grafik anschauen, um die Funktionsweise des Verfahrens und die Bestimmung der nächsten Nachbarn besser zu verstehen. Wir legen zunächst fest, dass k=1 sein soll. Wir suchen jetzt also nur den einen nächsten Nachbarn unserer neuen Beobachtung. In der Grafik ist beispielhaft eine neue Beobachtung durch den orangenen Punkt gekennzeichnet.

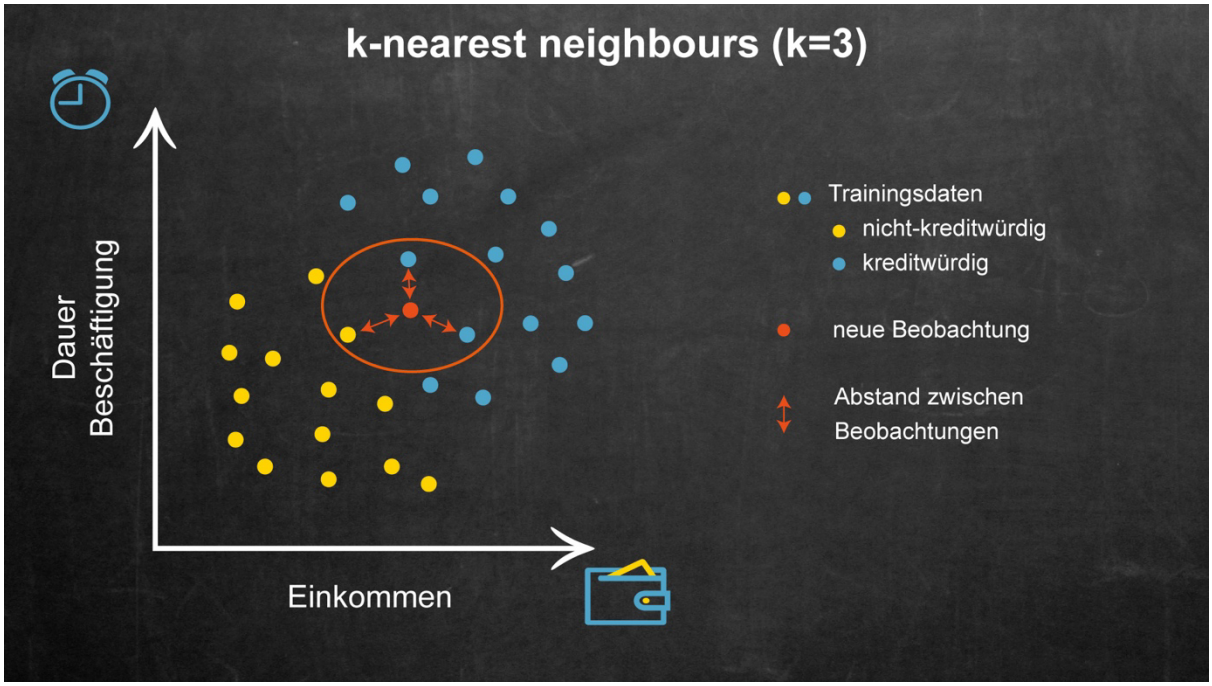



Jetzt ist es noch wichtig zu überlegen, was eigentlich genau hier mit Nähe gemeint ist und wie wir diese messen können. Wir meinen hier den Abstand zwischen zwei Beobachtungen bzw. Datenpunkten. Es gibt viele verschiedene Methoden, wie wir diesen Abstand messen könnten. Häufig wird die Euklidische Distanz verwendet. Sie wird auch als Fluglinie bezeichnet und ist hier einfach die Länge der Geraden, die wir zwischen den beiden Beobachtungen ziehen und theoretisch mit einem Lineal abmessen können.




Diese Distanz ist in unserem Beispiel durch den orangenen Pfeil gekennzeichnet. Unsere neue Beobachtung und ihr, dieser Distanz nach, nächster Nachbar sind orange umkreist. Jetzt haben wir also mit unserem gewählten Abstandsmaß den nächsten Nachbarn gefunden und ordnen unserer neuen Beobachtung die Klasse dieses Nachbarn zu. Wir sehen hier, dass unser nächster Nachbar die Klasse „kreditwürdig“ hat. Diese Klasse prognostizieren wir jetzt auch für unsere neue Beobachtung.

Neben dem Abstandsmaß müssen wir noch unser k wählen, also die Anzahl an Nachbarn, die wir zur Klassifizierung verwenden wollen. In unserem Beispiel von vorhin hatten wir z. B. k=1 gewählt.



Wenn wir jetzt stattdessen z. B. k=3 wählen, suchen wir diesmal die drei nächsten Nachbarn und weisen der neuen Beobachtung die häufigste Klasse zu, die unter den drei Nachbarn vorkommt. Daher ist es bei zwei Klassen auch nützlich, für k eine ungerade Anzahl zu wählen. Hier ist dann immer eine Klasse in der Mehrheit.

In unserem Beispiel haben wir unsere neue Beobachtung und ihre drei nächsten Nachbarn wieder orange umkreist. Wir sehen, dass unter den drei Nachbarn die blaue Klasse zweimal und die gelbe Klasse nur einmal vorkommt. Daher ordnen wir unserer neuen Beobachtung die blaue Klasse, also die Klasse „kreditwürdig“ zu.

Ein Vorteil des k-nearest neighbours Verfahrens ist seine Einfachheit und intuitive Herangehensweise. Wir müssen auch keine Annahmen über die genauen Zusammenhänge zwischen Zielgröße und features treffen, die im Falle von falschen Annahmen zu erheblichen Biases in den Ergebnissen führen können. Gerade wenn wir komplexe Zusammenhänge vermuten, ist das von Vorteil und steht im Gegensatz zur Regression. Ein Nachteil ist, dass das Verfahren bei vielen Beobachtungen und features durch die notwendige Speicherung der Daten sehr langsam wird und einen großen Arbeitsspeicher benötigt.

## Abschluss

Wir kennen jetzt das sehr intuitive k-nearest neighbours Verfahren zur Klassifikation. Es vergleicht Beobachtungen und sucht Ähnlichkeiten zwischen ihnen, so wie wir es im Prinzip alle aus unserem Alltag kennen. Anwendung findet das k-nearest neighbours Verfahren neben der bereits besprochenen Prognose der Kreditwürdigkeit z. B. auch häufig in der automatisierten Textanalyse.

| Einblendung kleine Grafik k-nearest neighbours mit k=3 |
|---|

# Weiterführendes Material

## Fachbücher:

Guter Einstieg ins Thema, anschaulich erläutert, keine Formeln oder tiefergehende methodische Erläuterungen:

Aberham, J., & Kossen, J. (2019). Klassifikation - Schubladendenken. In K. Kersting, C. Lampert, & C. Rothkopf (Hrsg.), *Wie Maschinen lernen* (S. 45-52). Springer, Wiesbaden.

Neumann, M. (2019). k-Nächste-Nachbarn – Nachbarschaftshilfe mal anders. In K. Kersting, C. Lampert, & C. Rothkopf (Hrsg.), *Wie Maschinen lernen* (S. 73-79). Springer, Wiesbaden.

Auch wenn dieses Buch mit R anstatt Python arbeitet, anschauliche Erklärung der Methoden, tiefergehende methodische Erläuterungen:

Lantz, B. (2015). *Machine learning with R* (2. Auflage). Packt Publishing Ltd, Birmingham.
- Chap. 3: Lazy Learning – Classification using Nearest Neighbors

Klassisches Werk für Statistisches/Maschinelles Lernen, tiefergehende methodische Erläuterungen:

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2. Auflage). Springer.
- Chap. 4: Classification

## Videos/Kurse:

Kurzer Einstieg ins Thema, anschaulich erläutert:

So lernen Maschinen: #4 Überwachtes Lernen - Klassifikation.
https://ki-campus.org/videos/solernenmaschinen

Etwas weitergehender Einstieg ins Thema, anschaulich erläutert:

AMALEA - Angewandte Machine Learning Algorithmen, Woche 3, Kap. 3: Schöne Nachbarschaft.
https://learn.ki-campus.org/courses/amalea-kit2021/items/5sG8I5AKowvfscFtvckpYw

Elements of AI, Chap. 4: Machine learning, II. The nearest neighbor classifier.
https://course.elementsofai.com/4/2

## Disclaimer

Transkript zu dem Video „Woche 07 Theorie: k-nearest neighbours Verfahren“, Katja Theune.