Nutzen und Funktionsweise maschineller Lernverfahren am Beispiel des Clustering

Dr. Ralf Höchenberger

DHBW Heidenheim
08. September 2021

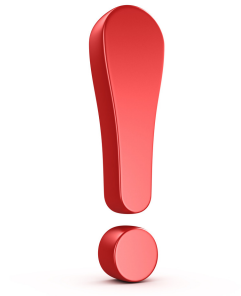
Was ist Machine Learning?

Definition (Fraunhofer-Gesellschaft 2018):

- *Lernalgorithmen* ***entwickeln aus Beispielen ein komplexes Modell***
- *Modell kann anschließend* ***auf neue, potenziell unbekannte Daten derselben Art angewendet*** *werden*
- ***Prozess****: Datenerhebung, Datenaufbereitung, Modellbildung, Modellevaluation*

Was nützt Machine Learning?

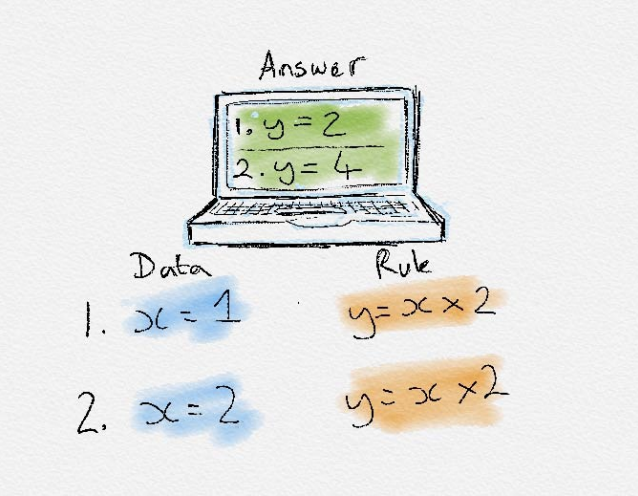
Steigerung des Unternehmenserfolgs!

... 

Voraussetzung ist aber, die Erkenntnisse in der Unternehmensstrategie zu verankern und operativ umzusetzen!

Wie funktioniert Machine Learning?

Überwachtes Lernen: Mensch gibt Daten und Regel vor, Maschine liefert Antwort.

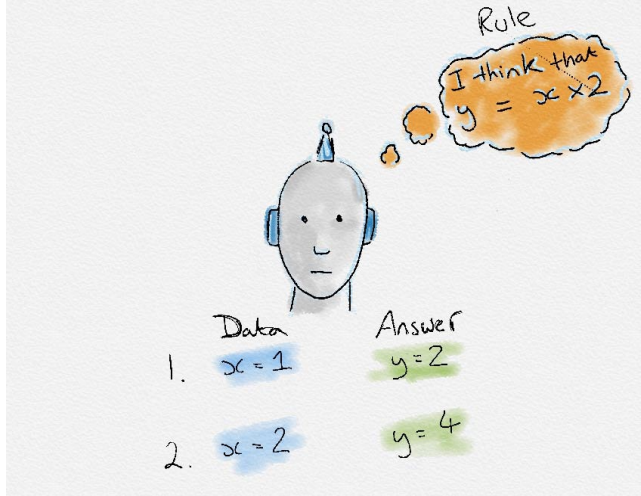


Kunde kauft eine Packung Weißwürste.
Wer Weißwürste kauft, kauft auch eher süßen Senf.

Kunde kauft ebenfalls zwei kleine Packungen süßen Senf.

→ Marketingstrategie: Weißwürste und süßen Senf in einer gemeinsamen Packung verkaufen.

Unüberwachtes Lernen: Mensch gibt Daten und Antwort vor, Maschine liefert Regel.



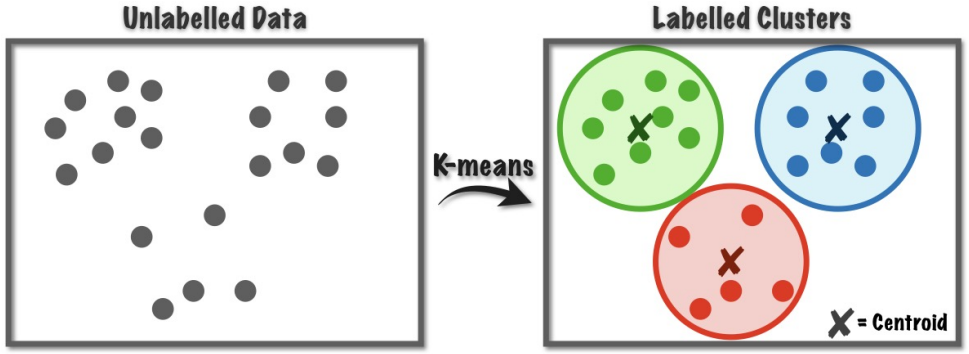
Männlicher Kunde kauft eine Packung Windeln.
Derselbe Kunde kauft auch einen Sixpack Bier.

Männliche Kunden, die Windeln kaufen, kaufen auch eher Bier!

→ Marketingstrategie für gestresste Väter von Kleinkindern: **Bier und Windeln nahe beieinander im Laden platzieren!**

Was ist k-Means-Clustering?

- **Identifizierung von Ähnlichkeitsstrukturen**, d.h. ähnliche Datenobjekte sollen zu Clustern zusammengenfasst werden
  - Daten **innerhalb eines Clusters** sollen möglichst **ähnlich** sein
  - Daten **zwischen Clustern** sollen möglichst **unähnlich** sein
- **Methode des unüberwachten Lernens**
- Datenobjekte sind zunächst ohne Cluster (bzw. Label) und werden in eine **zuvor definierte Anzahl von Clustern ($= k$) zugeordnet**, mit einem **Clustermittelpunkt (Centroid)** als Schwerpunkt
- **Ähnlichkeit** zweier Datenobjekte wird mit ihrem **Abstand zueinander** gemessen (mit dem Euklidischen Abstand)



Was ist maschinelles Lernen?

Was nützt maschinelles Lernen?

Wie funktioniert maschinelles Lernen?

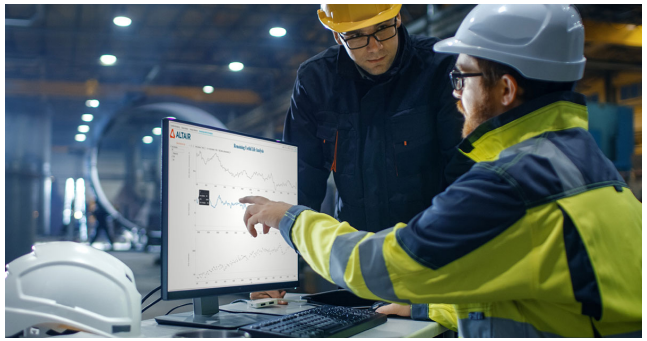
Was ist Clustering?

Was nützt Clustering?

Wie funktioniert Clustering?

Clustering: Use Cases aus der Praxis

Vorhersage von Maschinenausfällen („Remaining Useful Life“)



Erkennung von Kreditkartenbetrug („Fraud Detection“)

Umsetzung von Marketingstrategien („Customer Segmentation“)

Früherkennung von Krankheiten („Patient Diagnosis“)

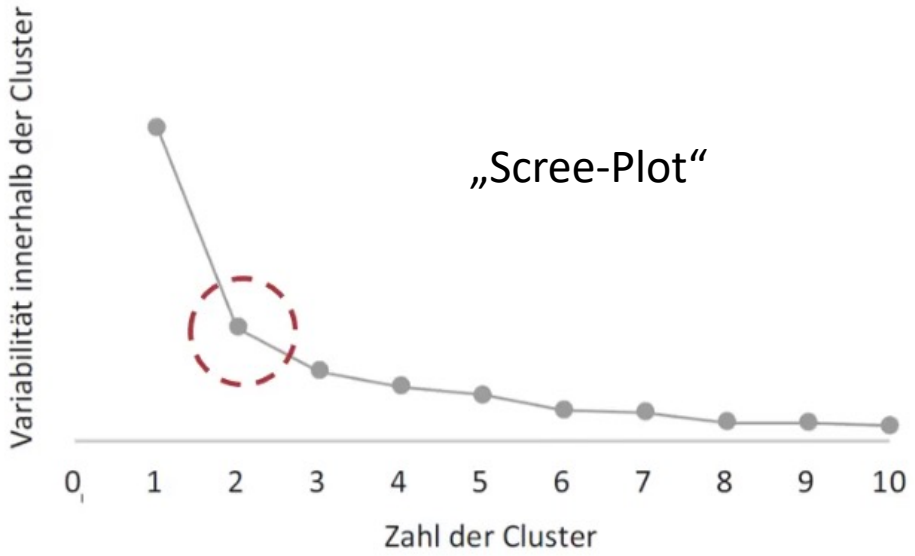
- Was ist maschinelles Lernen?
- Was nützt maschinelles Lernen?
- Wie funktioniert maschinelles Lernen?
- Was ist Clustering?
- **Was nützt Clustering?**
- Wie funktioniert Clustering?

Wie wird die Anzahl Cluster $(= k)$ bestimmt?

- **mehr Cluster**:
  - Mitglieder eines Clusters werden sich **zunehmend ähnlicher**, aber
  - benachbarte Cluster sind **immer schlechter voneinander unterscheidbar**
- **weniger Cluster:**
  - benachbarte Cluster sind **immer besser voneinander unterscheidbar,** aber
  - Mitglieder eines Clusters werden sich **zunehmend unähnlicher**

Es muss also ein **Mittelweg** gefunden werden, bei dem die Zahl der Cluster

- **groß genug** ist, um **aussagekräftige Muster abzuleiten**, und gleichzeitig
- so **klein bleibt**, dass sie sich noch **deutlich voneinander unterscheiden**



„Scree-Plot“

Was ist maschinelles Lernen?

Was nützt maschinelles Lernen?

Wie funktioniert maschinelles Lernen?

Was ist Clustering?

Was nützt Clustering?

Wie funktioniert Clustering?

k-Means-Clustering: Lloyd-Algorithmus

Gegeben: Anzahl gewünschter Cluster $k < n$ und maximale Laufzeit t_{max}

1. **Initialisierung**:
   Setze $t = 0$ und wähle k zufällige Clusterzentren aus den n Datenpunkten aus
2. **Zuordnung**:
   Ordne jeden Datenpunkt jeweils demjenigen Cluster mit dem nächstgelegenen Zentrum (= Zentrum, zu dem minimaler Abstand besteht, an Hand des Euklidischen Abstandes) zu
3. **Aktualisieren**:
   Berechne die Zentren der resultierenden Cluster neu
4. Wiederhole Schritte 2 und 3, bis sich die Zuordnungen nicht mehr ändern, oder $t \geq t_{max}$ gilt

Was ist maschinelles Lernen?

Was nützt maschinelles Lernen?

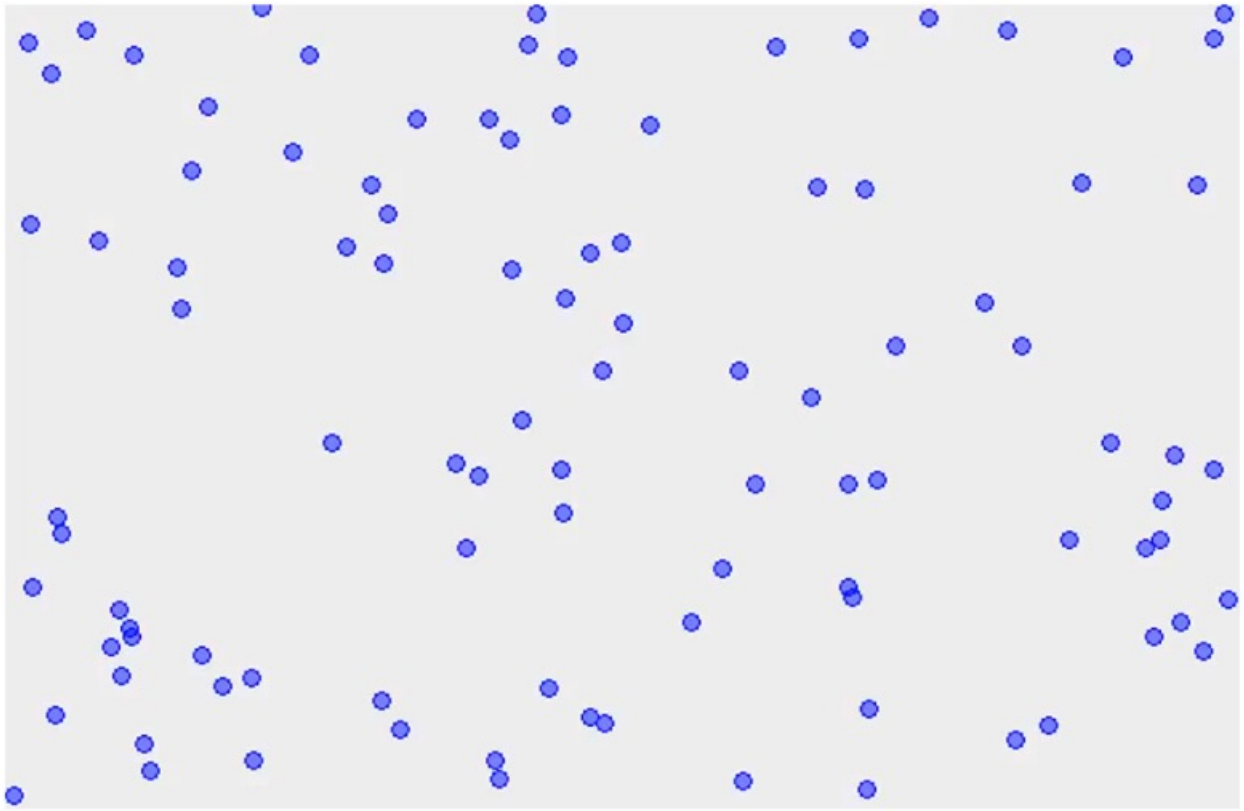
Wie funktioniert maschinelles Lernen?

Was ist Clustering?

Was nützt Clustering?

Wie funktioniert Clustering?

k-Means-Clustering: Beispiel



Nächster Schritt: **zufällige initiale Auswahl der k Clusterzentren** (hier $k = 6$)

Was ist maschinelles Lernen?

Was nützt maschinelles Lernen?

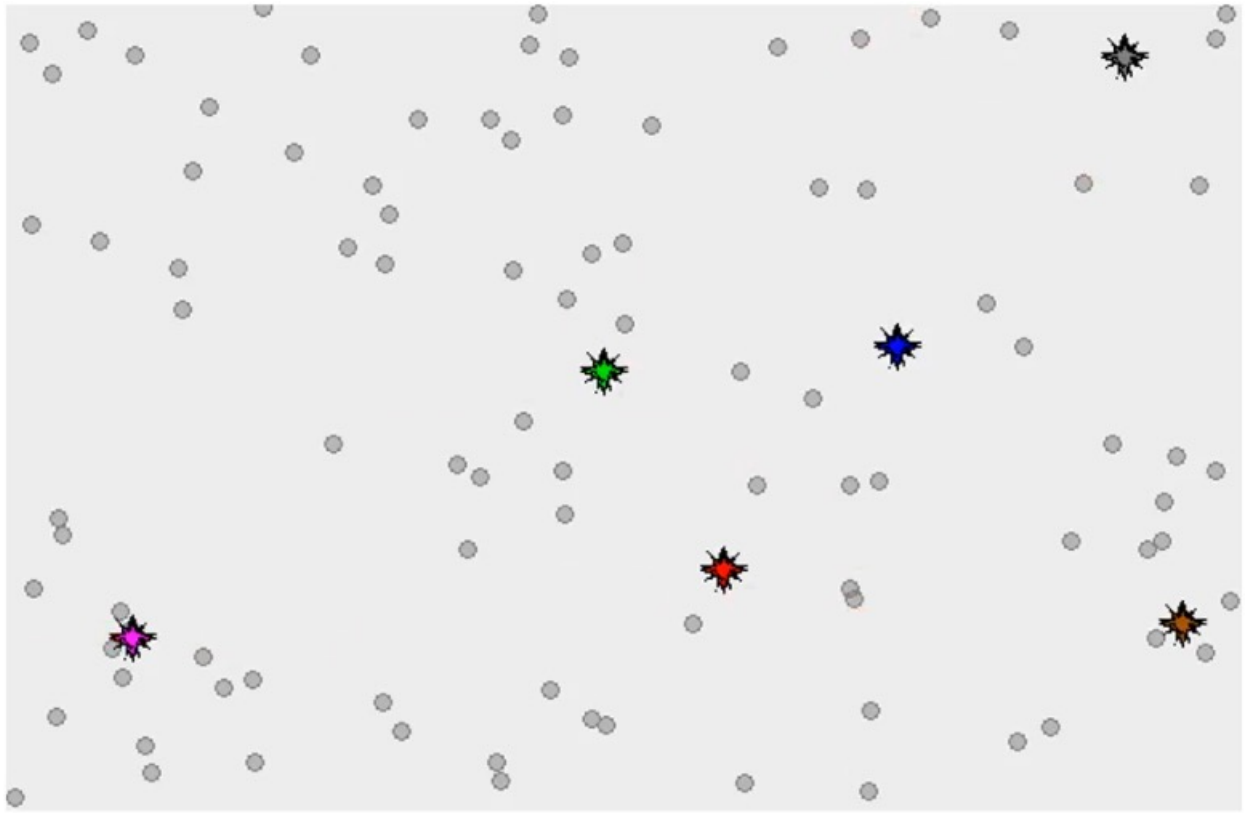
Wie funktioniert maschinelles Lernen?

Was ist Clustering?

Was nützt Clustering?

Wie funktioniert Clustering?

k-Means-Clustering: Beispiel



Nächster Schritt: berechne für jeden Datenpunkt die **Distanz zu jedem Clusterzentrum** und **weise ihn dem Cluster zu**, für das diese **Distanz am kleinsten** ist

Was ist maschinelles Lernen?

Was nützt maschinelles Lernen?

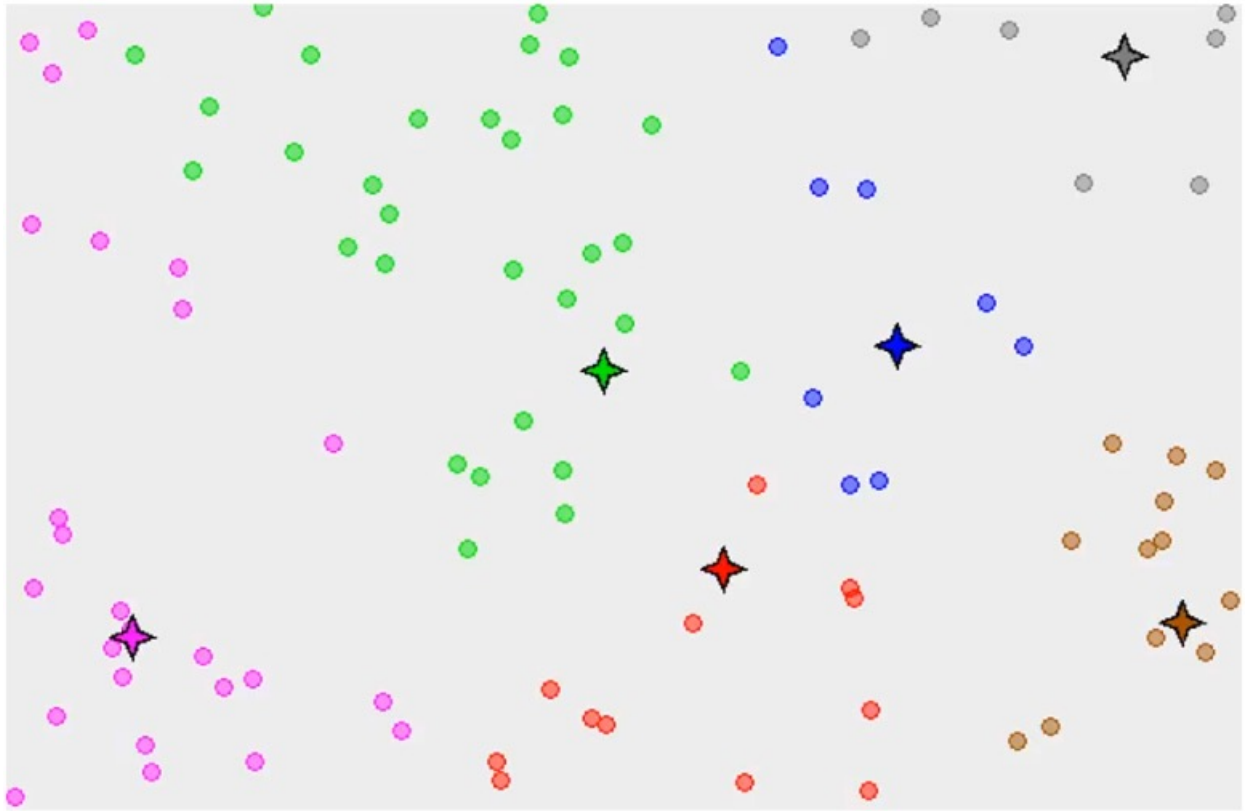
Wie funktioniert maschinelles Lernen?

Was ist Clustering?

Was nützt Clustering?

Wie funktioniert Clustering?

k-Means-Clustering: Beispiel



Nächster Schritt**: berechne für alle Cluster die Zentren neu**

Was ist maschinelles Lernen?

Was nützt maschinelles Lernen?

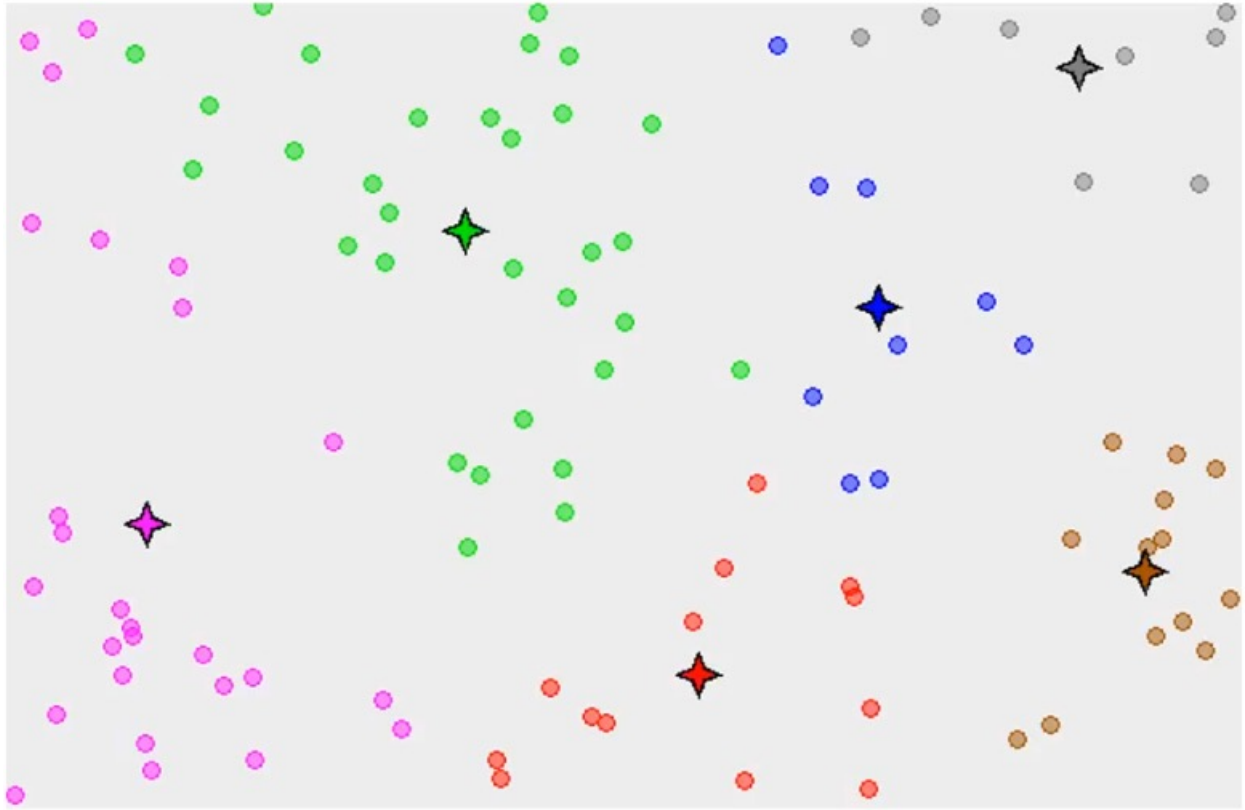
Wie funktioniert maschinelles Lernen?

Was ist Clustering?

Was nützt Clustering?

Wie funktioniert Clustering?

k-Means-Clustering: Beispiel



Nächster Schritt: berechne für jeden Datenpunkt die **Distanz zu jedem Clusterzentrum** und **weise ihn dem Cluster zu**, für das diese **Distanz am kleinsten** ist

Was ist maschinelles Lernen?

Was nützt maschinelles Lernen?

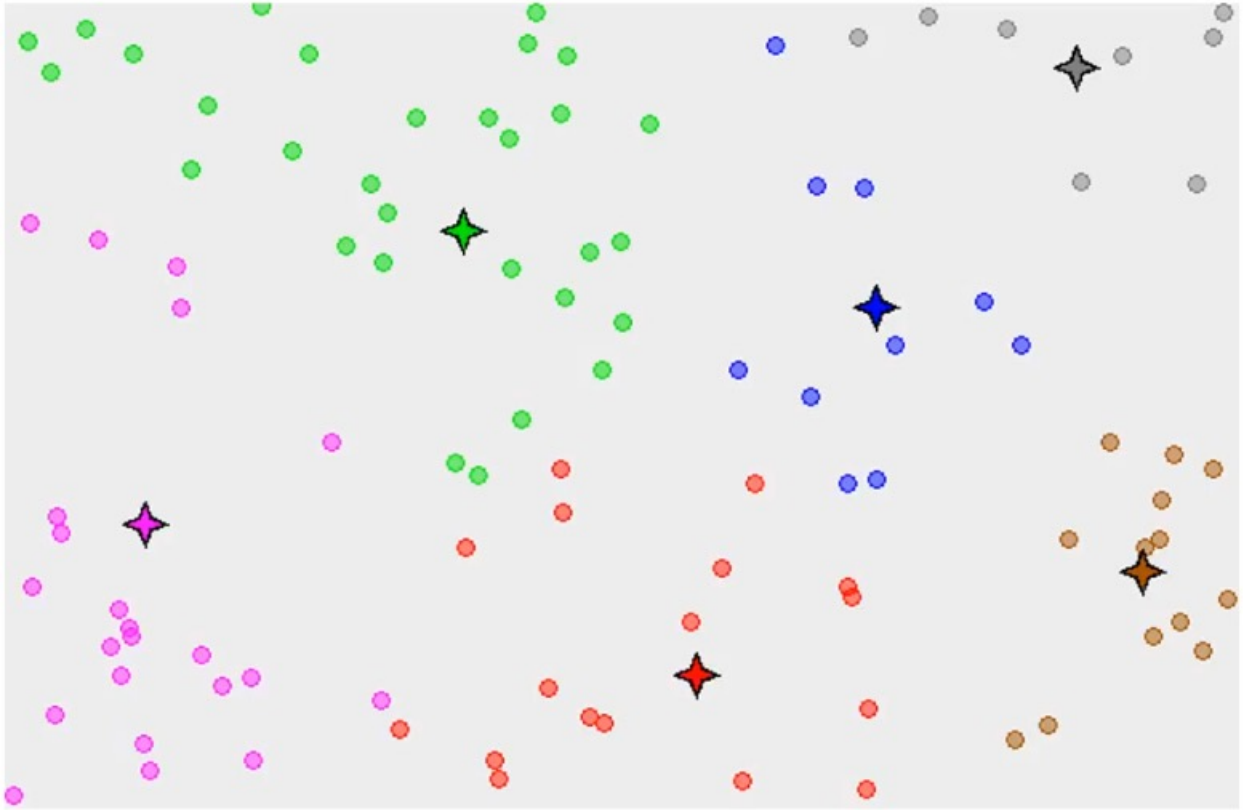
Wie funktioniert maschinelles Lernen?

Was ist Clustering?

Was nützt Clustering?

Wie funktioniert Clustering?

k-Means-Clustering: Beispiel



Nächster Schritt**: berechne für alle Cluster die Zentren neu**

Was ist maschinelles Lernen?

Was nützt maschinelles Lernen?

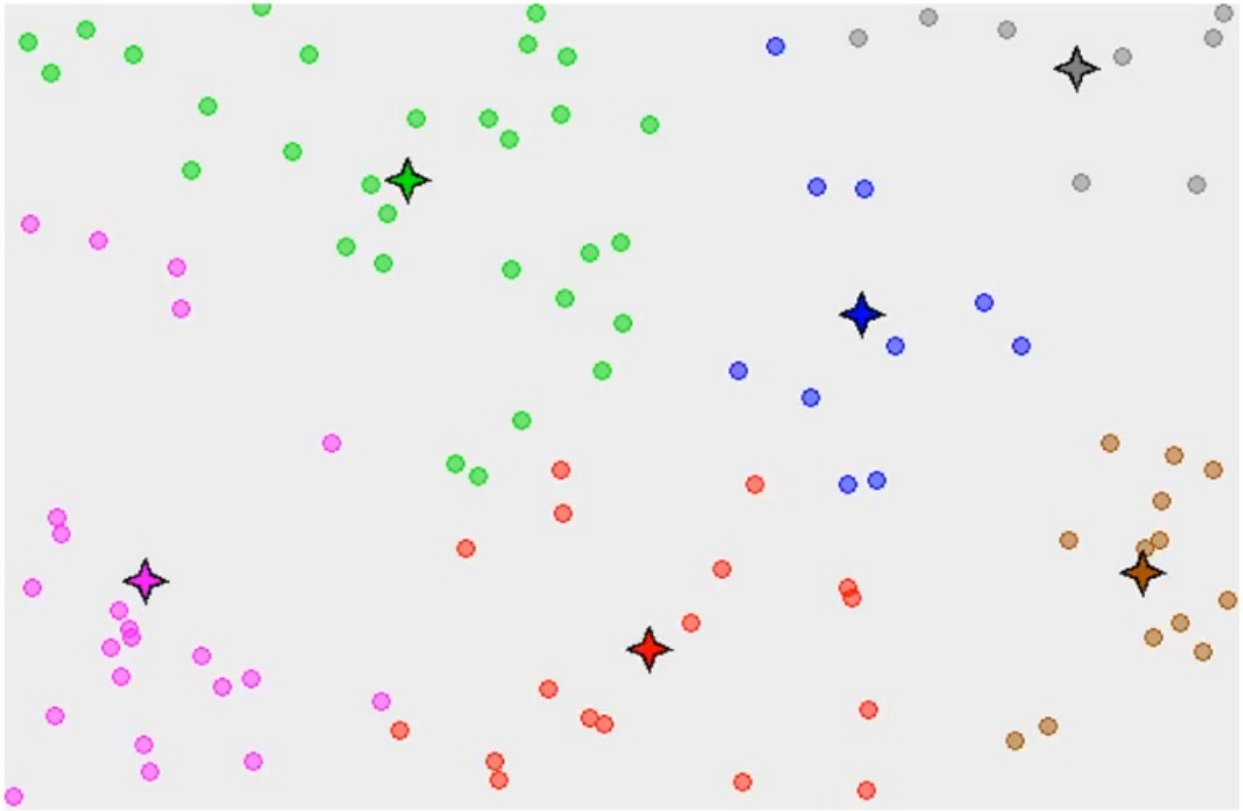
Wie funktioniert maschinelles Lernen?

Was ist Clustering?

Was nützt Clustering?

Wie funktioniert Clustering?

k-Means-Clustering: Beispiel

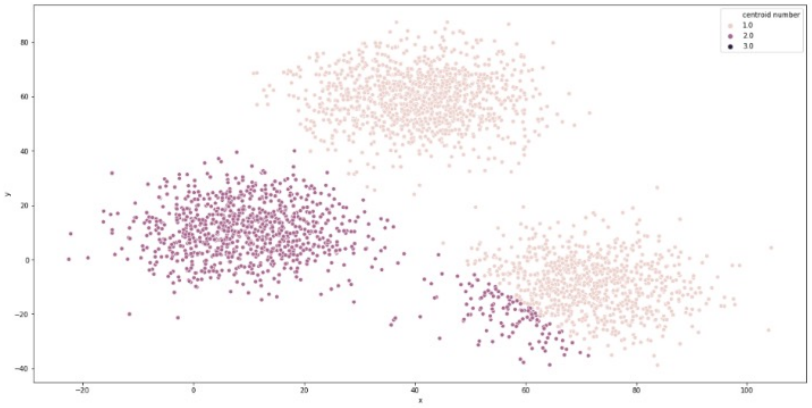


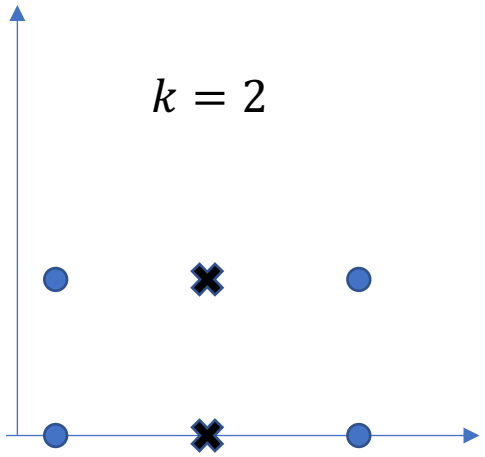
usw …

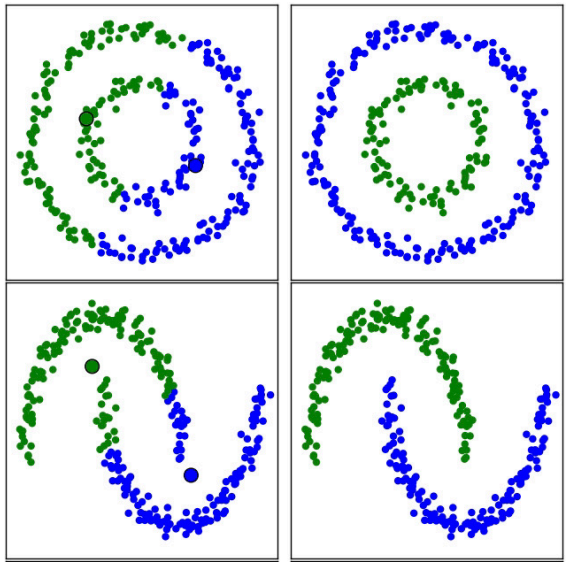
- Was ist maschinelles Lernen?
- Was nützt maschinelles Lernen?
- Wie funktioniert maschinelles Lernen?
- Was ist Clustering?
- Was nützt Clustering?
- **Wie funktioniert Clustering?**

k-Means-Clustering: Bewertung

Vorteile	Nachteile
einfach zu verstehen und zu programmieren	Anzahl der Cluster muss vorgegeben werden
auch für Big Data geeignet	anfällig gegenüber initialen Clusterzentren
effizient in der Laufzeit	anfällig gegenüber Ausreißern
nutzbar mit vielen verschiedenen Datentypen	anfällig gegenüber nicht-sphärischen Clustern







- Was ist maschinelles Lernen?
- Was nützt maschinelles Lernen?
- Wie funktioniert maschinelles Lernen?
- Was ist Clustering?
- Was nützt Clustering?
- **Wie funktioniert Clustering?**

Zusammenfassung

- **Machine Learning** befasst sich mit der **Erkennung unbekannter Muster in Daten**
- Es findet **Anwendung in zahlreichen Praxissituationen** und **steigert** bei korrekter Umsetzung den **Unternehmenserfolg**
- **Clustering** als eine konkrete Technik des Machine Learning zählt zu den **Methoden des unüberwachten Lernens** und **fasst ähnliche Objekte in Gruppen (sog. Cluster) zusammen**
- Clustering kann **somit im Marketing, im Finanzbereich und in medizinischen Anwendungsfällen verwendet** werden

- … lassen Sie uns das k-Means-Clustering doch mal mit einem realen Datensatz in Python implementieren!

Was ist maschinelles Lernen?

Was nützt maschinelles Lernen?

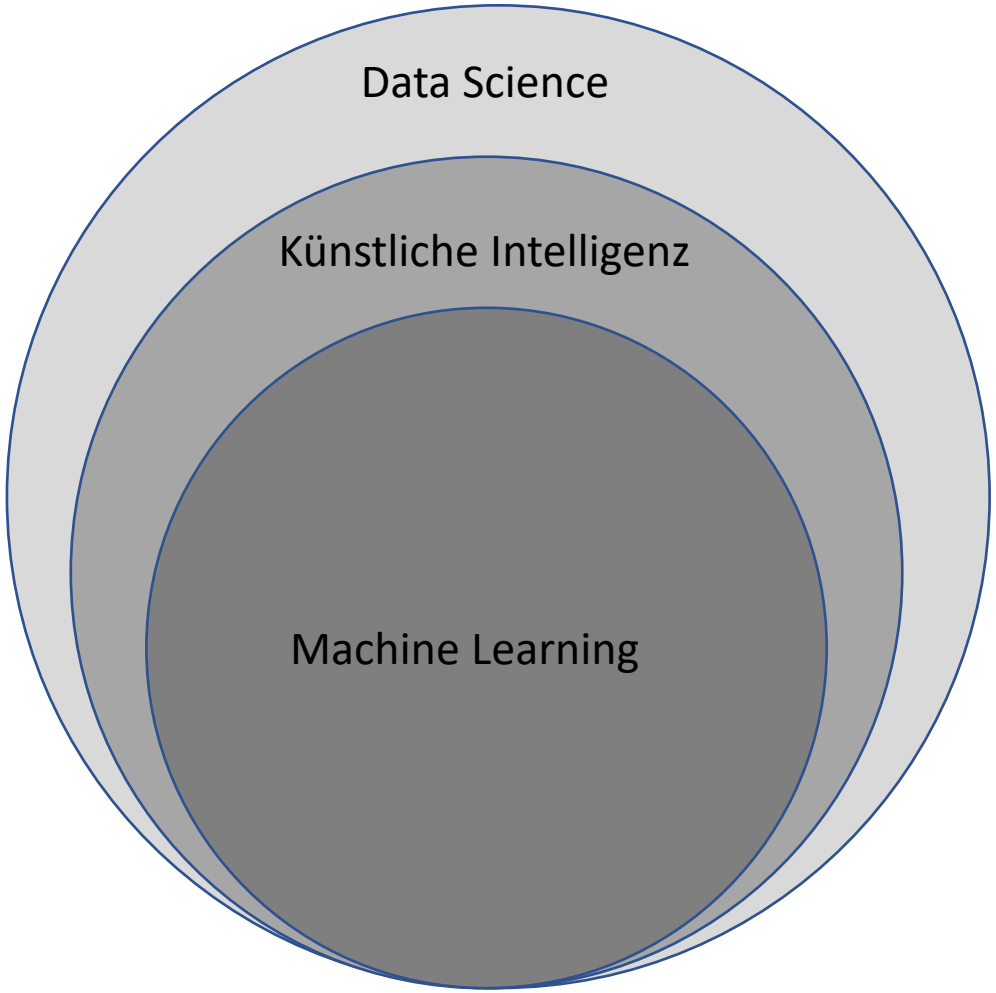
Wie funktioniert maschinelles Lernen?

Was ist Clustering?

Was nützt Clustering?

Wie funktioniert Clustering?

Einordnung der Thematik („Big Picture“)



Data Science (vgl. Dhar 2013, Leek 2014)

- *ein **interdisziplinäres Wissenschaftsfeld**, welches*
- ***wissenschaftlich fundierte Methoden**, Prozesse, Algorithmen und Systeme*
- *zur **Extraktion von Erkenntnissen, Mustern und Schlüssen** sowohl aus strukturierten als auch unstrukturierten **Daten** ermöglicht*

Künstliche Intelligenz (Bitkom e.V. 2021):

*Eigenschaft eines IT-Systems, **„menschenähnliche“ intelligente Verhaltensweisen** zu zeigen*

Was ist maschinelles Lernen?

Was nützt maschinelles Lernen?

Wie funktioniert maschinelles Lernen?

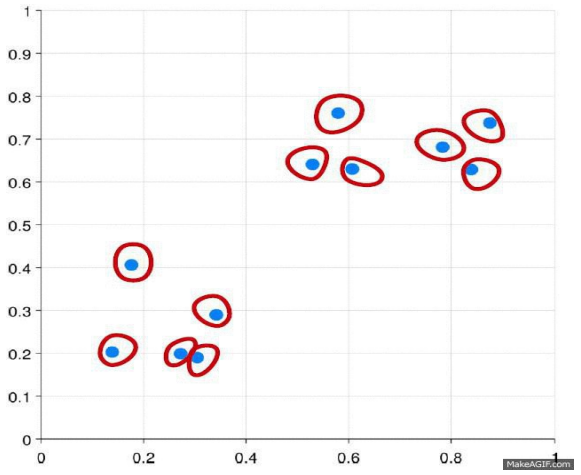
Was ist Clustering?

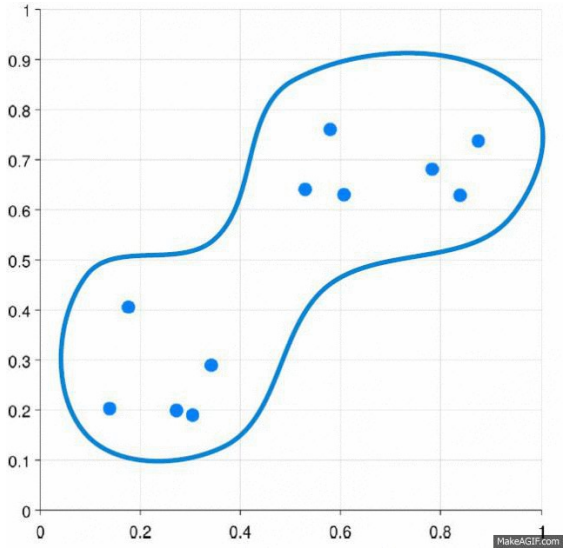
Was nützt Clustering?

Wie funktioniert Clustering?

Arten des Clustering

- Hierarchisches Clustering
  - agglomerativ:
    - jedes Objekt bildet zunächst ein Cluster
    - schrittweise Zusammenfassung zu Clustern
  - divisiv:
    - alle Objekte zunächst in einem Cluster
    - schrittweise Aufteilung in immer kleinere Cluster





Was ist maschinelles Lernen?

Was nützt maschinelles Lernen?

Wie funktioniert maschinelles Lernen?

Was ist Clustering?

Was nützt Clustering?

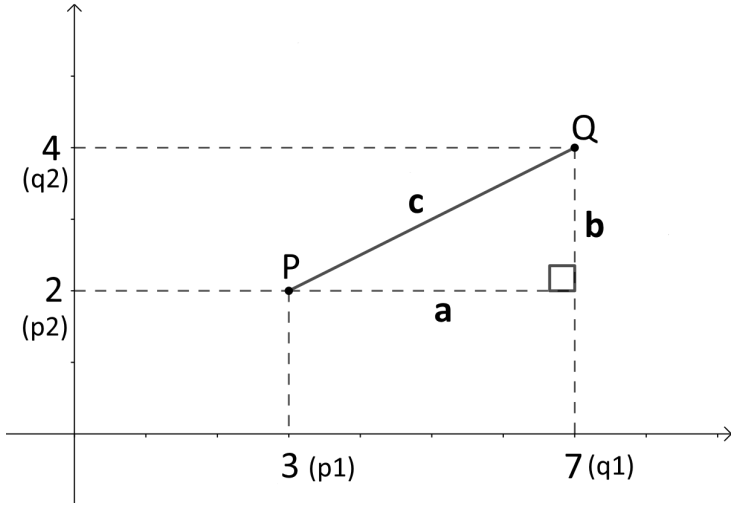
Wie funktioniert Clustering?

Wie wird der Abstand zwischen Punkten gemessen?

Euklidischer Abstand

- Herleitung über rechtwinkliges Dreieck, das durch die zwei Punkte entsteht

- Satz des Pythagoras $a^2 + b^2 = c^2$, wobei wir c genau die Strecke (=Verbindungslinie) zwischen den beiden Punkte bezeichnen



$$a^2 + b^2 = c^2$$

$$c = \sqrt{a^2 + b^2}$$

$$= \sqrt{(q1 - p1)^2 + (q2 - p2)^2}$$

$$= \sqrt{(7 - 3)^2 + (4 - 2)^2} \approx 4{,}47$$

Was ist maschinelles Lernen?

Was nützt maschinelles Lernen?

Wie funktioniert maschinelles Lernen?

Was ist Clustering?

Was nützt Clustering?

Wie funktioniert Clustering?

Anhang: Machine Learning-Algorithmen

Methodenklasse	Algorithmus	Funktionsweise
Überwachtes Lernen	Regression	Funktion abhängig von beschreibenden Variablen (z.B. lineare Regression: $y = a + bx$)
	k-nächste Nachbarn	Anzahl benachbarter Datenpunkte
	Neuronale Netze	Gewichte zwischen Neuronen, Aktivierungs- und Verlustfunktionen
	Random Forests	Entscheidungsregeln (z.B. „ja/nein“ oder „größer/kleiner“ entlang eines Entscheidungsbaums
Unüberwachtes Lernen	**Clustering**	**Daten werden in Gruppen ähnlicher Objekte zugeordnet.**
	Assoziationsanalyse	Berechnung von Korrelationen zwischen Objekten
	Hauptkomponentenanalyse	Beziehungen zwischen Akteuren innerhalb eines Netzwerks
	Soziale Netzwerkanalyse	Daten werden auf auf eine geringe Zahl möglichst aussagekräftiger Faktoren reduziert.

Was ist maschinelles Lernen?

Was nützt maschinelles Lernen?

Wie funktioniert maschinelles Lernen?

Was ist Clustering?

Was nützt Clustering?

Wie funktioniert Clustering?