Bitte beachten Sie die Lizenzhinweise auf der letzten Seite

Foto: Patricia Serna auf Unsplash

Distanzfunktionen

Wie ähnlich sind sich verschiedene Datenpunkte?

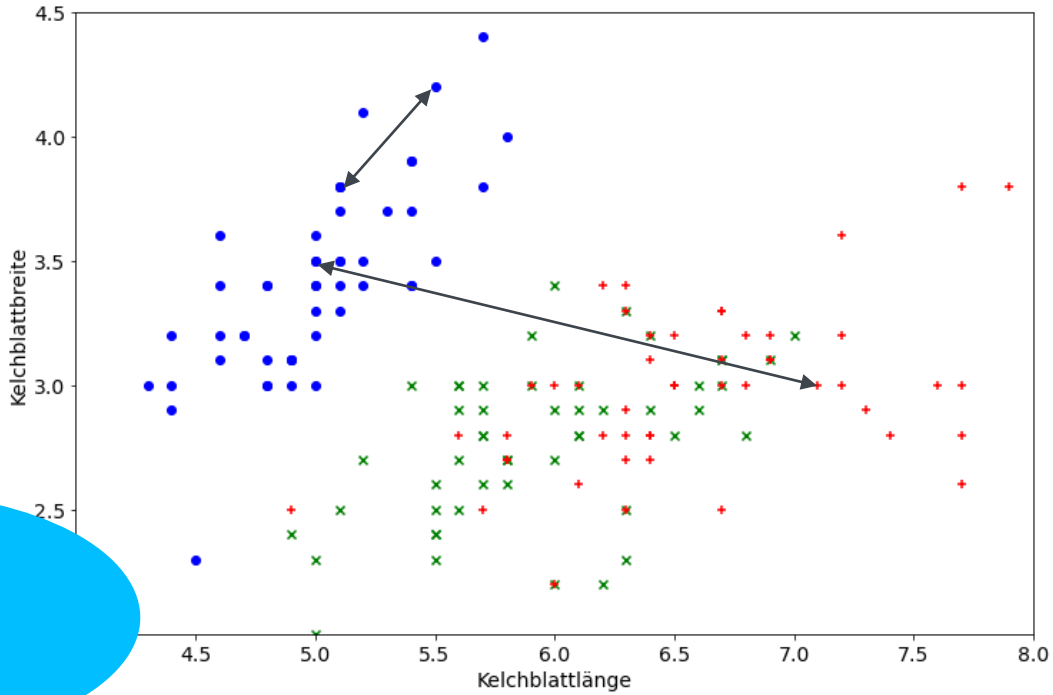
INNOVET
KI B³
Künstliche Intelligenz in die berufliche Bildung bringen

INNOVET
GEFÖRDERT VOM
Bundesministerium für Bildung und Forschung
bibb Bundesinstitut für Berufsbildung

Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

Ähnlich = nah

- Irisdaten
- Nahe beieinander liegende Datenpunkte repräsentieren ähnliche Blattmaße
- Wie misst man die Abstände?



Distanz-funktionen

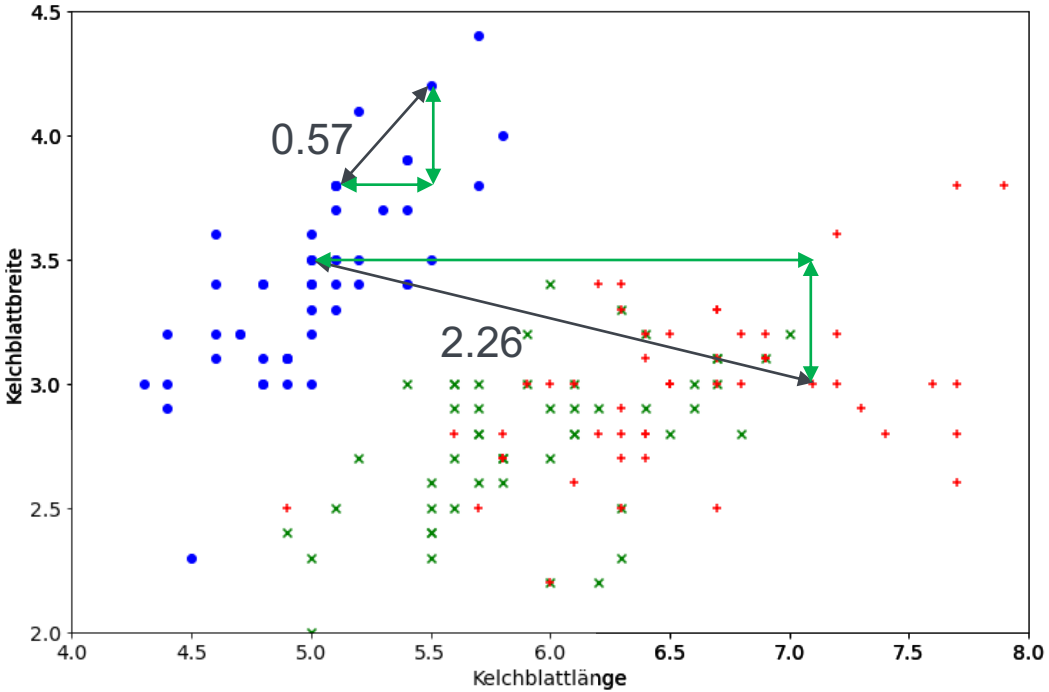
Distanzfunktionen berechnen den Abstand zwischen zwei Punkten. Sie werden deshalb auch Abstandsmaße genannt.

Euklidische Distanz

$$\sqrt{(5.5-5.1)^2+(4.2-3.8)^2} \approx 0.57$$

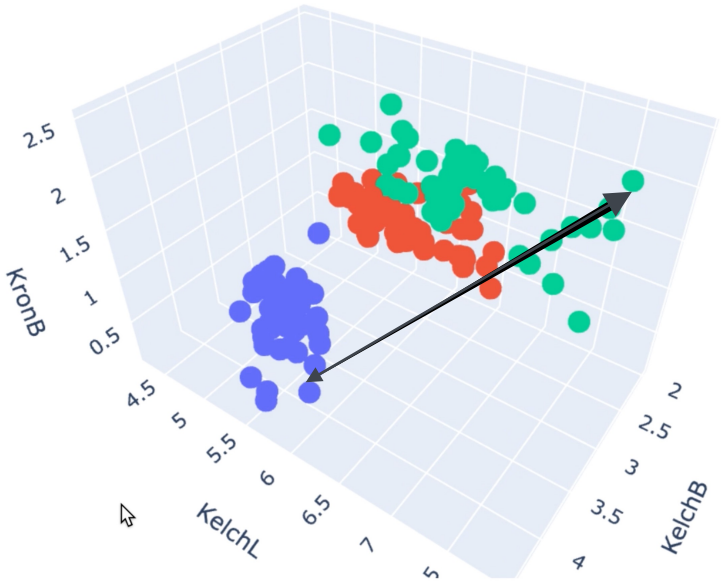
$$\sqrt{(7.2-5.0)^2+(3.0-3.5)^2} \approx 2.26$$

	Kelchblattlänge	Kelchblattbreite
44	5.1	3.8
33	5.5	4.2
40	5.0	3.5
129	7.2	3.0



Nähe im dreidimensionalen Raum

- Euklidischer Abstand in 3D
- Wie in 2D:
  - Differenzen der Koordinaten quadrieren
  - Addieren
  - Wurzel ziehen



Euklidische Distanz im dreidimensionalen Raum

- Distanz zwischen Datenpunkt 44 und 33

$$\sqrt{(5.5-5.1)^2+(4.2-3.8)^2+(0.2-0.4)^2} = 0.6$$

- Distanz zwischen Datenpunkt 40 und 129

$$\sqrt{(7.2-5.0)^2+(3.0-3.5)^2+(1.6-0.3)^2} \approx 2.60$$

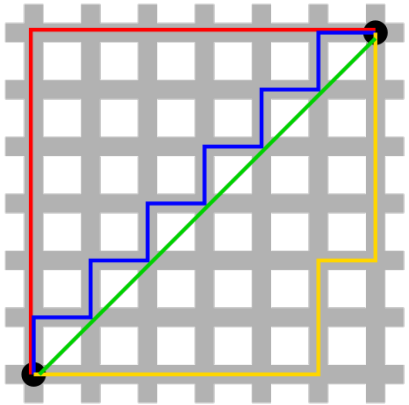
Genauso mit vier, fünf, … Dimensionen

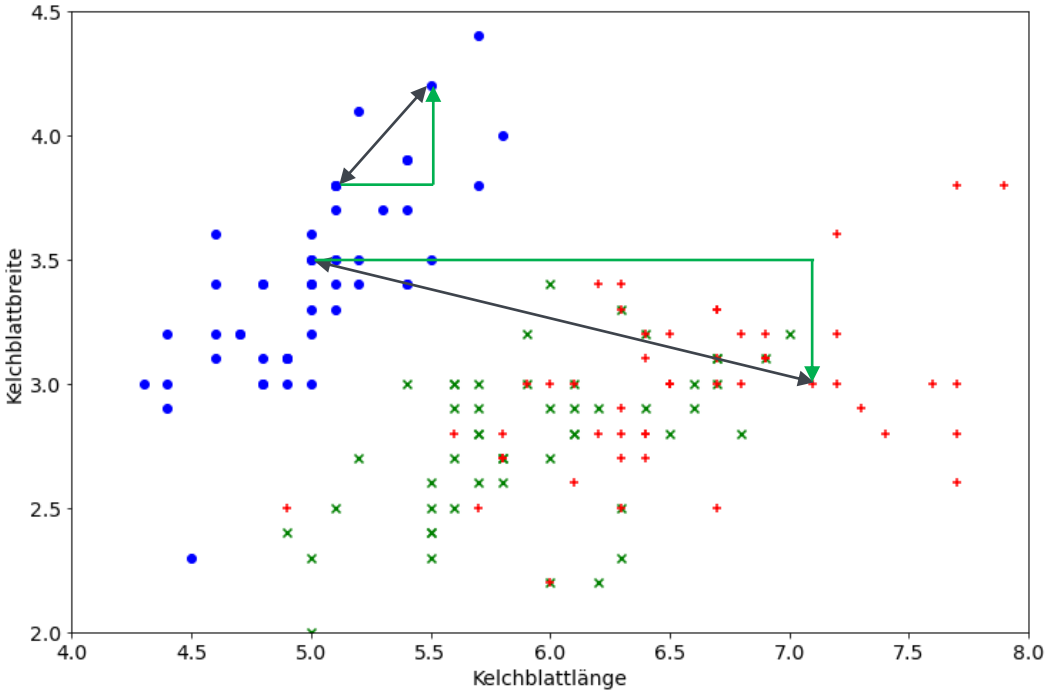
	Kelchblattlänge	Kelchblattbreite	Kronblattbreite	Kronblattlänge	Irisart
44	5.1	3.8	0.4	1.9	Iris-setosa
33	5.5	4.2	0.2	1.4	Iris-setosa
40	5.0	3.5	0.3	1.3	Iris-setosa
129	7.2	3.0	1.6	5.8	Iris-virginica

Die euklidische Distanz gibt die Länge der direkten Verbindung zwischen zwei Punkten in einem Koordinatensystem an.

Manhattan-Distanz

- gemessen parallel zu Achsen statt direkter Verbindung
- “Taxifahrer“-Distanz





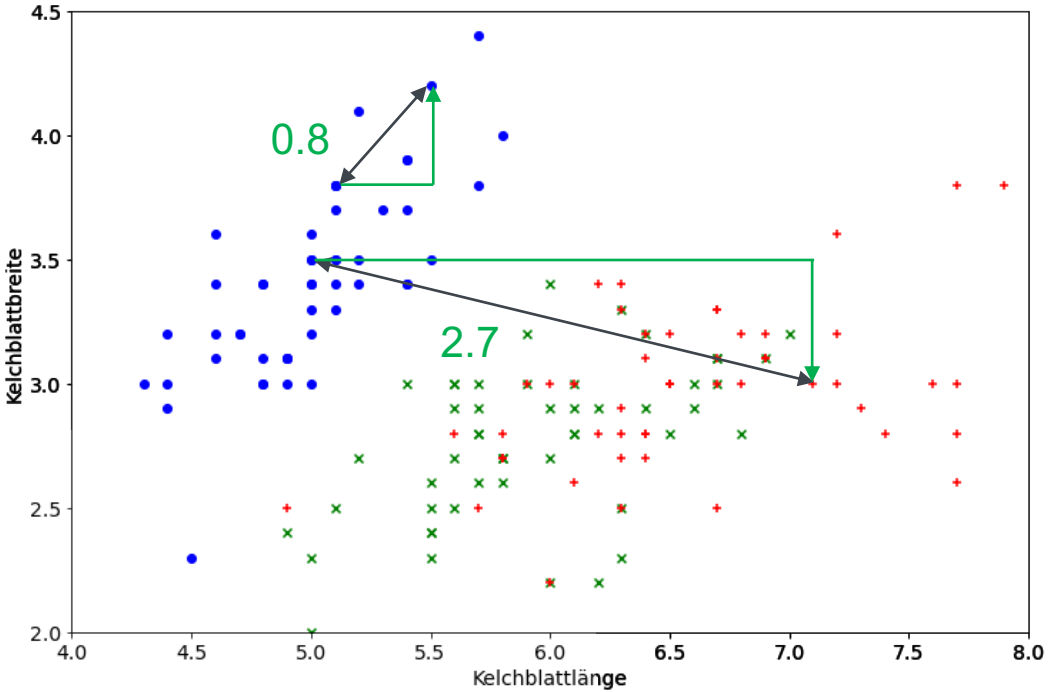
Manhattan-Distanz

Betrag der Differenz
(„absolute“ Differenz)

$$|5.5 - 5.1| + |4.2 - 3.8| = 0.8$$

$$|7.2 - 5.0| + |3.0 - 3.5| = 2.7$$

	Kelchblattlänge	Kelchblattbreite
44	5.1	3.8
33	5.5	4.2
40	5.0	3.5
129	7.2	3.0



Die Manhattan-Distanz berechnet den Abstand zwischen zwei Punkten als Länge des kürzesten Weges parallel zu den Achsen.

Sie wird auch Taxifahrer-Distanz genannt.

Die Manhattan-Distanz entspricht der Summe der absoluten Differenzen der Koordinaten.

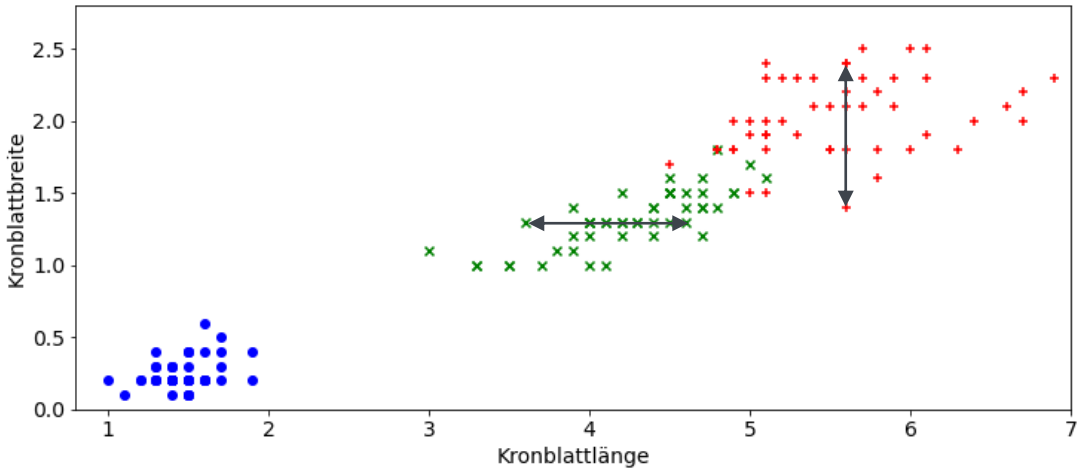
Question Set: Distanzen bei den Kelchblattmaßen

Multiple Choice: Euklidische Distanz im mehrdimensionalen Raum

Drag the Words: Euklidische- und Manhattan-Distanz

Beitrag verschiedener Dimensionen

- Sind Differenzen in verschiedenen Dimensionen gleich wichtig?
- Hier beide Distanzen 1
- Bei Kronblattbreite aber maximale Distanz 2.5
- Dagegen bei Kronblattlänge maximale Distanz 6

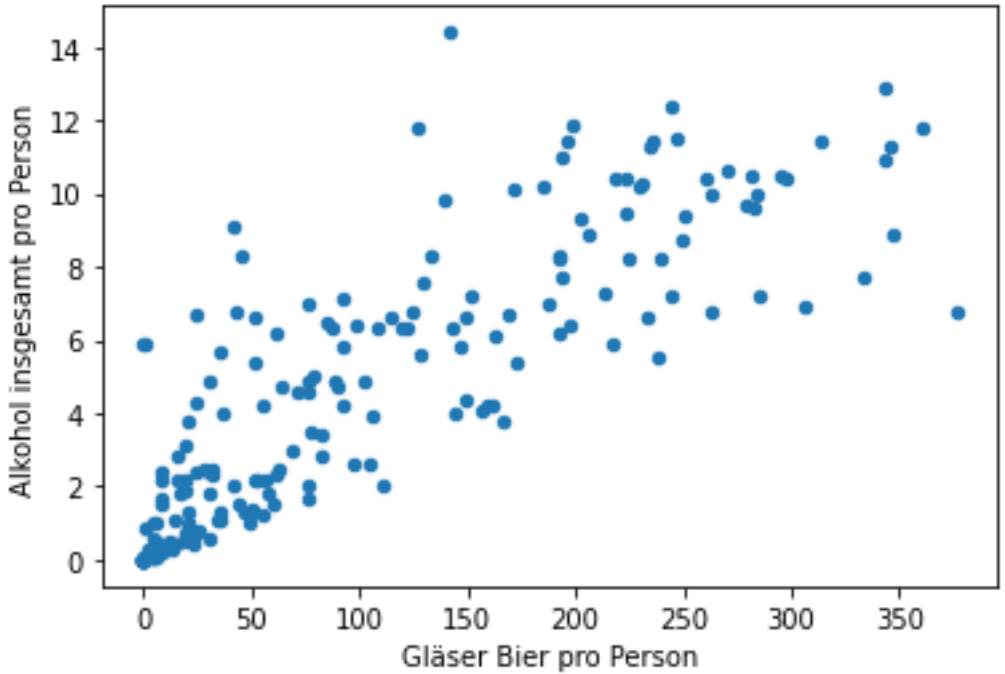


Beispiel: die “Alkoholdaten”

- Datenbank mit Daten zum Alkoholverbrauch in verschiedenen Ländern
- Erhältlich z.B. hier https://github.com/fivethirtyeight/data/tree/master/alcohol-consumption

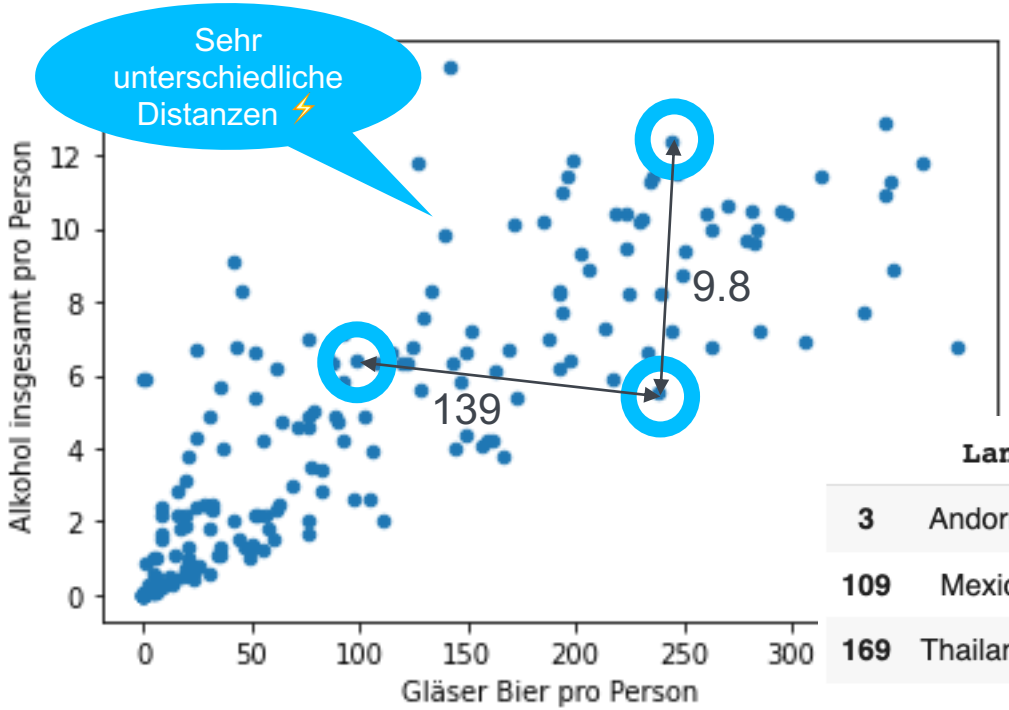
	Land	Gläser_Bier	Gläser_Schnaps	Gläser_Wein	reiner_Alkohol_insgesamt
0	Afghanistan	0	0	0	0.0
1	Albania	89	132	54	4.9
2	Algeria	25	0	14	0.7
3	Andorra	245	138	312	12.4
4	Angola	217	57	45	5.9

Bier und Alkohol pro Kopf in verschiedenen Ländern



- x-Achse: Werte von 0 bis 350
- y-Achse: Werte von 0 bis 14

Bier und Alkohol pro Kopf in verschiedenen Ländern



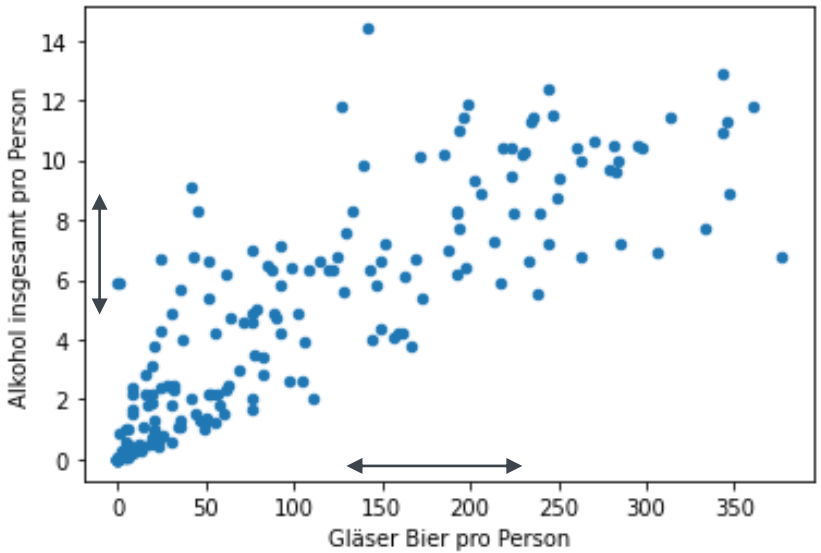
$$\sqrt{7^2 + 6.9^2} \approx 9.8$$

$$\sqrt{139^2 + 0.9^2} \approx 139$$

Differenz auf y-Achse trägt fast nichts bei

	Land	Gläser_Bier	reiner_Alkohol_insgesamt
3	Andorra	245	12.4
109	Mexico	238	5.5
169	Thailand	99	6.4

Distanzen im Verhältnis zur Standardabweichung



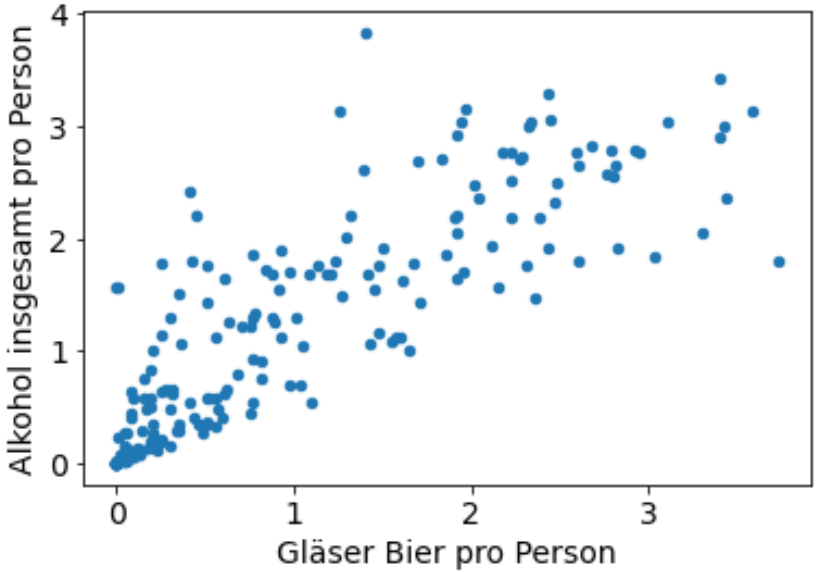
- Bei viel Streuung bedeutet eine kleine Distanz nicht viel
- Bei wenig Streuung bedeutet eine kleine Distanz viel
- Streuung erfasst durch Standardabweichung
  - Alkohol: ca. 3.8
  - Gläser Bier: ca. 101
- Hier: Veränderung um 3.8 bei Alkohol ähnlich wie Veränderung um 101 bei Bier

Multiple Choice: Kronblattmaße

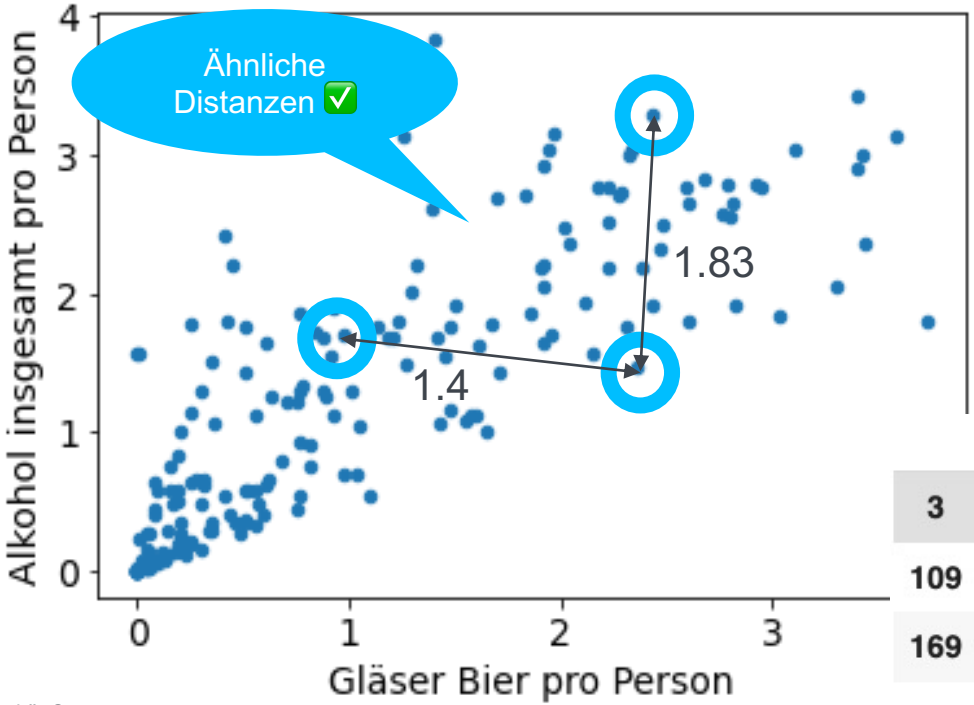
Multiple Choice: Streuung

Standardisierung von Daten

- Dividiere alle Werte einer Dimension durch ihre Standardabweichung
- Dadurch Standardabweichung = 1
- Differenz von 1 bedeutet dann: Differenz um eine Standardabweichung
- Hier: Veränderung um 3.8 bei Alkohol bzw. Veränderung um 101 bei Bier



Distanzen nach Division durch Standardabweichun



Differenz auf x-Achse trägt fast nichts bei ✅

$$\sqrt{0.07^2 + 1.83^2} \approx 1.83$$

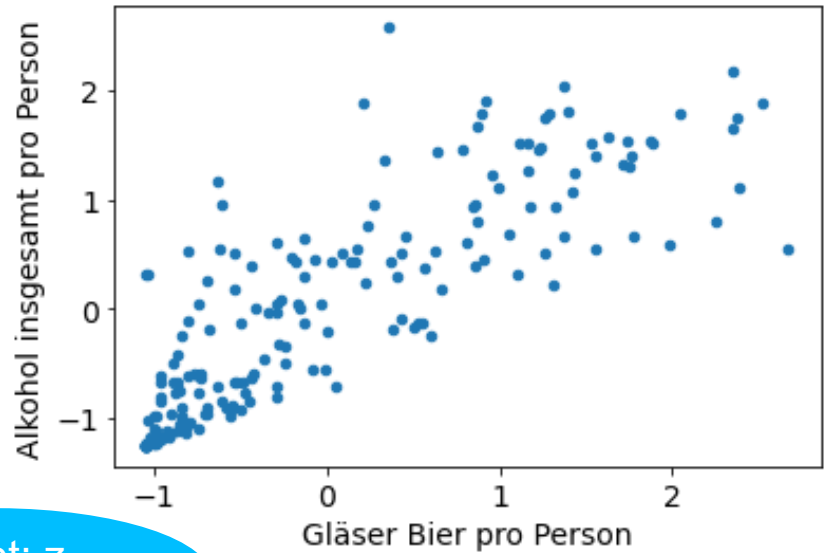
$$\sqrt{1.38^2 + 0.24^2} \approx 1.4$$

Differenz auf y-Achse trägt fast nichts bei ✅

	Gläser_Bier	reiner_Alkohol_insgesamt
3	2.428610	3.294797
109	2.359222	1.461402
169	0.981357	1.700540

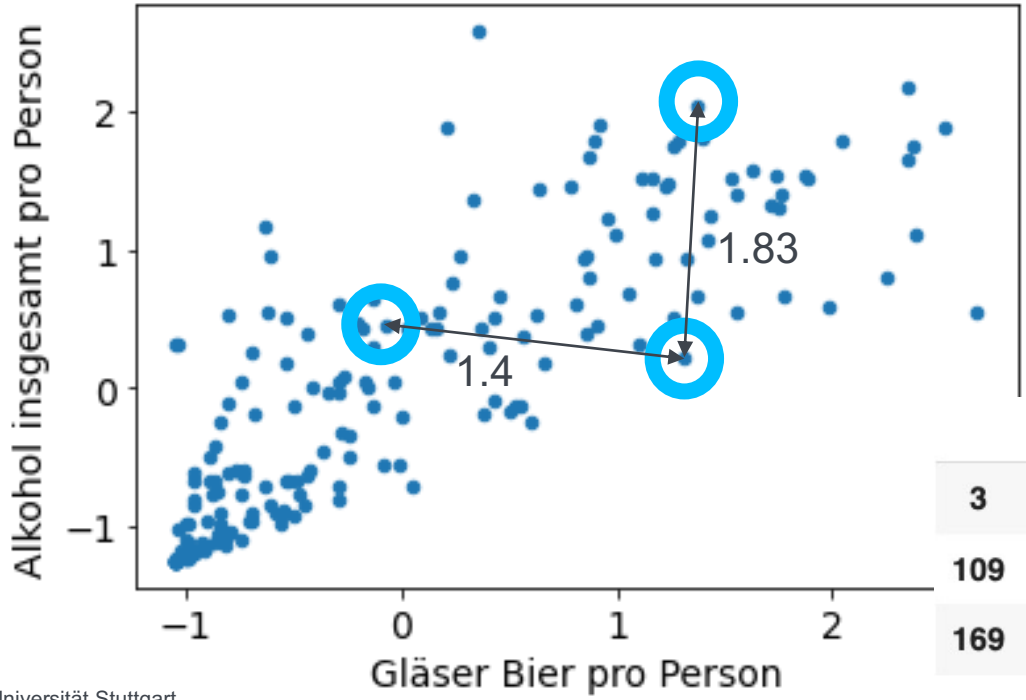
Standardisierung von Daten II

- Dividiere alle Werte einer Dimension durch ihre Standardabweichung
- Dadurch Standardabweichung = 1
- Differenz von 1 bedeutet dann: Differenz um eine Standardabweichung
- Häufig außerdem: ziehe den Mittelwert ab (hat keinen Einfluss auf die Distanz!)
- Dadurch neuer Mittelwert = 0
- **Ein z-Score > 0 bedeutet dann: mehr als der Durchschnitt**



Ergibt: z-Scores

Distanzen nach Standardisierung



$$\sqrt{0.07^2 + 1.83^2} \approx 1.83$$

$$\sqrt{1.38^2 + 0.24^2} \approx 1.4$$

Mittelwert verändert,
Differenzen unverändert!

	Gläser_Bier	reiner_Alkohol_insgesamt
3	1.376272	2.041419
109	1.306884	0.208024
169	-0.070981	0.447163

Standardisierung von Daten hat das Ziel, Werte leichter miteinander vergleichen zu können.

Durch die Standardisierung der Werte innerhalb aller Dimensionen tragen alle Dimensionen gleich zur Gesamt-Distanz bei.

Bei der Standardisierung teilt man üblicherweise durch die Standardabweichung und zieht den Mittelwert ab.

Die standardisierten Daten haben dadurch den Mittelwert 0 und die Standardabweichung 1.

Standardisierte Werte werden oft als z-scores oder z-Werte bezeichnet.

Ein z-score > 0 bedeutet, dass der Wert größer als der Durchschnitt ist.
Ein z-score < 0 bedeutet, dass der Wert kleiner als der Durchschnitt ist.

Drag the Words: Standardisierung

Single Choice: Standardisierung

Drag the Words: Distanzfunktionen

INNOVET

KI B³

Künstliche Intelligenz in die berufliche Bildung bringen

Dr. Antje Schweitzer

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung

Universität Stuttgart
Institut für Maschinelle Sprachverarbeitung
Institut für Software Engineering

IHK Industrie- und Handelskammer Reutlingen
Reutlingen | Tübingen | Zollernalb

IHK Region Stuttgart

IHK Industrie- und Handelskammer Karlsruhe

LMU Ludwig-Maximilians-Universität München

GEFÖRDERT VOM

INNOVET

Bundesministerium für Bildung und Forschung

bibb Bundesinstitut für Berufsbildung

Gefördert als InnoVET-Projekt aus Mitteln des Bundesministeriums für Bildung und Forschung.

Lizenzbestimmungen

“Distanzfunktionen" von Antje Schweitzer, KI B³ / Uni Stuttgart

Das Werk - mit Ausnahme der folgenden Elemente:

- Logos der Verbundpartner und des Förderprogramms
- im Quellenverzeichnis aufgeführte Medien

ist lizenziert unter:

CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/deed.de)
(Namensnennung 4.0 International)

Quellenverzeichnis

Titelfoto: Patricia Serna (https://unsplash.com/@sernaria), „colores a medida“, auf Unsplash (https://unsplash.com/photos/zPZ9vqqDNBA), ist lizenziert unter Unsplash-Lizenz (https://unsplash.com/license). Bildausschnitt verändert.

Seite 8, Psychonaut, “Manhattan-Distanz“, gemeinfrei, via Wikimedia Commons