ETH zürich

D-INFK Lehre

Unsupervised learning

Carlos Cotrini

Datenanalyse in der Physik

2024

Was wir heute lernen

- Wie geht man mit nicht-numerischen Daten um?
- Clustering
- Dimensionsreduktion (PCA)

ETH zürich

Kodierung kategorischer Daten

Wie geht man mit nicht-numerischen Daten um?

Class	Sex	Age	Survived?
Crew	F	Adult	1
Crew	F	Adult	1
First	M	Adult	0
First	M	Child	1
Second	F	Adult	0
Second	M	Child	1
Second	M	Adult	0

Ordinal encoding

Class	Sex	Age	Survived?
Crew	F	Adult	Y
Crew	F	Adult	Y
First	M	Adult	N
First	M	Child	Y
Second	F	Adult	N
Second	M	Child	Y
Second	M	Adult	N

Ordinal encoding

Class	Sex	Age	Survived?
1	F	Adult	Y
1	F	Adult	Y
2	M	Adult	N
2	M	Child	Y
3	F	Adult	N
3	M	Child	Y
3	M	Adult	N

Ordinal encoding

Class	Sex	Age	Survived?
1	1	Adult	Y
1	1	Adult	Y
2	2	Adult	N
2	2	Child	Y
3	1	Adult	N
3	2	Child	Y
3	2	Adult	N

Ordinal encoding

Class	Sex	Age	Survived?
1	1	1	Y
1	1	1	Y
2	2	1	N
2	2	2	Y
3	1	1	N
3	2	2	Y
3	2	1	N

Ordinal encoding

- Vorteile:
  - Einfach und effizient

- Nachteile:
  - Führt einen Begriff der Nähe ein, der falsch sein könnte. Zum Beispiel ist "Crew" (1) näher an "First" (2) als an "Second" (3).

Mean encoding

Class	Sex	Age	Survived?
Crew	F	Adult	Y
Crew	F	Adult	Y
First	M	Adult	N
First	M	Child	Y
Second	F	Adult	N
Second	M	Child	Y
Second	M	Adult	N

Mean encoding

Class	Sex	Age	Survived?
1.0	F	Adult	Y
1.0	F	Adult	Y
0.5	M	Adult	N
0.5	M	Child	Y
0.33	F	Adult	N
0.33	M	Child	Y
0.33	M	Adult	N

Mean encoding

Class	Sex	Age	Survived?
1.0	0.66	Adult	Y
1.0	0.66	Adult	Y
0.5	0.5	Adult	N
0.5	0.5	Child	Y
0.33	0.66	Adult	N
0.33	0.5	Child	Y
0.33	0.5	Adult	N

Mean encoding

Class	Sex	Age	Survived?
1.0	0.66	0.4	Y
1.0	0.66	0.4	Y
0.5	0.5	0.4	N
0.5	0.5	1.0	Y
0.33	0.66	0.4	N
0.33	0.5	1.0	Y
0.33	0.5	0.4	N

Mean encoding

- Vorteile:
  - Die Kodierung gibt dem Merkmalswert eine Bedeutung. Zum Beispiel bedeutet "Child", dass Sie in der Datenmenge mit einer Wahrscheinlichkeit von 1 überlebt haben.
- Nachteile:
  - Die Kodierung lässt Informationen aus den Features zu den Beispielen durchsickern, diese Durchsickerung führt zu Overfitting.

One-hot encoding

Class	Sex	Age	Survived?
Crew	F	Adult	Y
Crew	F	Adult	Y
First	M	Adult	N
First	M	Child	Y
Second	F	Adult	N
Second	M	Child	Y
Second	M	Adult	N

One-hot encoding

Class	Sex	Age	Survived?
Crew	F	Adult	Y
Crew	F	Adult	Y
First	M	Adult	N
First	M	Child	Y
Second	F	Adult	N
Second	M	Child	Y
Second	M	Adult	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex	Age	Survived?
1	0	0	F	Adult	Y
1	0	0	F	Child	Y
0	1	0	M	Adult	N
0	1	0	M	Child	Y
0	0	1	F	Adult	N
0	0	1	M	Child	Y
0	0	1	M	Adult	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex	Age	Survived?
1	0	0	F	Adult	Y
1	0	0	F	Child	Y
0	1	0	M	Adult	N
0	1	0	M	Child	Y
0	0	1	F	Adult	N
0	0	1	M	Child	Y
0	0	1	M	Adult	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex	Age	Survived?
1	0	0	F	Adult	Y
1	0	0	F	Child	Y
0	1	0	M	Adult	N
0	1	0	M	Child	Y
0	0	1	F	Adult	N
0	0	1	M	Child	Y
0	0	1	M	Adult	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex = F	Sex = M	Age	Survived ?
1	0	0	1	0	Adult	Y
1	0	0	1	0	Child	Y
0	1	0	0	1	Adult	N
0	1	0	0	1	Child	Y
0	0	1	1	0	Adult	N
0	0	1	0	1	Child	Y
0	0	1	1	0	Adult	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex = F	Sex = M	Age	Survived ?
1	0	0	1	0	Adult	Y
1	0	0	1	0	Child	Y
0	1	0	0	1	Adult	N
0	1	0	0	1	Child	Y
0	0	1	1	0	Adult	N
0	0	1	0	1	Child	Y
0	0	1	1	0	Adult	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex = F	Sex = M	Age	Survived ?
1	0	0	1	0	Adult	Y
1	0	0	1	0	Child	Y
0	1	0	0	1	Adult	N
0	1	0	0	1	Child	Y
0	0	1	1	0	Adult	N
0	0	1	0	1	Child	Y
0	0	1	1	0	Adult	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex = F	Sex = M	Age = Child	Age = Adult	Survived ?
1	0	0	1	0	0	1	Y
1	0	0	1	0	1	0	Y
0	1	0	0	1	0	1	N
0	1	0	0	1	1	0	Y
0	0	1	1	0	0	1	N
0	0	1	0	1	1	0	Y
0	0	1	1	0	0	1	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex = F	Sex = M	Age = Child	Age = Adult	Survived ?
1	0	0	1	0	0	1	Y
1	0	0	1	0	1	0	Y
0	1	0	0	1	0	1	N
0	1	0	0	1	1	0	Y
0	0	1	1	0	0	1	N
0	0	1	0	1	1	0	Y
0	0	1	1	0	0	1	N

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex = F	Sex = M	Age = Child	Age = Adult	Survived ?
1	0	0	1	0	0	1	Y
1	0	0	1	0	1	0	Y
0	1	0	0	1	0	1	N
0	1	0	0	1	1	0	Y
0	0	1	1	0	0	1	N
0	0	1	0	1	1	0	Y
0	0	1	1	0	0	1	N

Compute the one-hot encoding for the following passengers:

Class	Sex	Age
Crew	F	Adult
First	M	Child
Second	F	Adult

One-hot encoding

Class = Crew	Class = First	Class = Second	Sex = F	Sex = M	Age = Child	Age = Adult	Survived ?
1	0	0	1	0	0	1	Y
1	0	0	1	0	1	0	Y
0	1	0	0	1	0	1	N
0	1	0	0	1	1	0	Y
0	0	1	1	0	0	1	N
0	0	1	0	1	1	0	Y
0	0	1	1	0	0	1	N

Berechnen Sie die One-Hot-Encoding für die folgenden Passagiere:

Class	Sex	Age
Crew	F	Adult
First	M	Child
Second	F	Adult

1	0	0	1	0	0	1
0	1	0	0	1	1	0
0	0	1	1	0	0	1

One-hot encoding

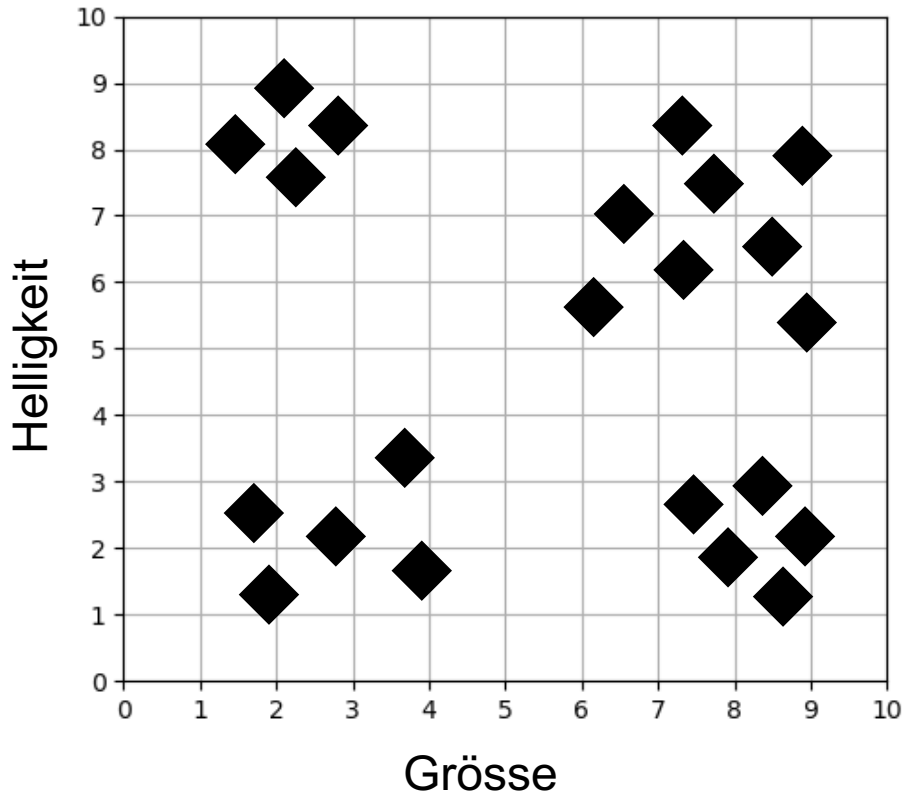
- Vorteile:
  - Die Kodierung induziert keine Art von "Nähe" zwischen Merkmalswerten. "Crew" (1, 0, 0), "First" (0, 1, 0) und "Second" (0, 0, 1) sind gleich weit entfernt.

- Nachteile:
  - Wenn es zu viele Merkmalswerte gibt, explodiert die Anzahl der Merkmale, was das Training des Modells verlangsamt.

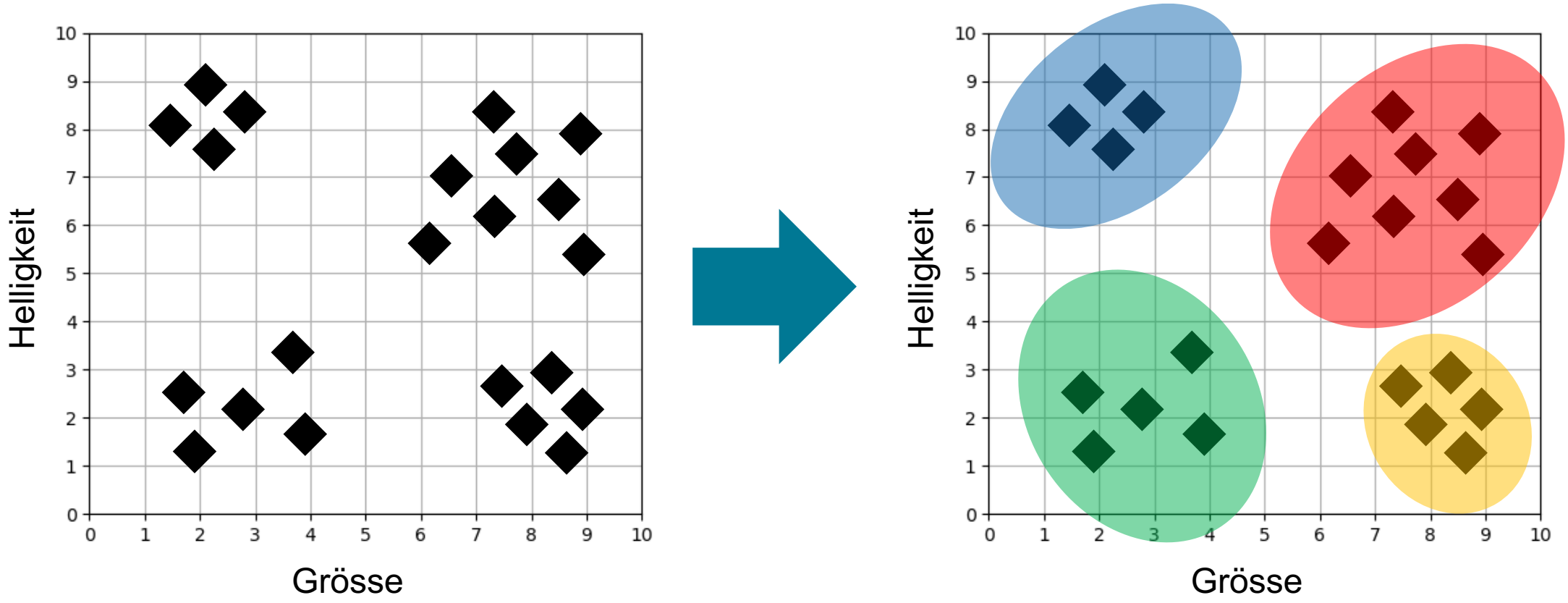
ETH zürich

Clustering

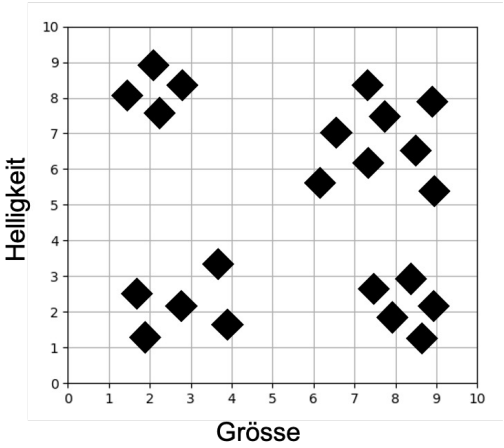
Clustering



Clustering



Trainieren von Modellen in ML



1. Datensatz

$$\mathcal{H} = \{ (\theta_1, \dots, \theta_k, c) \mid \theta_1, \dots, \theta_k \in \mathbb{R}^d \text{ und c ist eine Zuweisung} \}$$

2. Modell auswählen

3. Verlustfunktion

4. Training

f^*

5. Validierung

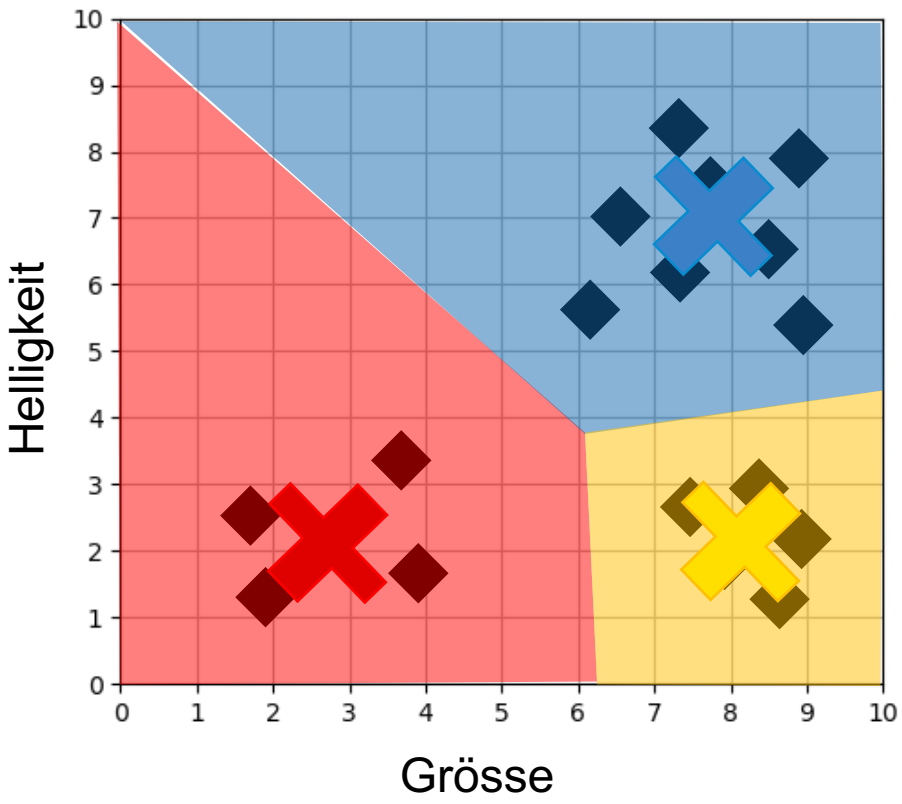
Schätzer und Modell

Ein Schätzer für Clustering besteht aus zwei Objekten:

- K Zentroide $\theta_1, \theta_2, \ldots, \theta_K \in \mathbb{R}^d$:
  - Der Durchschnitt aller Punkte innerhalb eines Clusters.

- Zuweisung $c: \mathbb{R}^d \to \{1, \ldots, K\}$
  - Jeder Punkt wird dem nächsten Zentroid-Cluster zugewiesen.

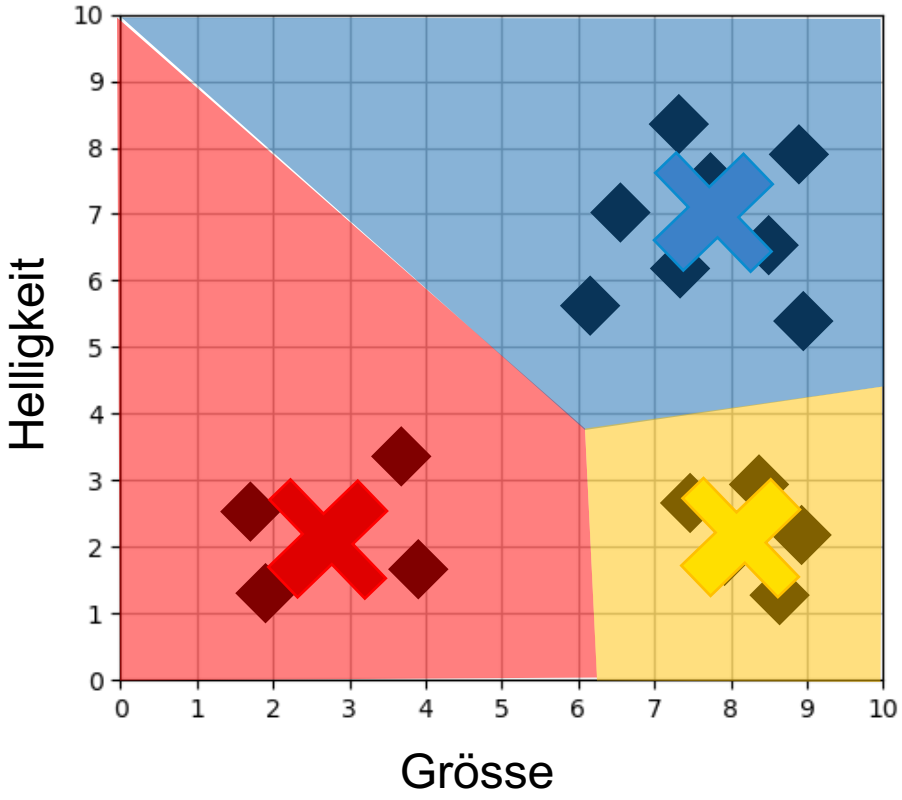
Das Modell besteht aus aller möglichen Schätzer mit genau K Zentroiden.



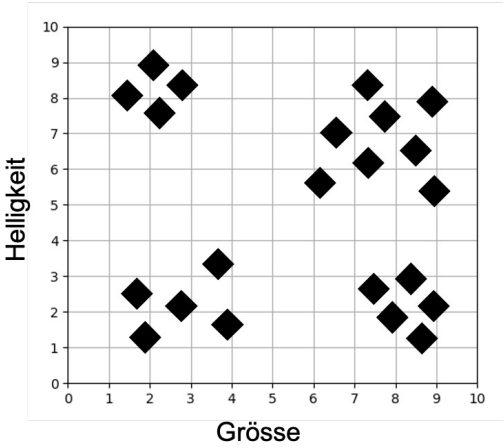
Schätzer und Modell

Das Modell besteht aus aller möglichen Schätzer mit genau K Zentroiden.

Sie müssen die Anzahl K von Clustern selbst festlegen!



Trainieren von Modellen in ML



1. Datensatz

$$\mathcal{H} = \{ (\theta_1, \dots, \theta_k, c) \mid \theta_1, \dots, \theta_k \in \mathbb{R}^d \text{ und c ist eine Zuweisung} \}$$

2. Modell auswählen

3. Verlustfunktion

4. Training

f^*

5. Validierung

Die Verlustfunktion

- Wir finden den Schätzer durch die Minimierung einer geeigneten Verlustfunktion.
- Diese bestraft die Summe der quadratischen Abstände von jedem Punkt zu seinem zugewiesenen Zentroid.
- Es kann gezeigt werden, dass diese Funktion minimiert wird, wenn jeder Zentroid im Schwerpunkt jedes Clusters liegt.

Formalisierung der Verlustfunktion

Denken Sie daran, dass die Verlustfunktion die Summe der quadratischen Abstände von jedem Punkt zu seinem nächsten Zentroid bestraft.

Formalisierung der Verlustfunktion

Denken Sie daran, dass die Verlustfunktion die Summe der quadratischen Abstände von jedem Punkt zu seinem nächsten Zentroid bestraft.

Sei $K \in \mathbb{N}$. Es bezeichnet die Anzahl der Cluster, die wir berechnen möchten.

Sei $\mathrm{X} = \{x_1, x_2, \ldots, x_n\} \subseteq \mathbb{R}^d$ eine Menge von Punkten.

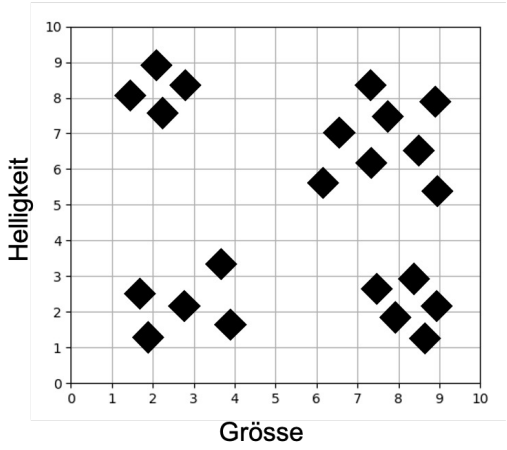
Seien $\{\theta_1, \theta_2, \ldots, \theta_K\} \subseteq \mathbb{R}^d$ die Zentroide.

Sei $c : \mathbb{R}^d \to \{1,2, \ldots, K\}$ die Zuweisung. Beachten Sie, dass $c(x) = \ell$, wenn θ_ℓ das Zentroid ist, das x am nächsten liegt. Daher ist $\theta_{c(x_i)}$ das Zentroid, das x_i am nächsten liegt.

Die Verlustfunktion ist dann

$$\mathcal{L}(X, \theta_1, \ldots, \theta_K, c) = \sum_{i \leq n} \left\| x_i - \theta_{c(x_i)} \right\|^2$$

Trainieren von Modellen in ML



1. Datensatz

$$\mathcal{H} = \{ (\theta_1, \dots, \theta_k, c) \mid \theta_1, \dots, \theta_k \in \mathbb{R}^d \text{ und c ist eine Zuweisung} \}$$

2. Modell auswählen

3. Verlustfunktion

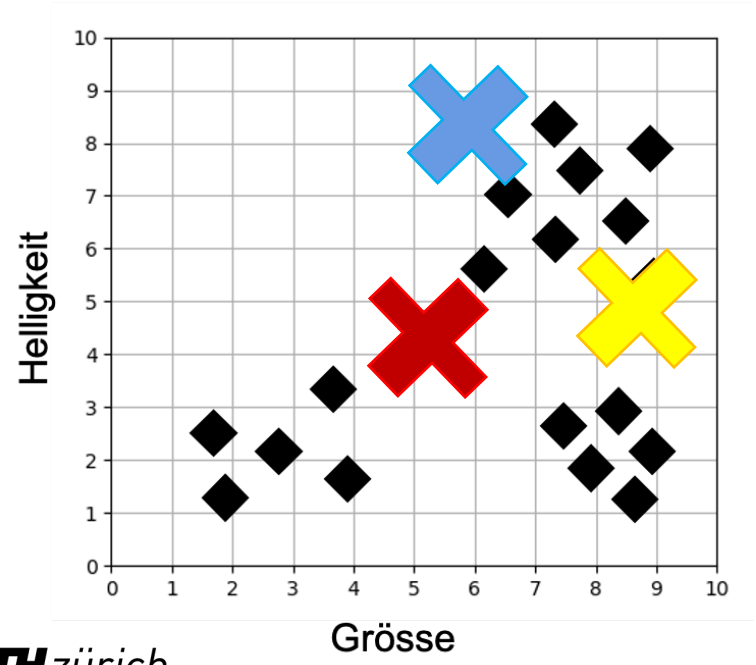
4. Training

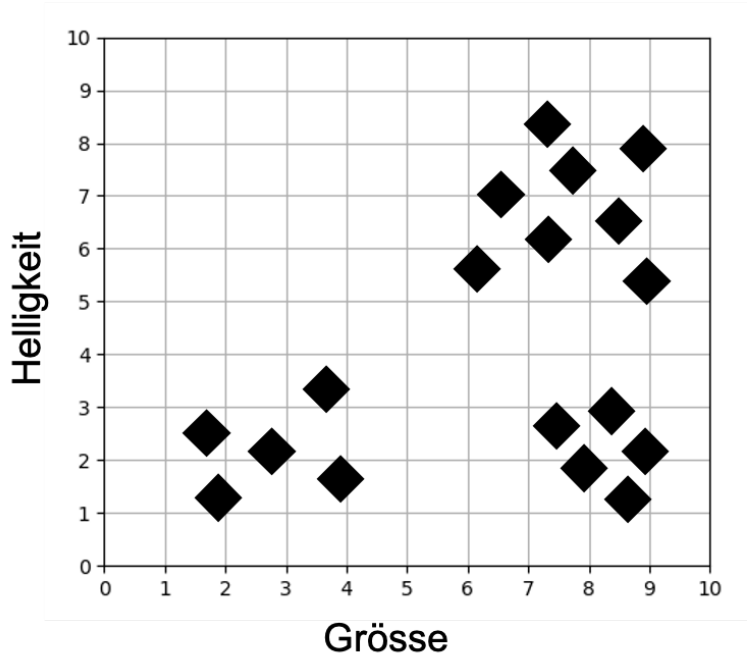
f^*

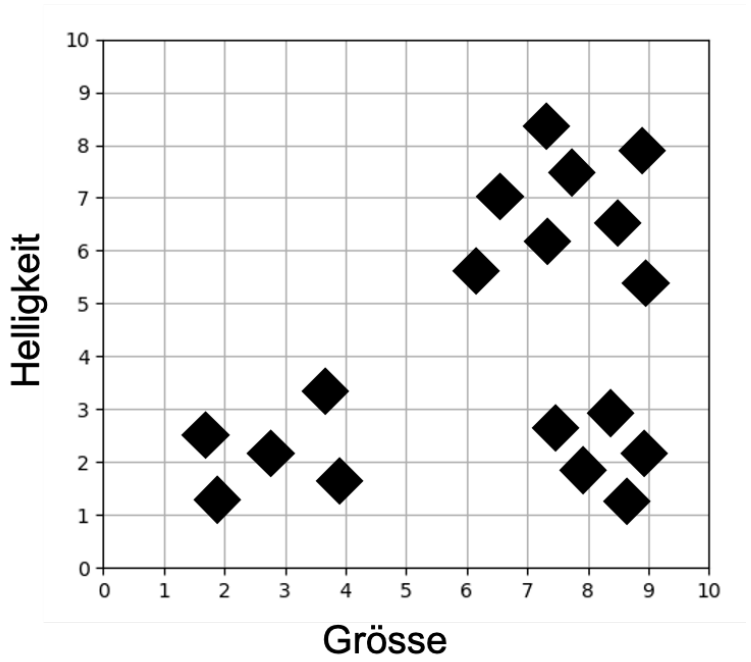
5. Validierung

Der Trainingsalgorithmus

- **Zufällig Initialisierung** der Zentroide
- **Aktualisierung der Zuweisung**
- **Aktualisierung der Zentroide**. Jedes Zentroid wird auf den Durchschnitt aller Punkte gesetzt, die diesem Zentroid-Cluster zugewiesen sind.

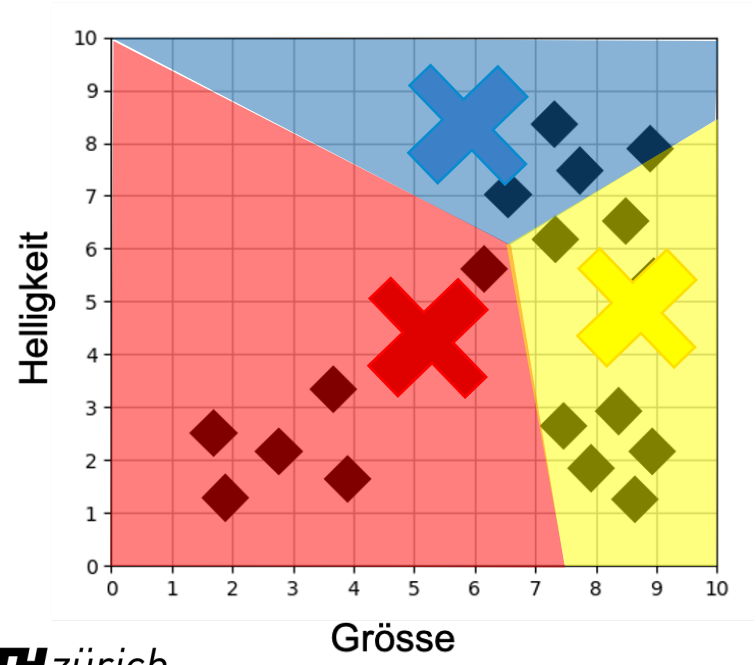


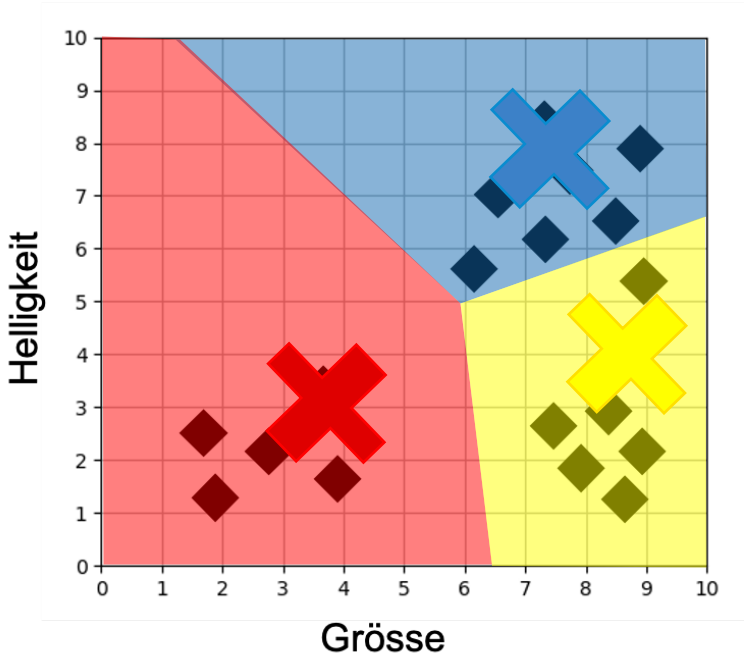


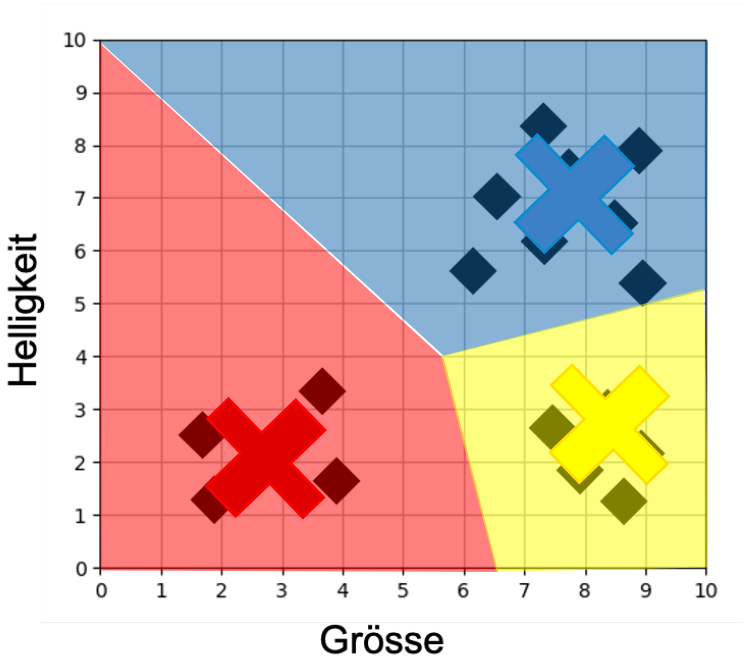


Der Trainingsalgorithmus

- **Zufällig Initialisierung** der Zentroide
- **Aktualisierung der Zuweisung**
- **Aktualisierung der Zentroide**. Jedes Zentroid wird auf den Durchschnitt aller Punkte gesetzt, die diesem Zentroid-Cluster zugewiesen sind.







Der Trainingsalgorithmus

Der Trainingsalgorithmus

- **Gegeben:** $\{x_1, x_2, \dots, x_n\} \subseteq \mathbb{R}^d$
- **Initialisieren** Sie Zentroide zufällig.
  - Für jedes $j \leq K$, definieren Sie θ_j durch das zufällig Auswählen eines Wertes aus $\mathbb{R}^d$.
- **Aktualisieren Sie die Zuweisung**
  - Für $x \in \mathbb{R}^d$, definieren Sie $c(x) = argmin_j \left\|x - \theta_j\right\|$.
- **Setzen Sie** Zentroide auf Durchschnitte.
  - Für $j \leq K$, definieren Sie $\theta_j = \frac{1}{|S_j|}\sum_{x \in S_j} x$, wobei S_j die Menge aller Punkte ist, die dem Cluster j durch c zugewiesen sind.
- Gehen Sie zurück zur Aktualisierung der Zuweisung und wiederholen Sie das, bis die Zentroide nicht mehr ändern.

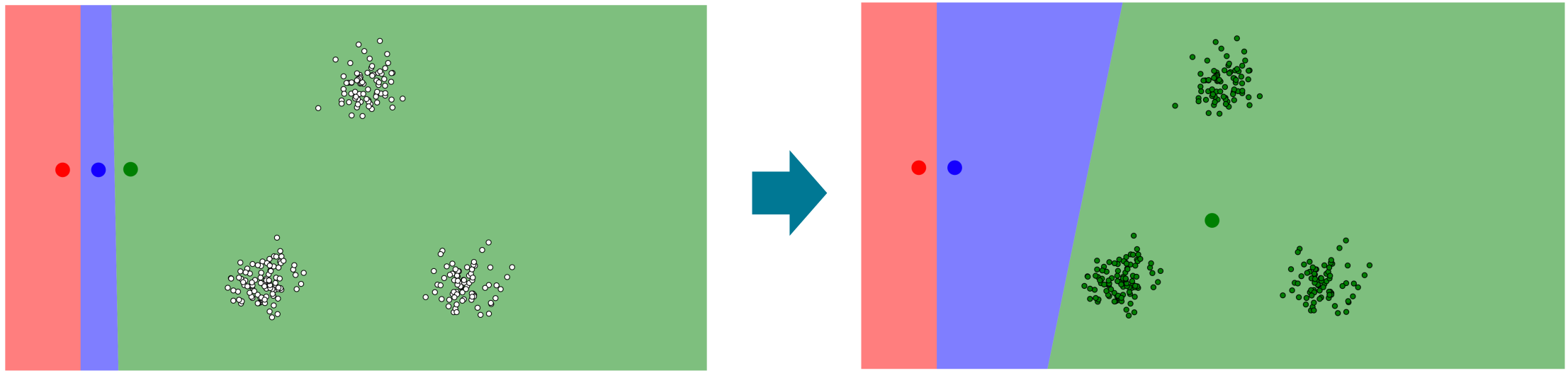
- Es kann formal bewiesen werden, dass K-Means immer konvergiert!

Visualisierung

- Naftali Harris's Webseite: https://www.naftaliharris.com/blog/visualizing-k-means-clustering/

Bemerkungen

- K-Means bestimmt nicht die Anzahl K von Clustern. Dies ist ein Hyperparameter.
- K-means ist empfindlich gegenüber der Initialisierung.



- Wie üblich müssen Sie sehr vorsichtig sein, um nicht zu überanpassen. Wie können Sie überanpassen?

Overfitting in K-means (nicht klausurrelevant)

- Wenn Sie K zu gross wählen (z.B. K = Anzahl der Punkte in den Daten), dann ist die minimale Clusterzuordnung eine, bei der jedem Punkt sein eigener Cluster zugewiesen ist.
- **Warnung:** Cross-Validation, mit der Verlustfunktion, wird hier nicht funktionieren, um Overfitting zu erkennen. Warum?

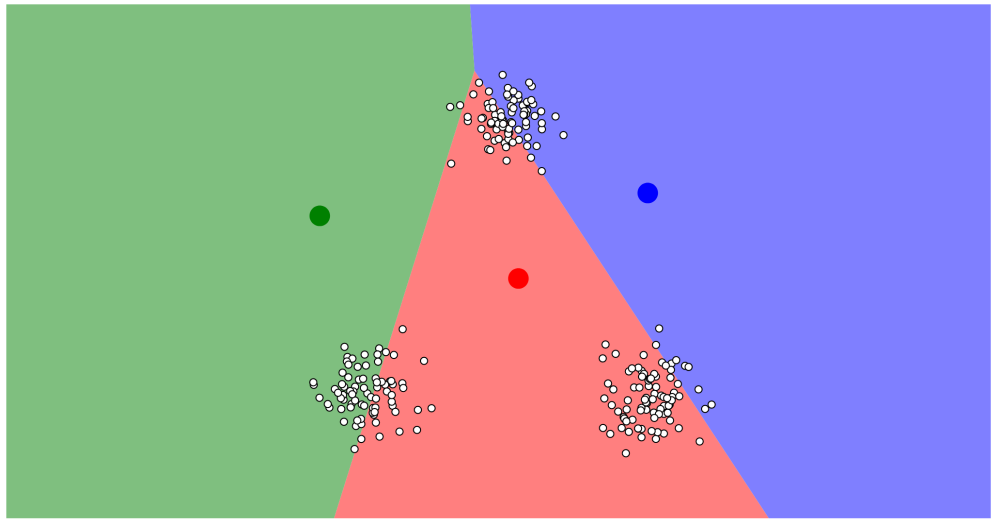
Overfitting in K-means (nicht klausurrelevant)

- Wenn Sie K zu gross wählen (z.B. K = Anzahl der Punkte in den Daten), dann ist die minimale Clusterzuordnung eine, bei der jedem Punkt sein eigener Cluster zugewiesen ist.
- **Warnung:** Cross-Validation, mit der Verlustfunktion, wird hier nicht funktionieren, um Overfitting zu erkennen. Warum?
- **Wenn jedem Punkt sein eigener Cluster zugewiesen ist, beträgt die Verlustfunktion 0!**
- Die Lösung besteht dann darin, eine Verlustfunktion zu verwenden, die auch die Anzahl der verwendeten Cluster bestraft. Zum Beispiel definieren wir eine ausreichend grosse Konstante $\lambda > 0$ und verwenden dann

$$\mathcal{L}(X, \theta_1, \ldots, \theta_K, c) + \lambda\, e^K$$

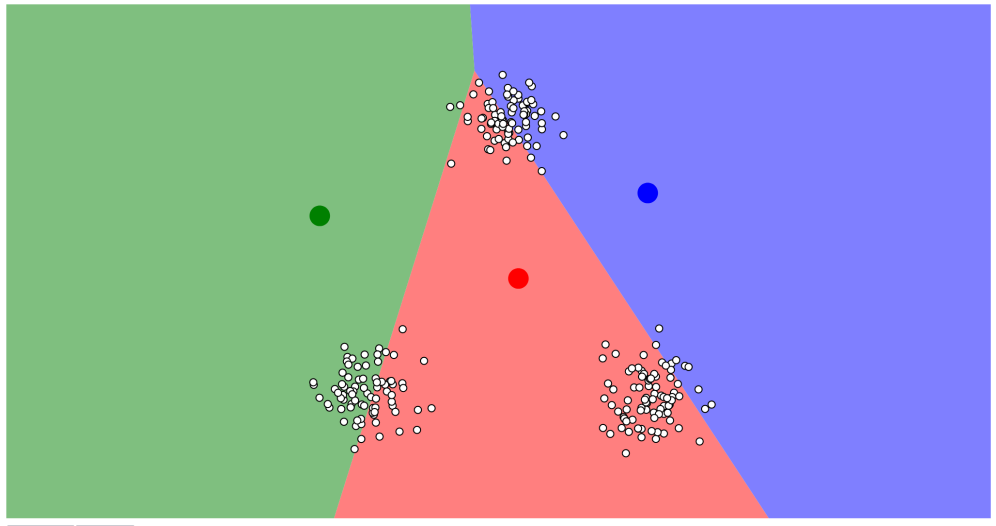
Quiz

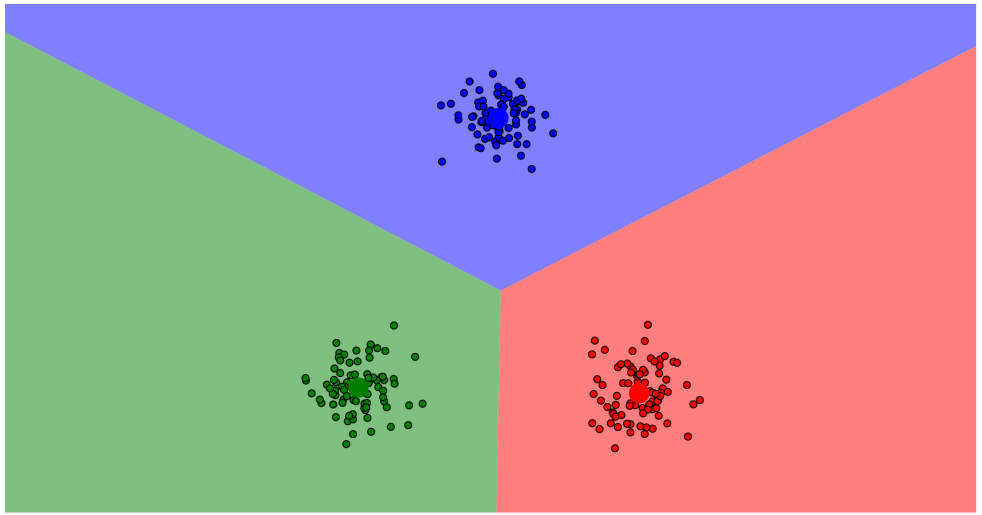
- Welche ist die Ausgabe von K-Means bei der unten abgebildeten Initialisierung?



Quiz

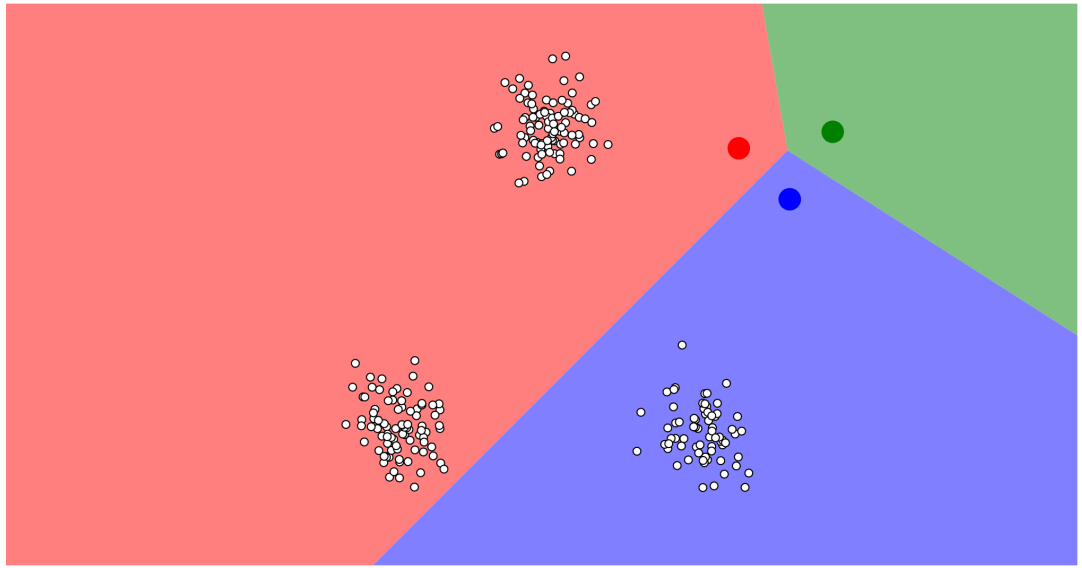
- Welche ist die Ausgabe von K-Means bei der unten abgebildeten Initialisierung?





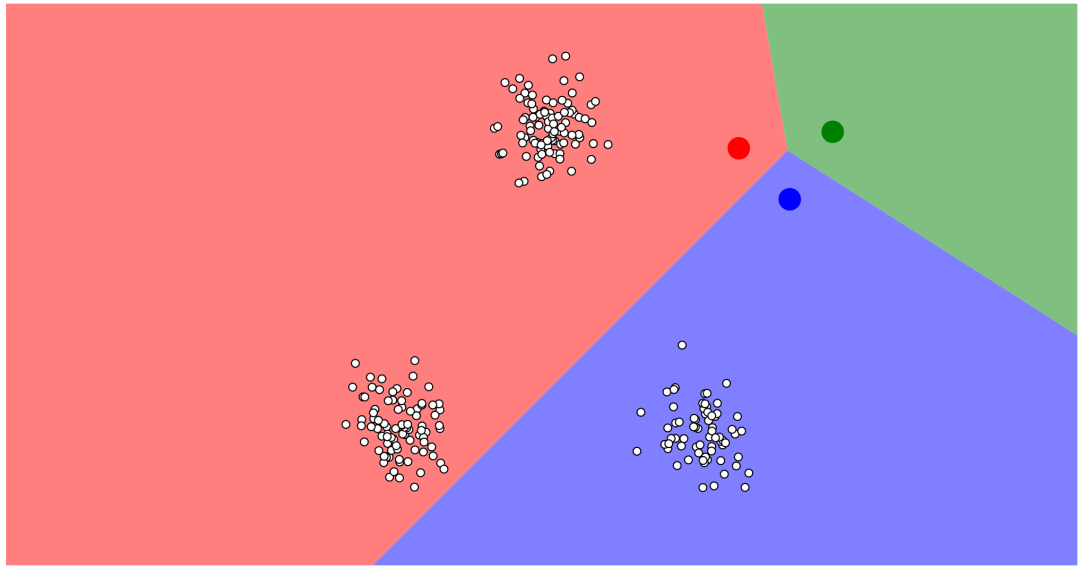
Quiz

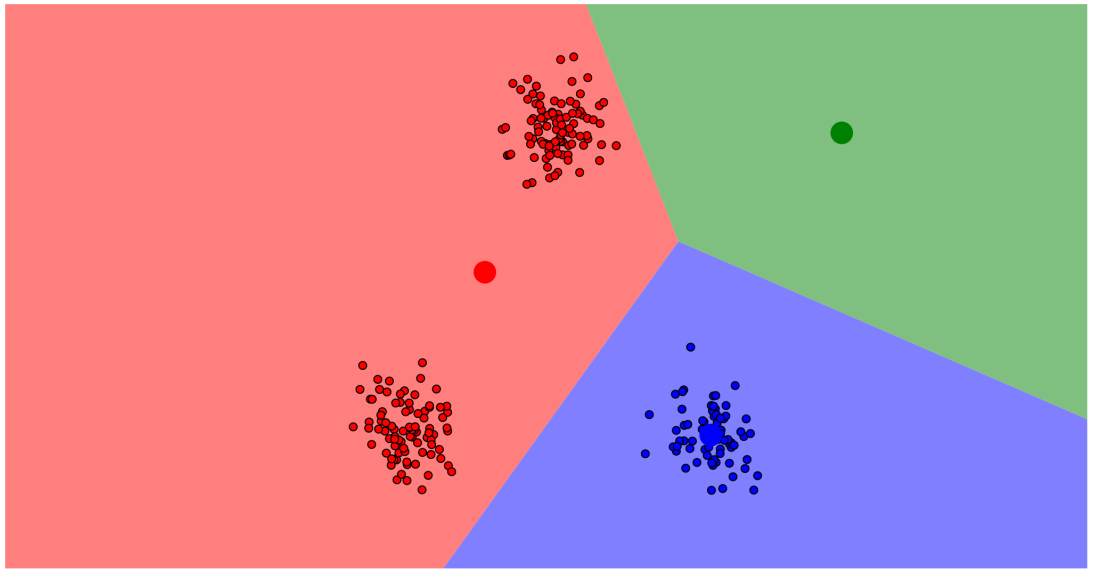
- Welche ist die Ausgabe von K-Means bei der unten abgebildeten Initialisierung?



Quiz

- Welche ist die Ausgabe von K-Means bei der unten abgebildeten Initialisierung?





ETH zürich

Dimensionsreduktion

Der Fluch der Dimensionalität

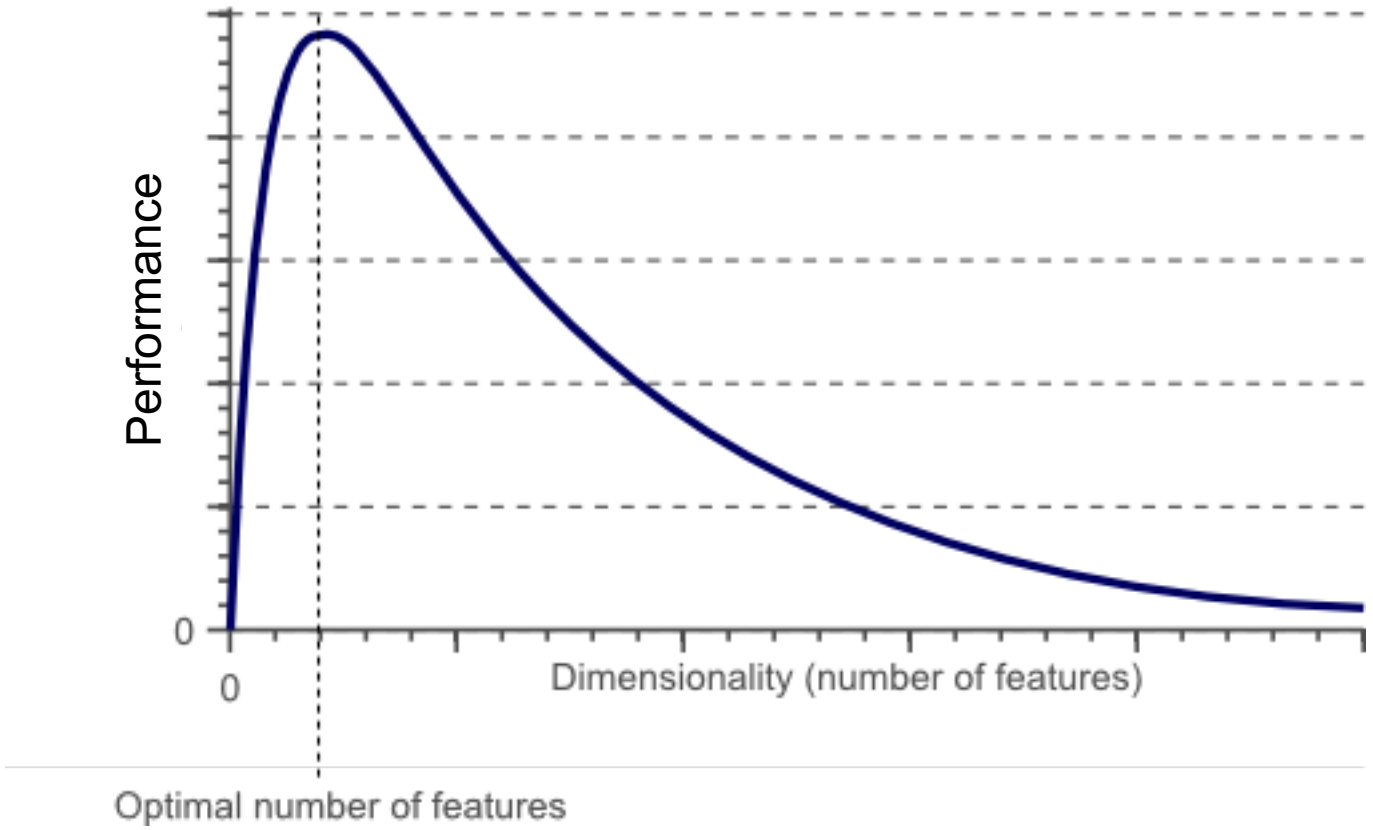
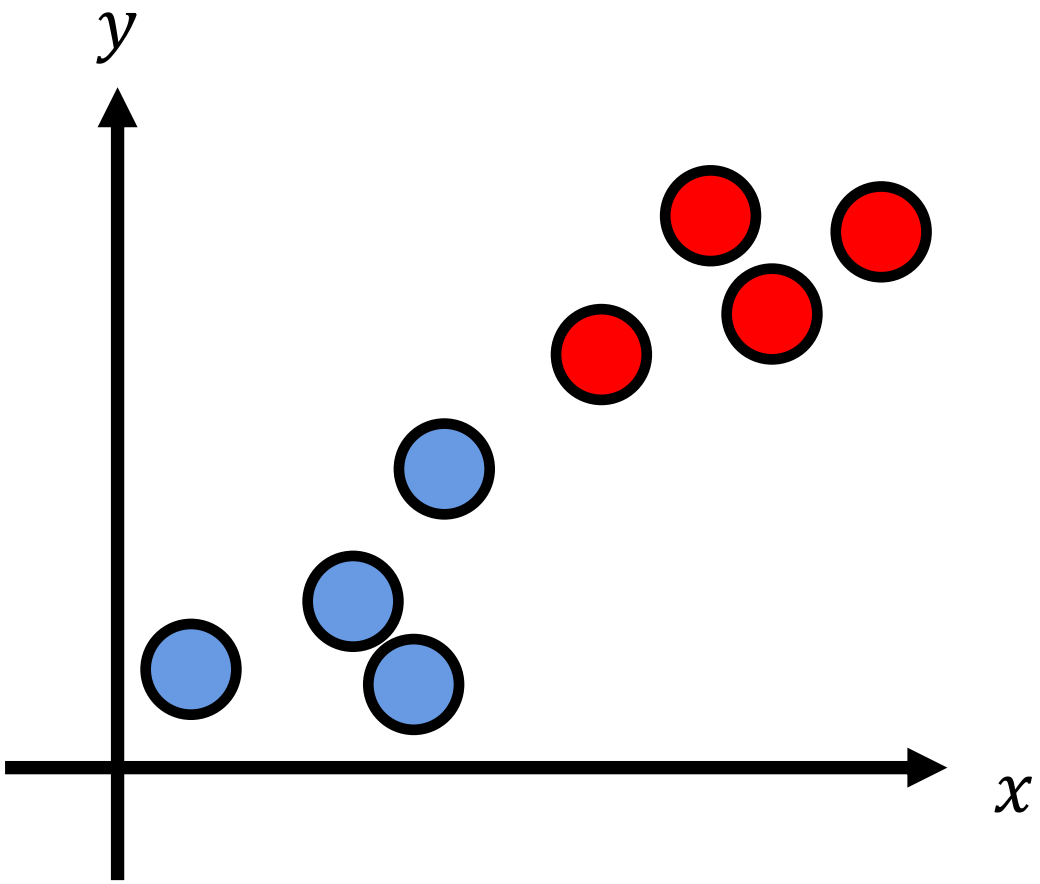


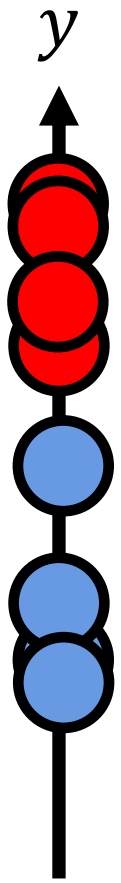
Bild von: https://goldinlocks.github.io/Basic-Dimensionality-Reduction/

Der Fluch der Dimensionalität

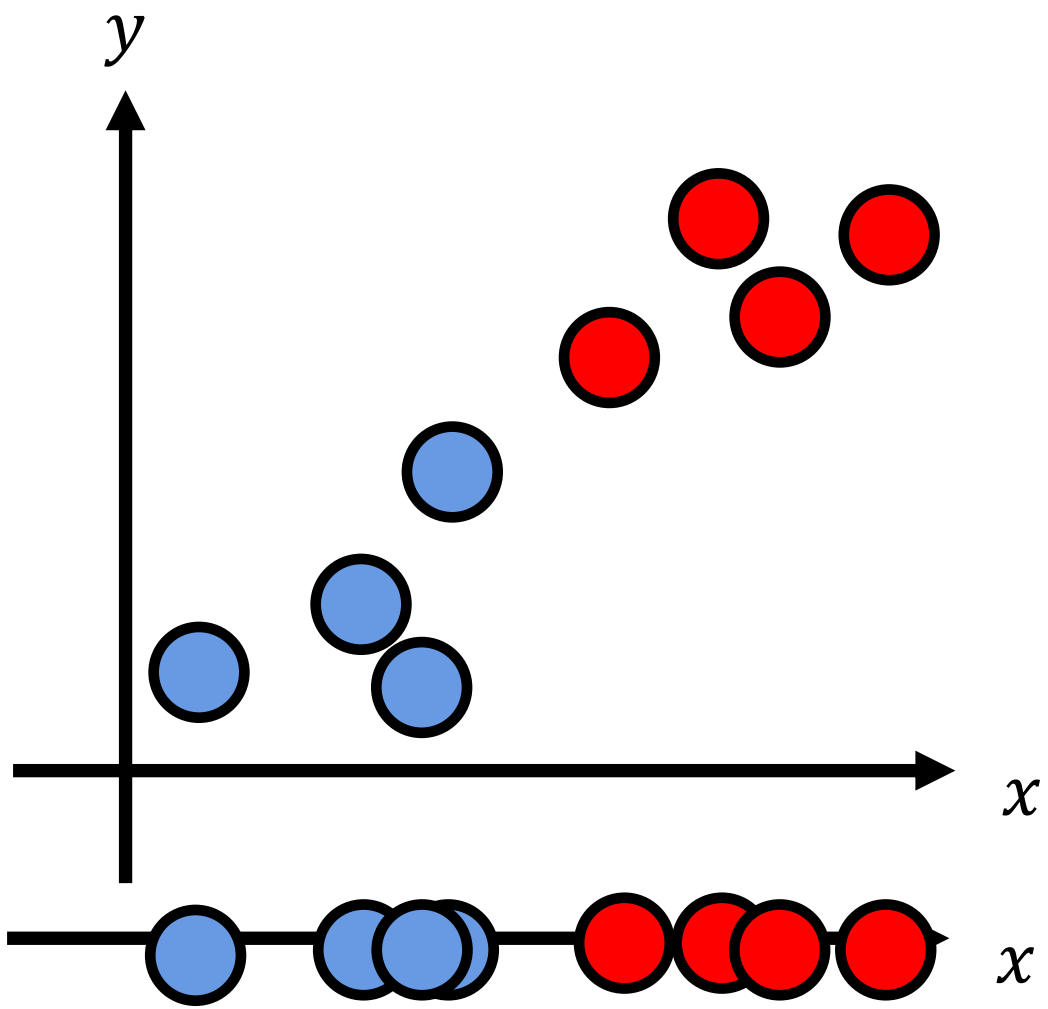
- Mehr Merkmale erhöhen den Suchraum für Trainingsalgorithmen dramatisch.
  - Trainingsalgorithmen werden langsamer.
  - Trainierte Modelle werden schlechter.
- Logistische Regression mit allen Merkmalen im Brustkrebsdatensatz: 0.93
- Logistische Regression mit einer sorgfältig konstruierten Menge von 6 Merkmalen: 0.96

Principal component analysis (PCA)

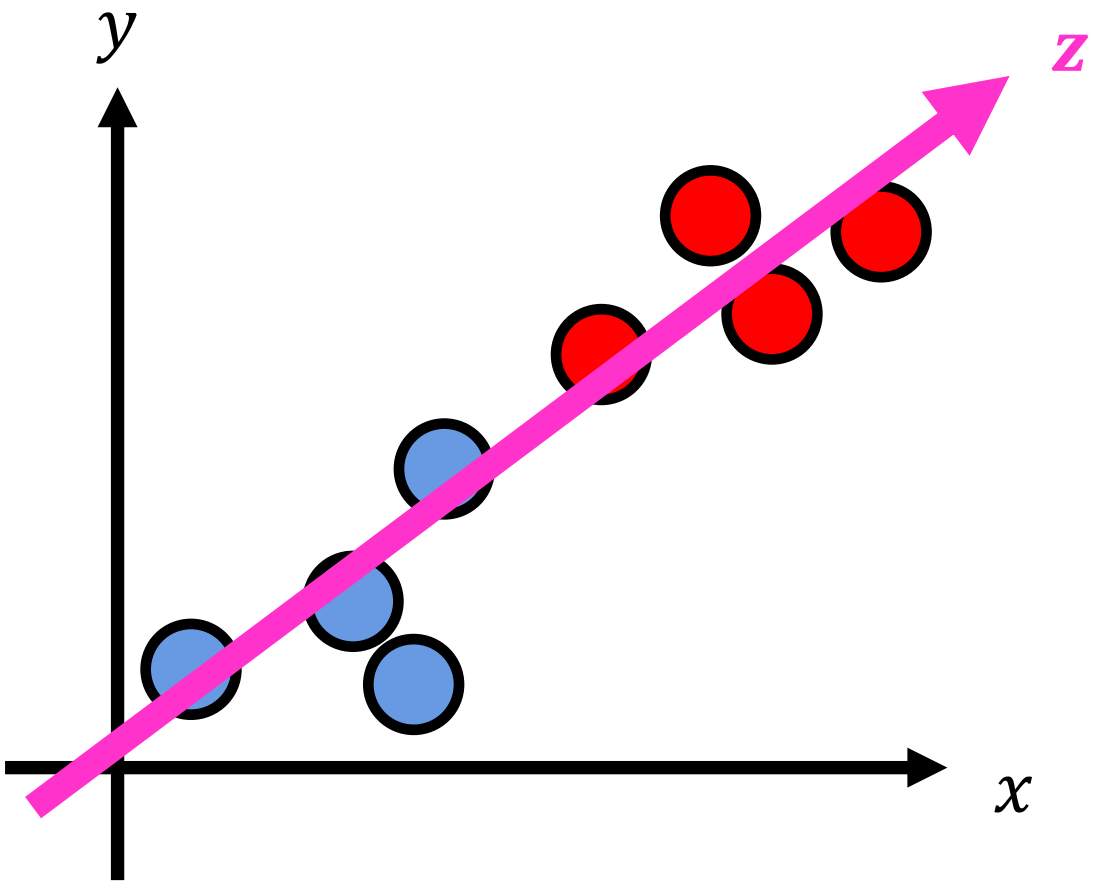




Principal component analysis (PCA)

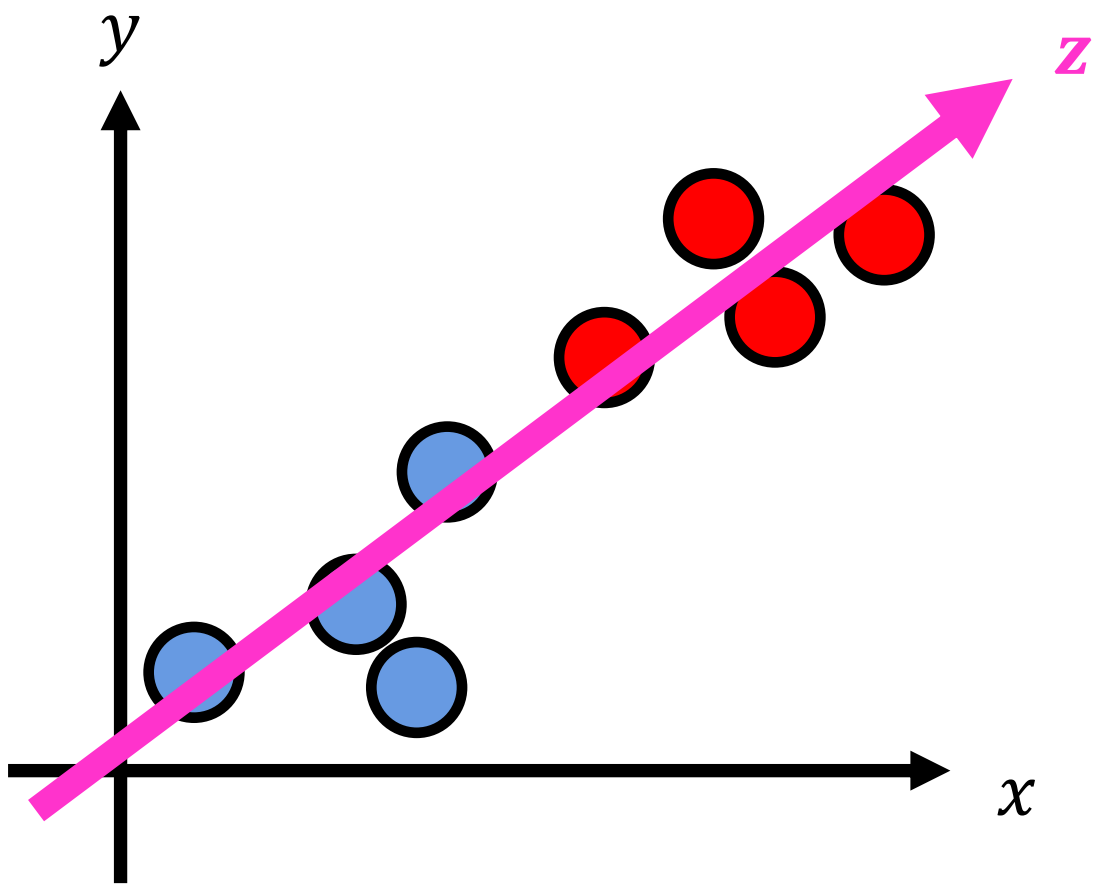


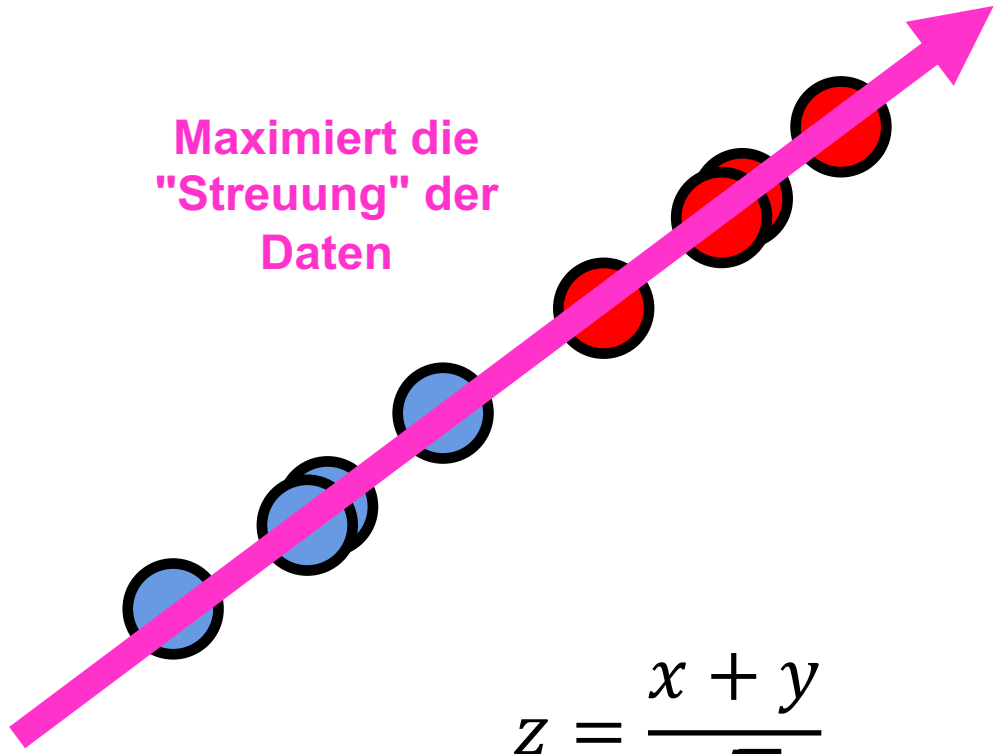
Principal component analysis (PCA)



Maximiert die "Streuung" der Daten

Principal component analysis (PCA)

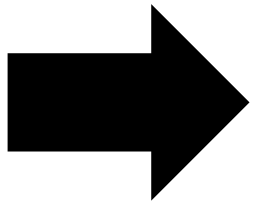




$$z = \frac{x + y}{\sqrt{2}}$$

Ein weiteres motivierendes Beispiel

A	B	C	D
0.1	0.12	-5.6	-5.64
0.2	0.21	-6.1	-6.13
0.3	0.34	-7.3	-7.31
0.4	0.40	-8.1	-8.15
0.5	0.52	-9.2	-9.22
0.6	0.61	-10.1	-10.13
0.7	0.73	-10.3	-10.31



B	D
0.12	-5.64
0.21	-6.13
0.34	-7.31
0.40	-8.15
0.52	-9.22
0.61	-10.13
0.73	-10.31

PCA: formalization

- Gegeben $\{x_1, x_2, \dots, x_n\} \subseteq \mathbb{R}^D$ und $d < D$, berechnen Sie einen Unterraum $\mathcal{H} \subseteq \mathbb{R}^D$ der Dimensionalität d, so dass $\{proj_{\mathcal{H}} x_1, proj_{\mathcal{H}} x_2, \dots, proj_{\mathcal{H}} x_n\} \subseteq \mathcal{H}$ eine grosse "Streuung" hat.

- Mit einer geeigneten Wahl von d gibt es eine Reduktion der Dimensionen ohne signifikanten Qualitätsverlust der Daten.
- Hauptvorteile:
  - Trainingsalgorithmen laufen schneller und produzieren bessere Modelle.
  - Daten können visualisiert werden.

Basisfall d = 1

Basisfall d = 1

- Der Maximierer dieser quadratischen Funktion ist der grösste Eigenwert λ_1^* von S!
- Die optimale Projektion ist die auf einen Einheitseigenvektor u_1^* von λ_1^*.

Allgemeiner Fall d > 1

Allgemeiner Fall d > 1

- Berechnen Sie u_1^* aus X wie zuvor.
- Sei $X_1 = \{x - proj_{u_1^*} x : x \in X\}$.
- Berechnen Sie u_2^* aus X_1 wie zuvor.
- Sei $X_2 = \{x - proj_{u_2^*} x : x \in X_1\}$.
- …
- Berechnen Sie u_d^* aus X_{d-1} wie zuvor.

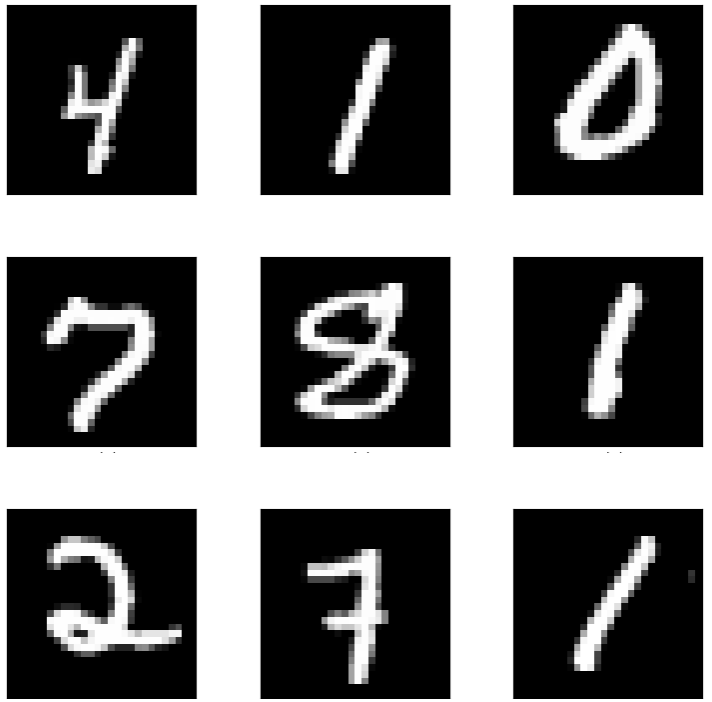
- Definieren Sie $\pi\,(x) = (x^\top u_1^*, \dots)$

ETH zürich

Unüberwachtes Erkennen von Ziffern

PCA + K-Means für MNIST

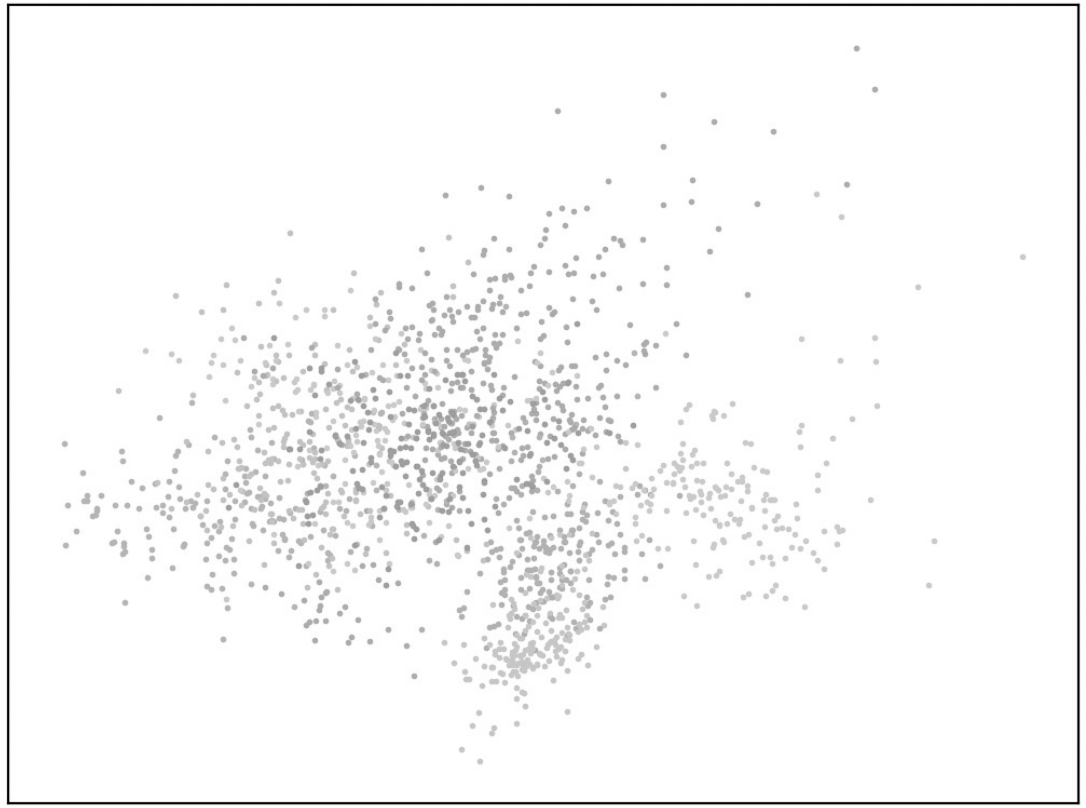
- Jedes Bild ist 28 * 28 Pixel gross.
- Jedes Bild kann als Vektor in $\mathbb{R}^{784}$ dargestellt werden.



Strategie

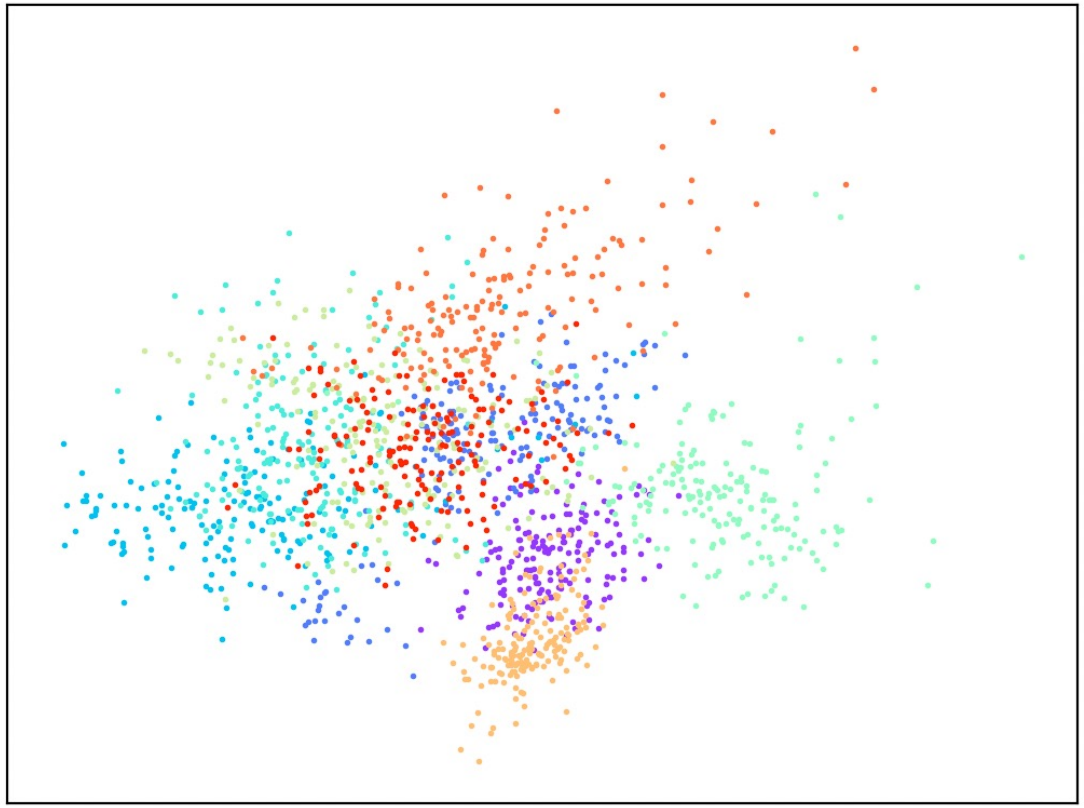
- Wir wenden PCA auf die Bilder an, um sie in Punkte in $\mathbb{R}^2$ zu transformieren.
- Wir wenden dann K-Means auf diese Punkte an, wobei wir 10 Cluster verwenden.

PCA für MNIST

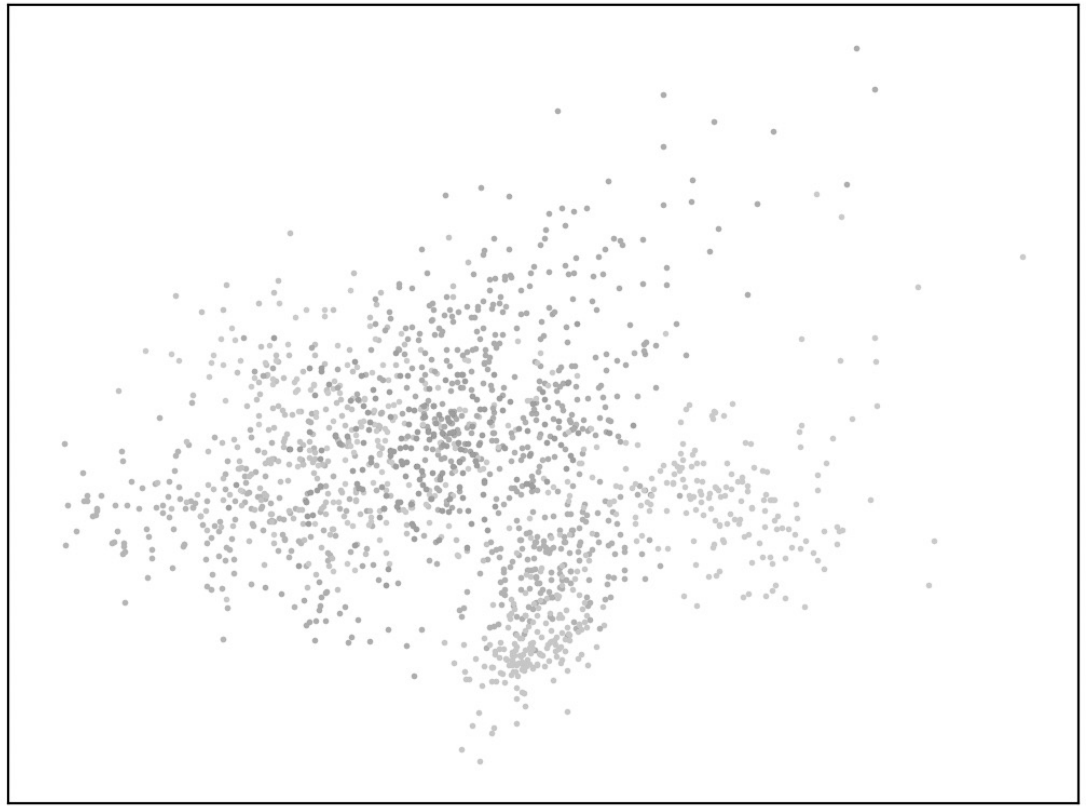


PCA on MNIST

PCA für MNIST

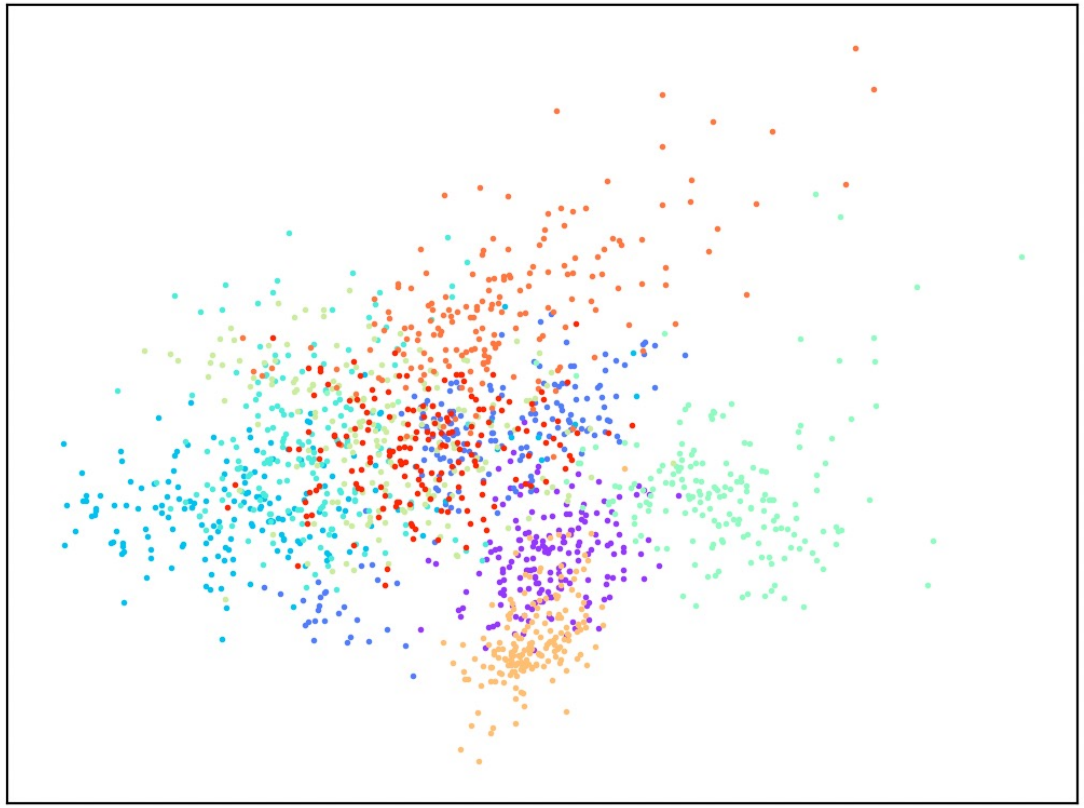


PCA on MNIST. Each point is colored according to the corresponding digit.

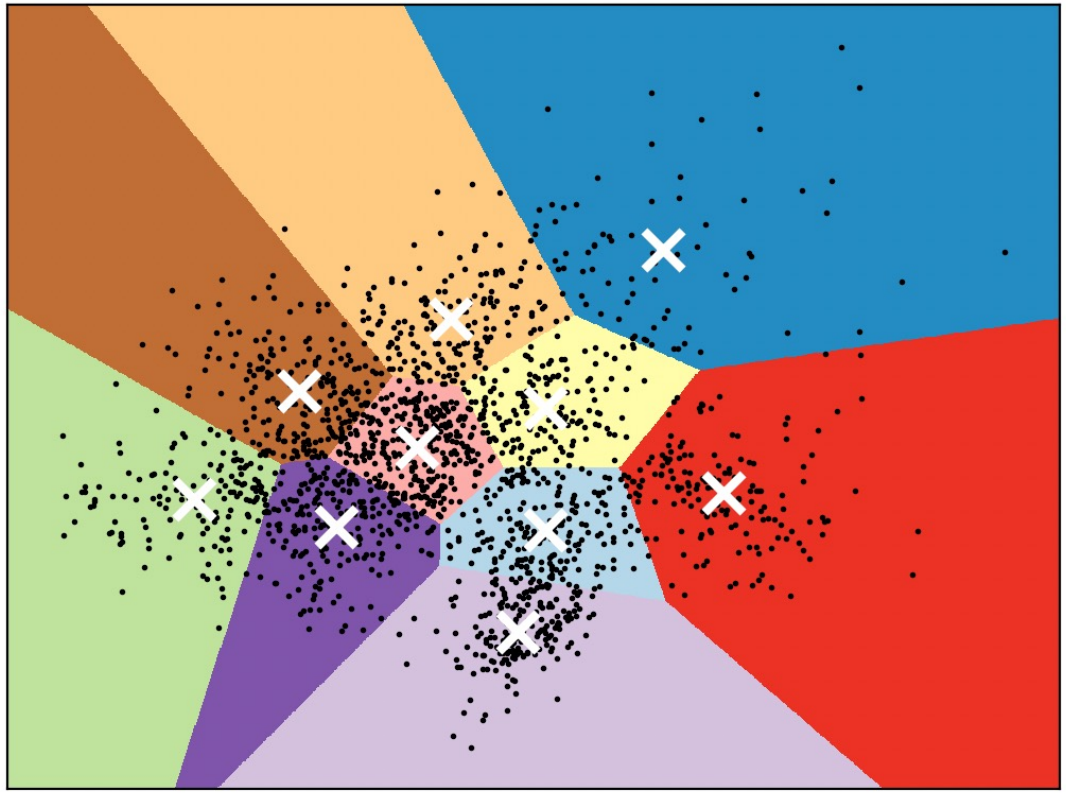


PCA on MNIST

K-means für MNIST



PCA on MNIST. Each point is colored according to the corresponding digit.



K-means on PCA on MNIST

Ergebnisse

- K-Means hat 10 Cluster entdeckt. Jeder Cluster erfasst mehr oder weniger alle Darstellungen, die einer einzelnen Ziffer entsprechen!
- Mit PCA und K-Means konnten wir einen Ziffernklassifikator ohne jegliche Überwachung erstellen!

Was haben wir heute gelernt?

- Wie geht man mit nicht-numerischen Daten um?
  - Ordinal encoding
  - Mean encoding
  - One-hot encoding
- Unüberwachtes Lernen
  - Clustering mit K-Means
  - Dimensionsreduktion mit PCA