FFHS
Fernfachhochschule Schweiz
Zürich | Basel | Bern | Brig
Mitglied der SUPSI

Einführung in Data Science
Unüberwachtes Lernen – Block 4

PVA 4 – Programm

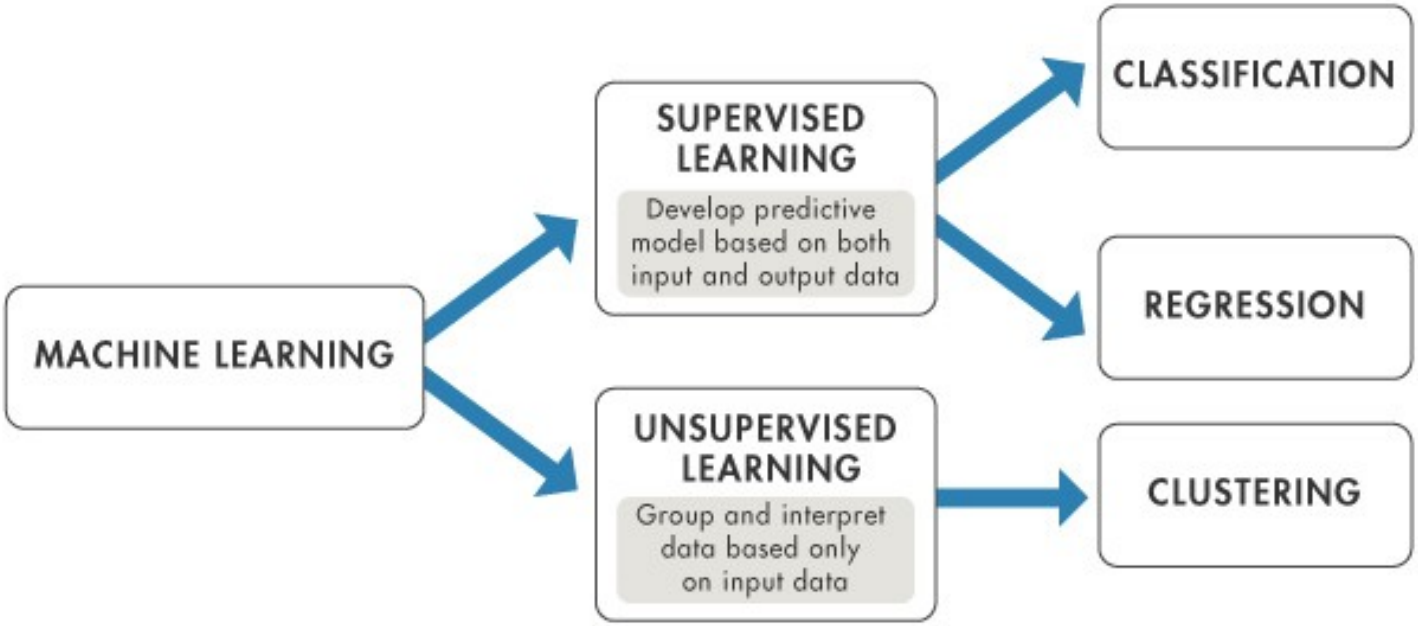
Thema	Form	Zeit
Besprechen der Semesterarbeit	Diskussion	13:45 – 14:00
Verständnisfragen	Diskussion	14:00 – 14:15
Unsupervised Learning	Vorlesung	14:15 – 15:00
Pause		15:00 – 15:15
Zusammenfassung B1-B3	Vorlesung	15:15 – 15:45
Workshop	Workshop	15:45 – 16:45
Diskussion Workshop	Diskussion	16:45 – 17:00

Machine Learning

Machine Learning Algorithmen

- Supervised Learning (überwachtes Lernen)
- Unsupervised Learning (unüberwachtes Lernen)

Supervised Learning vs. Unsupervised Learning



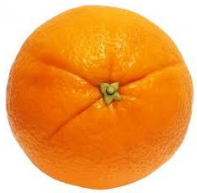
Supervised Learning

- Daten mit k Attributen und eine Klasse

Features:
1. Color: **Radish/Red**
2. Type : **Fruit**
3. Shape
etc...

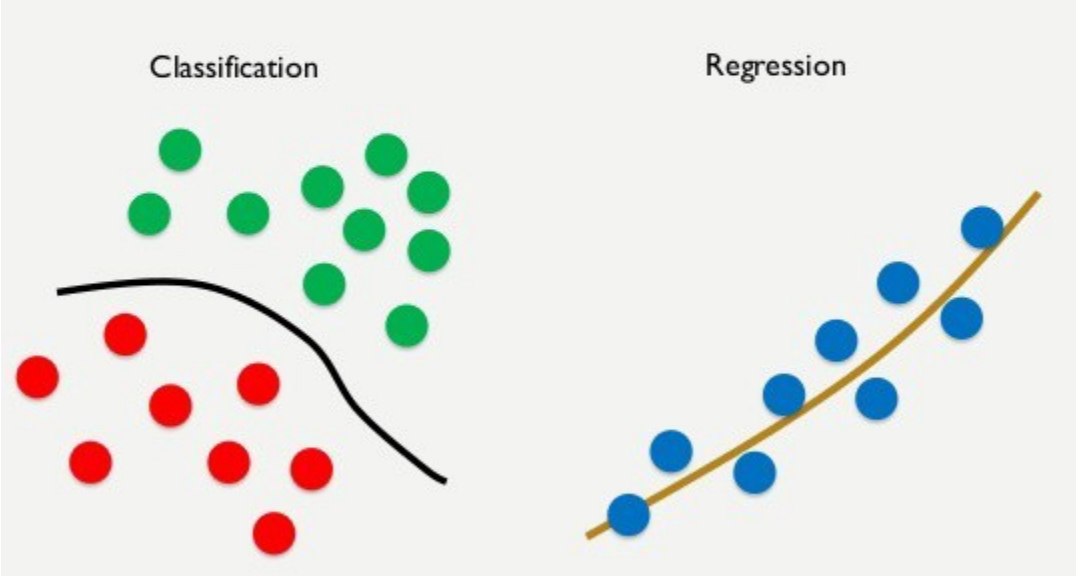
Features:
1. Sky Blue
2. **Logo**
3. Shape
etc...

Features:
1. **Yellow**
2. **Fruit**
3. Shape
etc...



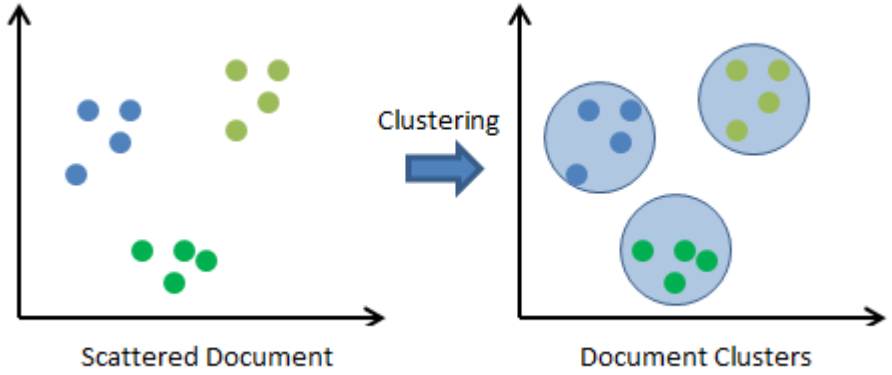
Klassifikation vs Regression

- Klassifikation: Die Ausgangsvariable nimmt Klassenbezeichnungen. (Beispiel: ist eine Email Spam oder nicht?)
- Regression: Die Ausgangsvariable nimmt kontinuierliche Werte an. (Beispiel: geeignete Hauspreise kennen). Trainingsdaten sind nicht mit Klassen annotiert.



Unsupervised Learning

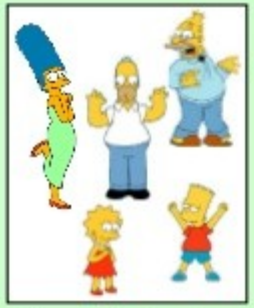
- Die bekannten Daten haben keine Labels (im Unterschied zum Supervised Learning)
- Wir wissen auch nicht, was die Datenpunkten bedeuten
- Die Aufgabe besteht darin, in den Daten Muster (Clusters) zu finden



Unsupervised Learning

What is a natural grouping among these objects?

Clustering is subjective



Simpson's Family

School Employees

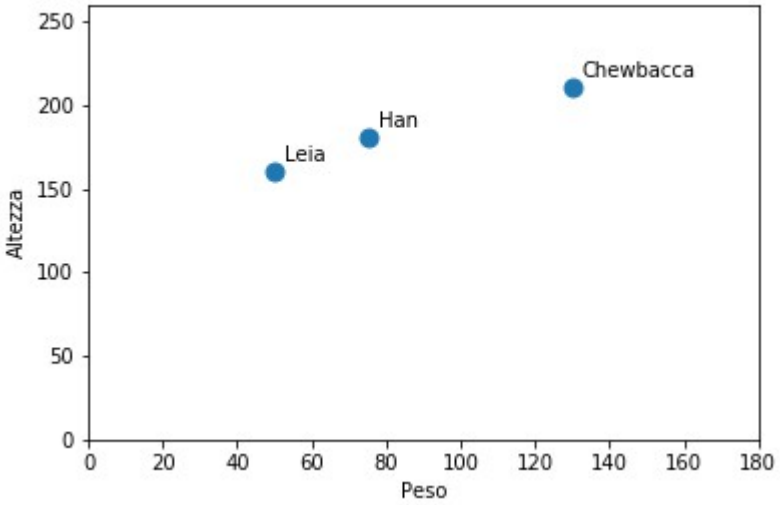
Females

Males

Beispiele

- News.google.com → Links mit vergleichbarem Inhalt werden geclustert
- Analyse des Genoms verschiedener Menschen: z.B. auf das Vorhandensein eines Gens
- Grosse Computer Cluster finden – Maschinen, die zusammenarbeiten, ins gleiche Rack tun
- Social Network Analyse
- Markt Segmentierung
- Astronomische Daten analysieren

Metriken



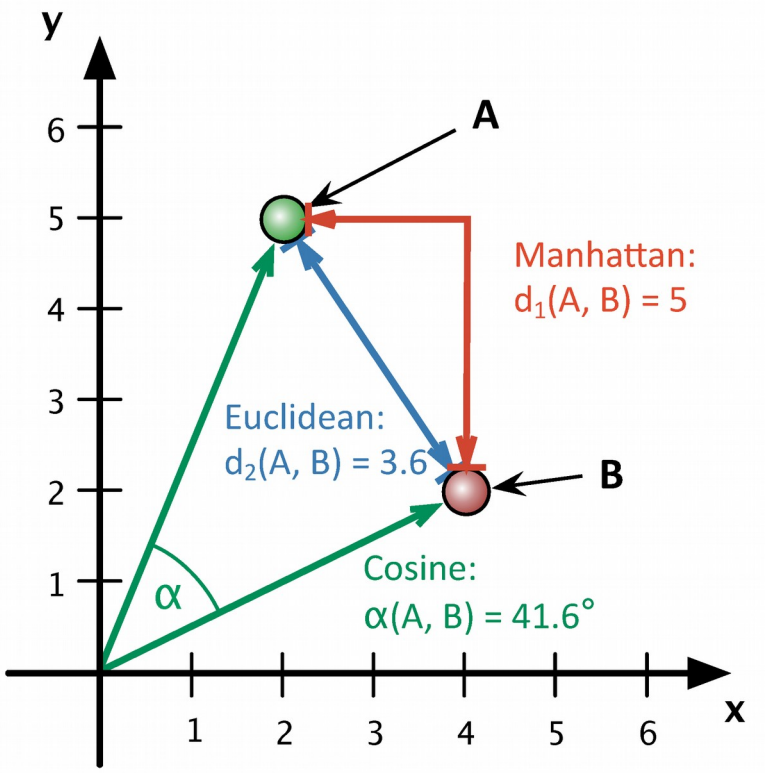
Han: (180, 75)

Leia: (160, 50)

Chewbacca: (210, 130)

- Je grösser die Anzahl der Attribute, desto grösser wird der betreffende Raum sein.
- Sobald die Objekte in der Sammlung mit Punkten im Raum verbunden sind, ist es möglich, ihre "Ähnlichkeit" durch das Konzept der Distanz zu überprüfen.
- Je näher zwei Punkte sind, desto ähnlicher sind sie.

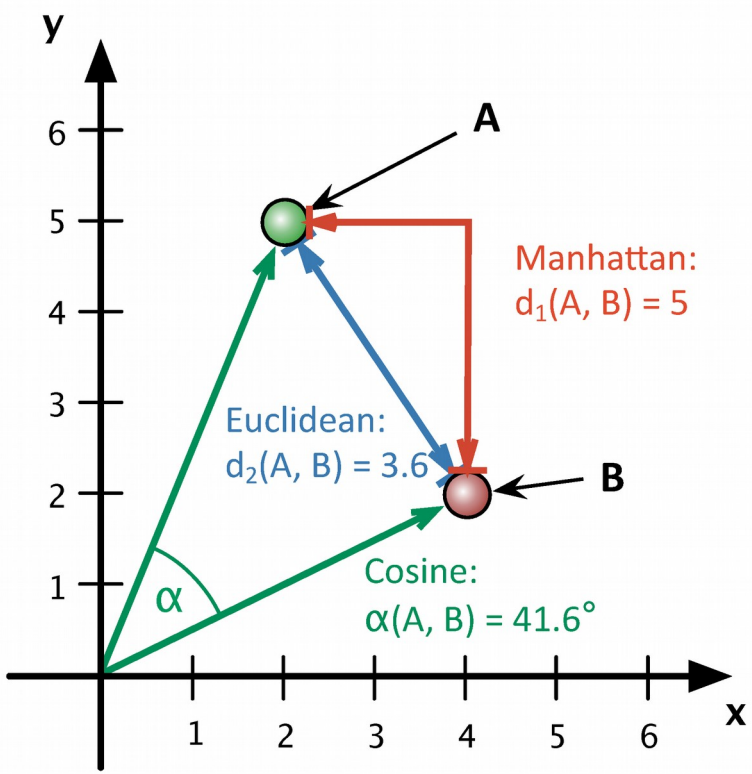
Euklidische Distanz



Euklidische Distanz: geometrische Distanz zwischen zwei Punkte

$$d(x, y) = \|x - y\|_2 = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

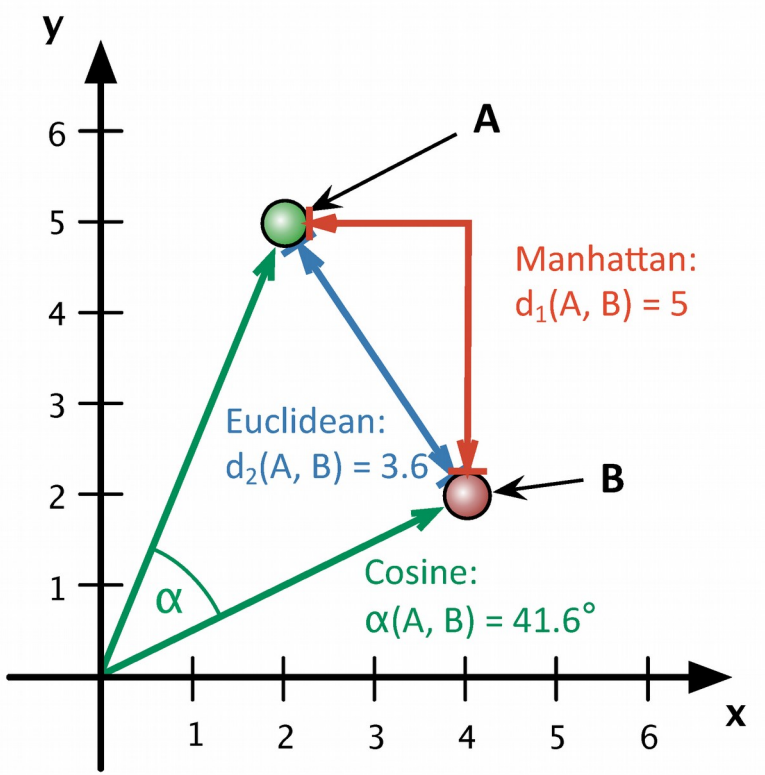
Manhattan Distanz



Manhattan Distanz: City-Block-Distanz. Distanz zwischen zwei Punkten als die Summe der absoluten Differenzen ihrer Einzelkoordinaten

$$d(a,b) = \sum_i |a_i - b_i|$$

Cosinus Distanz



$$\frac{\sum_{k=1}^{n} A(k)B(k)}{\sqrt{\sum_{k=1}^{n} A(k)^2}\sqrt{\sum_{k=1}^{n} B(k)^2}}$$

Edit Distanz

- Die Levenshtein-Distanz zwischen zwei Strings ist die minimale Anzahl von Einfüge-, Lösch- und Ersetz-Operationen, um die erste String in die zweite umzuwandeln.
- In der Praxis wird die Levenshtein-Distanz zur Bestimmung der Ähnlichkeit von Strings beispielsweise zur Rechtschreibprüfung oder bei der Duplikaterkennung angewandt.

Cost of change

D(i,j)= min

D(i-1,j-1) + d(si,tj) *//subst/copy*
D(i-1,j)+1 *//insert*
D(i,j-1)+1 *//delete*

		C	O	H	E	N
	0	1	2	3	4	5
M	1	1	2	3	4	5
C	2	1	2	3	4	5
C	3	2	2	3	4	5
O	4	3	2	3	4	5
H	5	4	3	2	3	4
N	6	5	4	3	3	3

T1	T2	Cost
M	-	1
C	-	1
C	C	0
O	O	0
H	H	0
-	E	1
N	N	0

T1	T2	Cost
M	-	1
C	C	0
C	-	1
O	O	0
H	H	0
-	E	1
N	N	0

D = 3

kitten → sitten (substitute “s” for “k”)

sitten → sittin (substitute “i” for “e”)

sittin → sitting (insert “g” at the end)

N-gram Similarity

- N-Gramme sind das Ergebnis der Zerlegung eines Textes in Fragmente. Der Text wird dabei zerlegt, und jeweils N aufeinanderfolgende Fragmente werden als N-Gramm zusammengefasst. Die Fragmente können Buchstaben, Wörter und Ähnliches sein.
- Die Firma Google veröffentlichte im Jahr 2006 sechs DVDs mit englischsprachigen N-Grammen von ein bis fünf Wörtern, die bei der Indexierung des Webs entstanden.

N-Gramm-Name	N	Beispiel
Monogramm	1	A
Bigramm	2	AB
Trigramm	3	UNO
Tetragramm	4	HAUS
Pentagramm	5	HEUTE
Hexagramm	6	SCHIRM
Heptagramm	7	TELEFON
Oktogramm	8	COMPUTER
...	...	...
Multigramm	N	BEOBACHTUNGSLISTE ($N = 17$)

Distance of strings s1 and s2:

$$|G(s1)| + |G(s2)| - 2|G(s1) \cap G(s2)|$$

Example:

G(rodney) = {rod, odn, dne, ney}

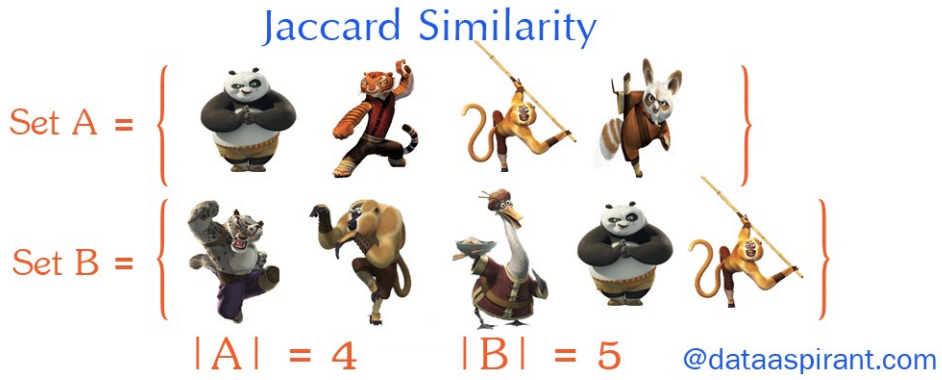
G(rhodnee) = {rho, hod, odn, dne, nee}

distance (rodney, rhodnee) = 4 + 5 – 2*2 = 5

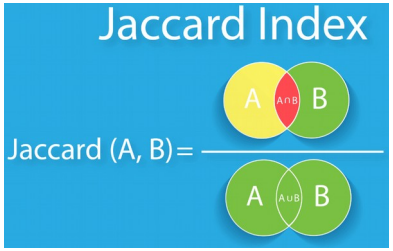
Jaccard Index und Distanz

Kennzahl für die Ähnlichkeit von Mengen.

Jaccard Index = (the number in both sets) / (the number in either set) * 100



2/7=0.29



Jaccard Distanz - Dissimilarity

D(X,Y) = 1 – J(X,Y) D=1-0.29=0.71

Clustering Algorithmen

- Greedy Algorithmen
- Hierarchical Clustering
- K-Means Clustering

Greedy Algorithmen

- Greedy Algorithmen treffen immer die Wahl, die in diesem Moment die beste zu sein scheint. Das bedeutet, dass sie eine lokal optimale Wahl trifft, in der Hoffnung, dass diese Wahl zu einer global optimalen Lösung führt.
- Der Greedy-Algorithmus hat nur einen Shot, um die optimale Lösung zu berechnen, so dass er nie zurückkehrt und die Entscheidung ändert.
- Beispiele: Prim, Kruskal Algorithm (Spanning Trees); Clustering Job Titles (VB).
- Manchmal gelingt es nicht, die optimale Lösung zu finden, weil der gesamte Datensatz nicht berücksichtigt wird.

Greedy Algorithmus zum Problem der minimalen Münzwechsel

- Das Ziel ist es, die minimale Anzahl von Münzen (mit einem bestimmten Wert) zu finden, die sich zu einem bestimmten Geldbetrag addieren.

- Beispiel: 40 bezahlen mit {1,5,10,20,25} Münzen.

- Optimal wäre 2x20 (2 Münzen).

- Greedy Ansatz: Münzen sortieren → {25,20,10,5,1}

- 40-25=15
- 15-10=5
- 5-5=0

- Greedy Lösung {25,10,5} (3 Münzen)

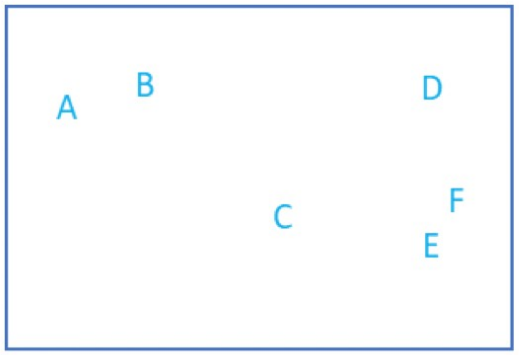
Hierarchical Clustering

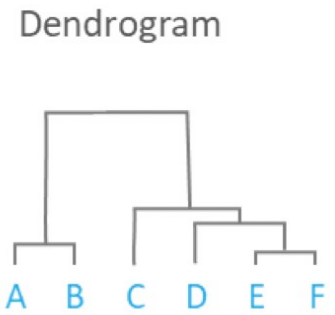
- Hierarchical Clustering ist ein Algorithmus, der ähnliche Objekte in Gruppen gruppiert.
- Der Endpunkt ist eine Reihe von Clustern, bei denen sich jeder Cluster von jedem anderen Cluster unterscheidet und die Objekte innerhalb jedes Clusters weitgehend ähnlich sind.
- Ziel: Inter-Cluster Ähnlichkeit maximieren, Intra-Cluster Ähnlichkeit minimieren.
- Berechnet die vollständige Matrix der Abstände zwischen allen Gegenständen.

Darstellung: Dendrogramm

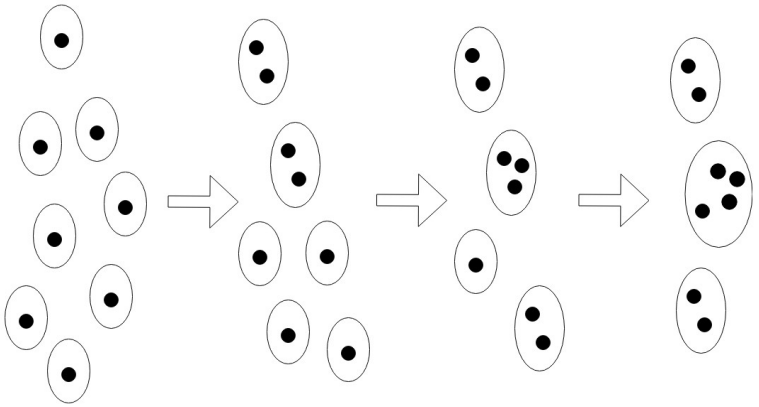
Ein Dendrogramm ist ein Diagramm, das einen Baum darstellt.

	A	B	C	D	E
B	16				
C	47	37			
D	72	57	40		
E	77	65	30	31	
F	79	66	35	23	10



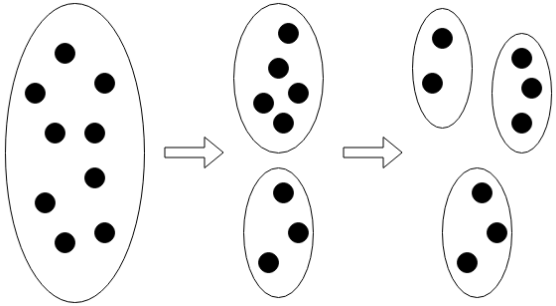


Agglomerative Clustering-Algorithmen (bottom-up)



Sie beginnen damit, jedes Objekt in der Menge in einen eigenen Cluster einzufügen und gruppieren es dann iterativ, bis eine bestimmte Bedingung erreicht ist (z.B. Anzahl der gewünschten Cluster).

Divisive Clustering-Algorithmen (top-down)



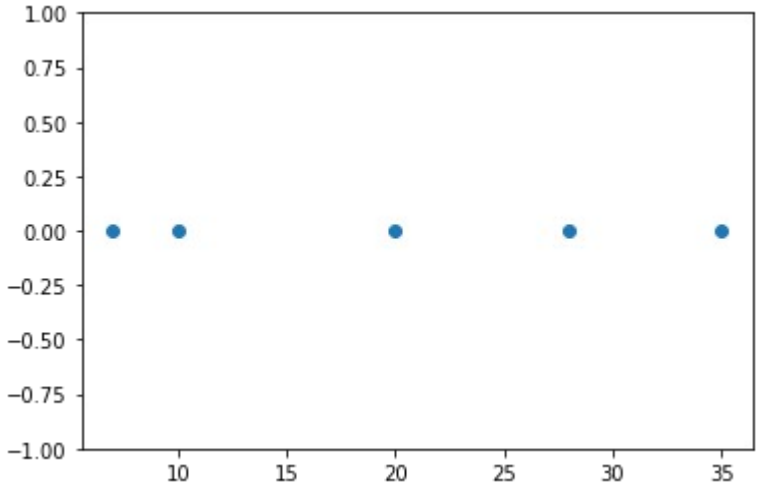
Sie beginnen damit, alle Objekte in der Menge in einem einzigen Cluster zu platzieren und trennen ihn dann iterativ in kleinere Cluster, bis eine bestimmte Bedingung erreicht ist.

Linkage

Nach der Auswahl einer guten Metrik ist es notwendig zu bestimmen, woher die Entfernung berechnet wird.

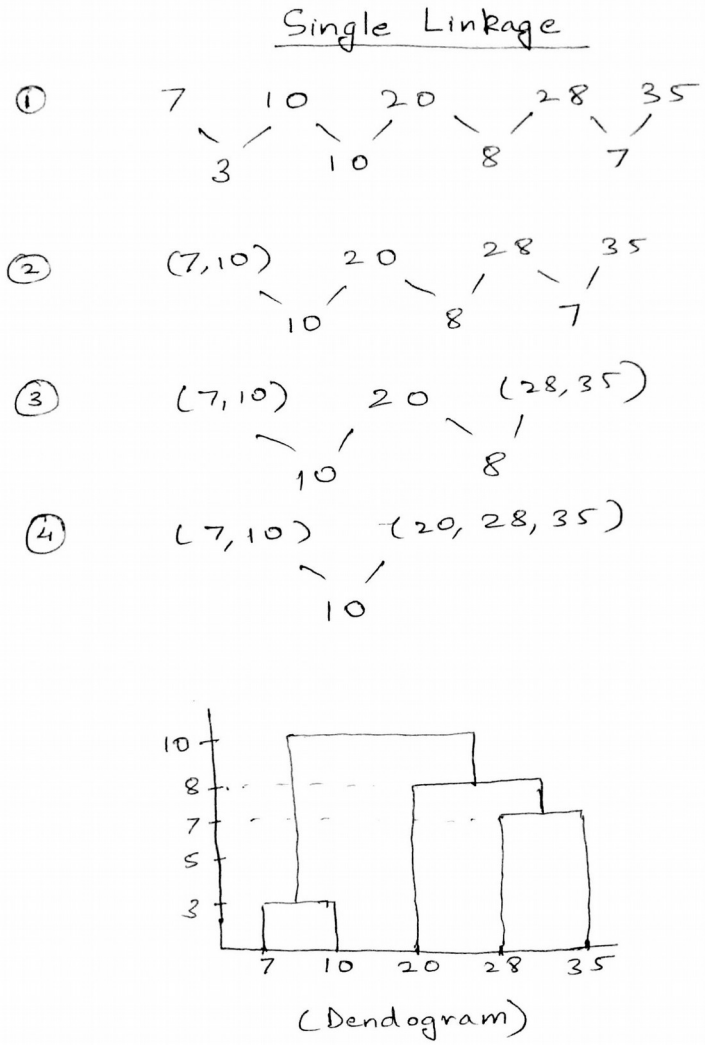
Single-Link: berechnet zwischen den beiden ähnlichsten Teilen eines Clusters.

Complete-Link: berechnet zwichen den beiden am wenigsten ähnlichen Teilen eines Clusters.



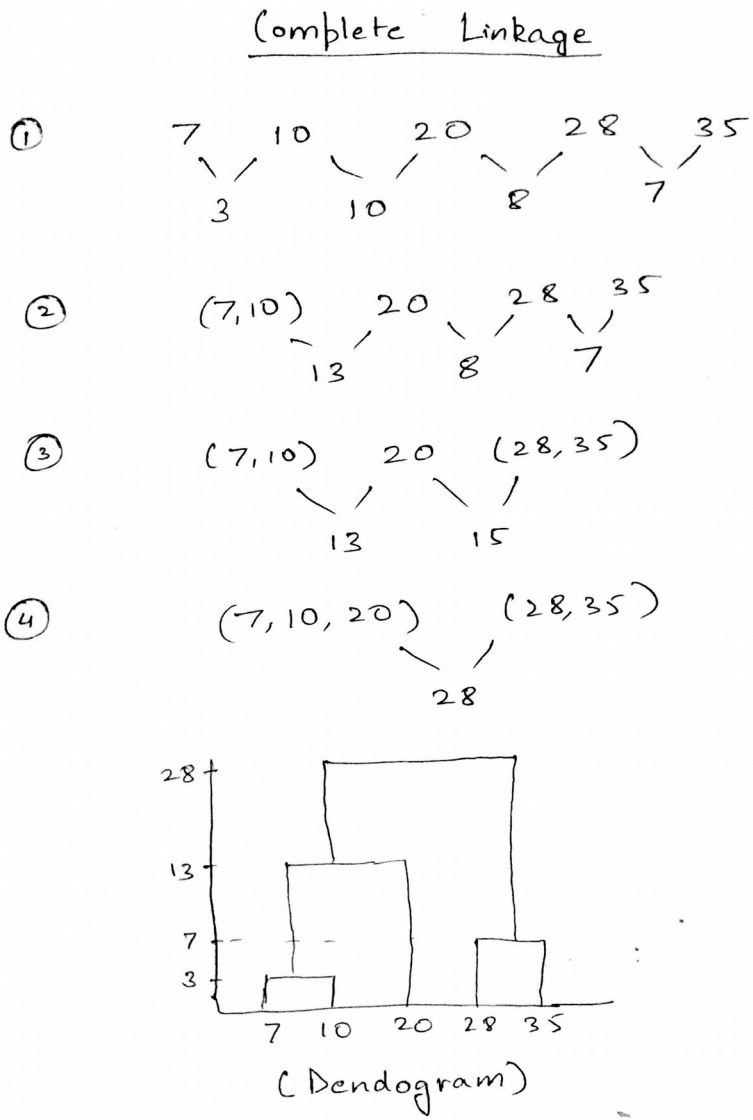
- 7 und 10 liegen nahe beieinander und sollten im gleichen Cluster liegen.
- 28 und 35 liegen nahe beieinander und sollten im gleichen Cluster liegen.
- Cluster des Mittelpunktes (20) ist nicht einfach zu schliessen.

Single Linkage



Im Single Link Clustering verschmelzen (merge) wir in jedem Schritt die beiden Cluster, deren zwei nächstgelegene Mitglieder den kleinsten Abstand haben.

Complete Linkage



Beim Complete-Link Clustering verschmelzen wir in jedem Schritt die Mitglieder der Cluster, die den kleinsten maximalen paarweisen Abstand bieten.

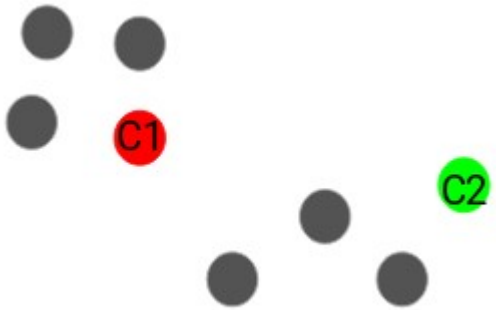
K-means Clustering

K-Means ist ein Centroid-basierter Algorithmus oder ein bei dem wir die Abstände berechnen, um einem Cluster einen Punkt zuzuordnen. In K-Means ist jeder Cluster mit einem Centroid verbunden.

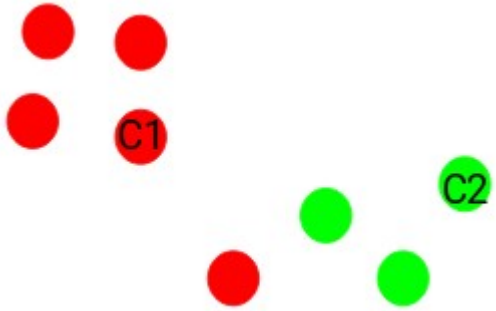
K-means Clustering

Schritt 1: Die Anzahl der Cluster k wählen → k=2

Schritt 2: k zufällige Punkte aus den Daten als Centroids auswählen.

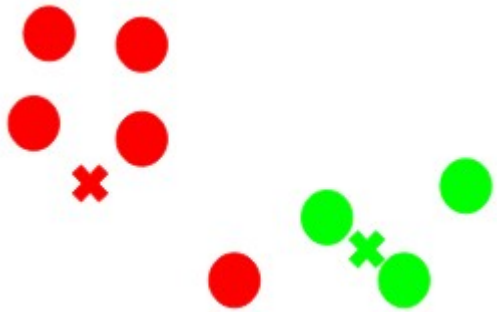


Schritt 3: Alle andere Punkte dem nächstgelegenen Cluster-Centroid zuweisen.



K-means Clustering

Schritt 4: Neuberechnung der Centroids neu gebildeter Cluster.



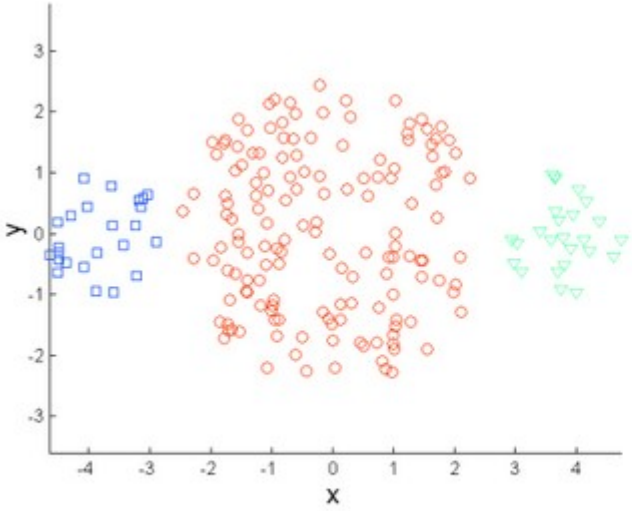
Schritt 5: Schritte 3 und 4 wiederholen.

Wann sollten wir diesen Prozess stoppen?

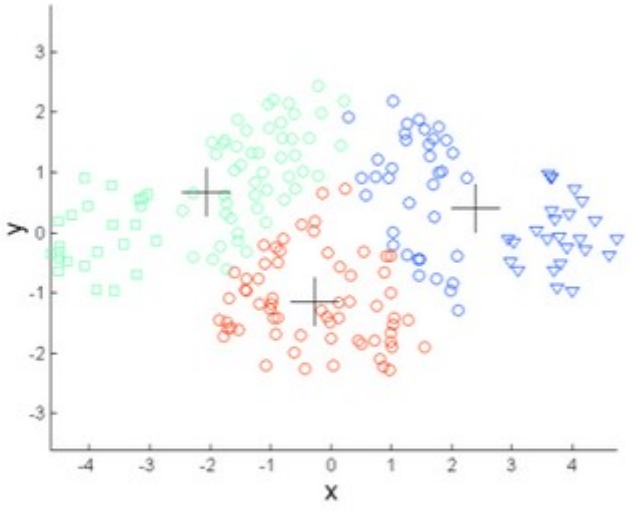
Stop Kriterien

- Die Centroids der neu gebildeten Cluster ändern sich nicht.
- Die Punkte bleiben im gleichen Cluster.
- Maximale Anzahl der Iterationen ist erreicht.

Herausforderung 1:
Die Grösse der Cluster ist unterschiedlich

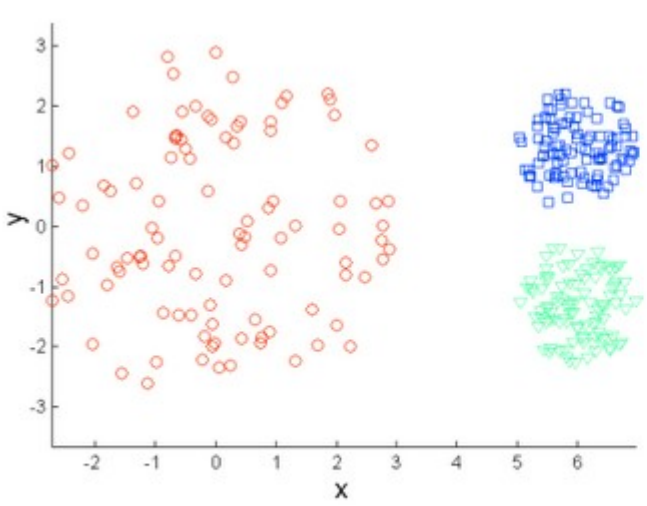


Original Points

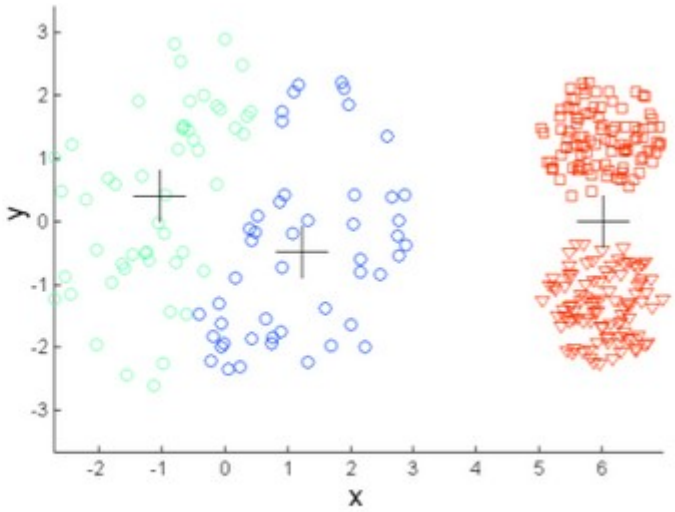


K-means (k = 3)

Herausforderung 2:
Die Dichten der Ursprungspunkte sind unterschiedlich

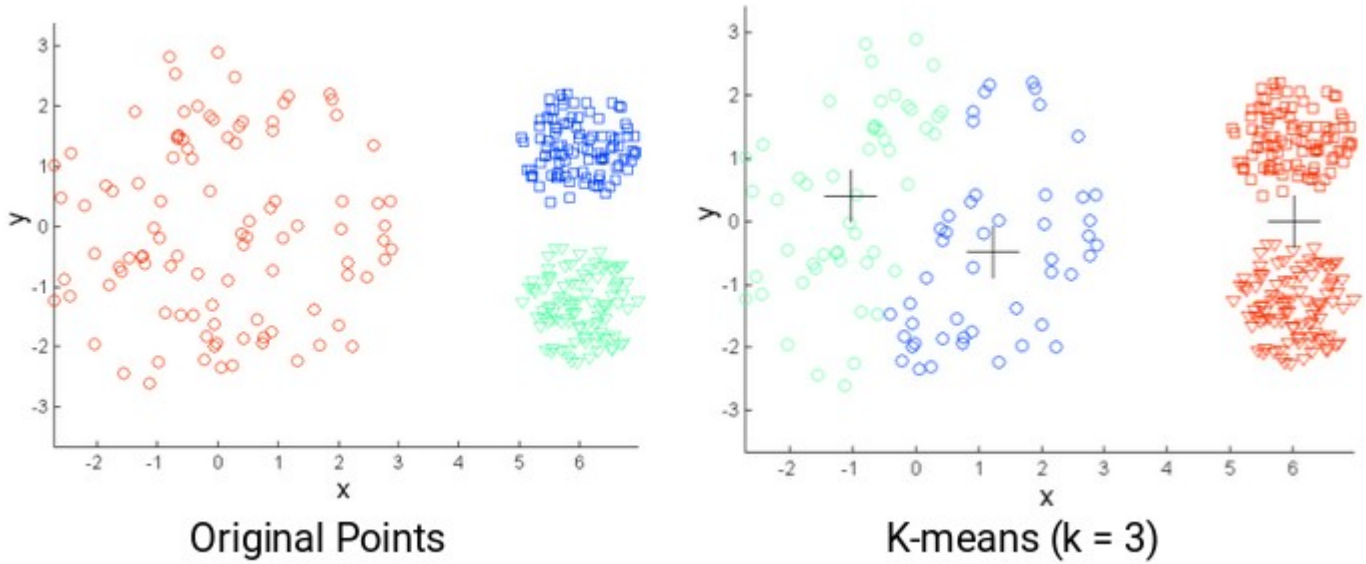


Original Points



K-means (k = 3)

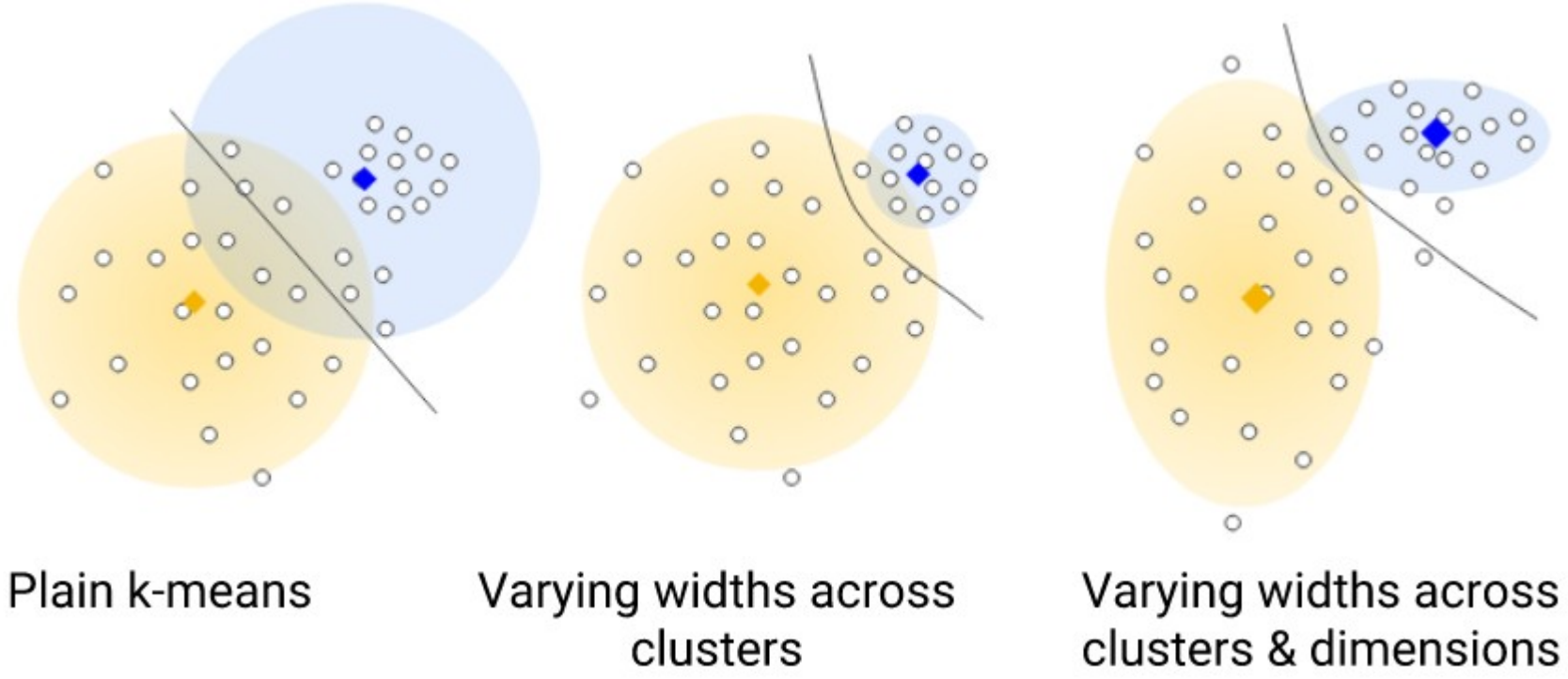
Herausforderung 2:
Die Dichten der Ursprungspunkte sind unterschiedlich



Eine Lösung ist die Verwendung einer höheren Anzahl von Clustern (k>3)

Lösung: K-means Gaussian mixture Modelle

Um natürlich unausgewogene Cluster zu clustern, kann man die k-means anpassen (generalisieren).



k-means++

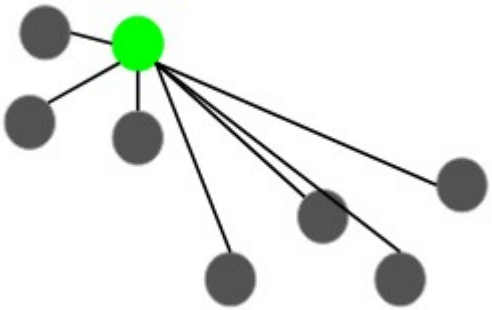
- In k-means die Centroids sind zufälling initialisiert.
- Problem: Jedes Mal könnten wir andere Cluster bekommen.
- Lösung: k-means++. Nur der erste Centroid wird zufälling initialisiert.

k = 3

k-means++

Schritt 1: Centroid 1 zufällig ausgewählt

Schritt 2 Berechnen der Distanz D jedes Datenpunktes von dem bereits gewählten Clusterzentrum aus.



k-means++

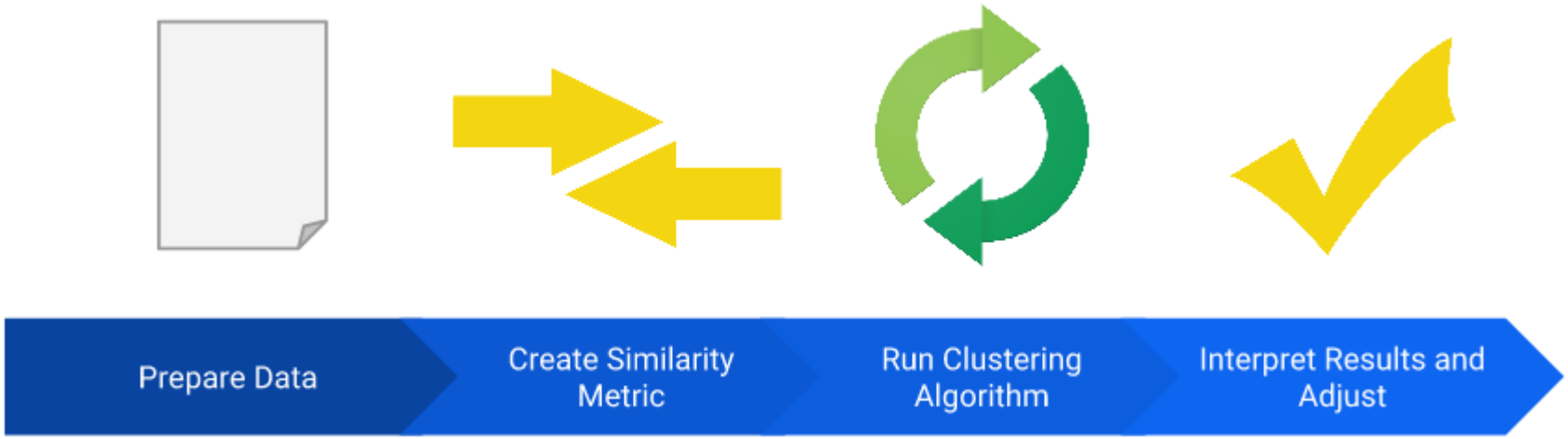
Schritt 3: Der nächste Centroid ist derjenige, dessen quadratischer Abstand (D(x)2) der weiteste vom aktuellen Centroid ist

Schritt 4: um den letzten Centroid auszuwählen, nehmen wir den Abstand jedes Punktes von seinem nächsten Centroid und der Punkt mit dem grössten quadratischen Abstand wird als nächster Centroid ausgewählt

Mit k-means fortfahren

Clustering Workflow

- Data Preparation
- Similarity
- Run Algorithmus
- Ergebnisse Interpretation und Clustering Anpassung



Fragen

- Welche der folgenden Beispiele könnten mit einem unsupervised Learning Algorithmus gelöst werden:

  A) E-Mails sind mit Labels Spam/Nicht-Spam versehen, der Algorithmus soll lernen, Spam zu filtern

  B) Eine Datenbank mit Kundendaten ist vorhanden und es sollen automatisch Markt-Segmente gefunden werden und die Kunden sollen in diese Marktsegmente eingeordnet werden

  C) Ein File mit Patientendaten ist vorgegeben, welche entweder Diabetes haben oder nicht und der Algorithmus soll lernen, von neu dazukommenden Patienten zu entscheiden, ob diese Diabetes haben oder nicht

Fragen

- Eine Firma möchte Learning Algorithmen entwickeln um die folgenden beiden Probleme anzugehen
- Problem 1: Die Firma hat ein grosses Lager von gleichen Artikeln. Sie sollen vorhersagen, wie viele dieser Artikel in den nächsten 2 Monaten verkauft werden.
- Problem 2: Eine Software soll User-Accounts analysieren und für jeden Account entscheiden, ob es gehackt wurde.

Sind dies Klassifikations oder Regressionsprobleme?

A) Beides sind Klassifikationsprobleme

B) P1 ist ein Klassifikationsproblem, P2 ist ein Regressionsproblem

C) P1 ist ein Regressionsproblem, P2 ist ein Klassifikationsproblem

D) Beides sind Regressionsprobleme

Workshop

- 15- Fitness dataset from a user’s Fitbit device. The data are collected in one year from May 2015 to May 2016.

- There is included different kind of activity such as walking, intense training and sleeping.

- Walking and training activity: OneYearFitBitData.csv

- Sleep activity: OneYearFitBitDataSleep.csv