

Explorative Datenanalyse:

EDA Univariate explorative Analyse

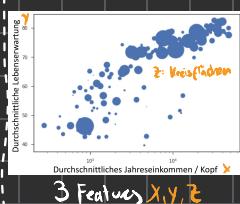
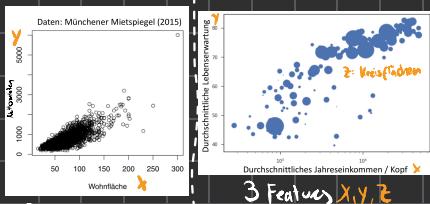
Ziele: ①

1. Identifikation Problem
 2. Prüfen ob beantwortbar
 3. Antrittsskizze
 - ⇒ 4. neue fragt! Hypothesen
- Erwartungshaltung (Fragen)
 - Deskriptive Statistik
 - Visualisierung
 - Dokumentation Ergebniswege

Univariat eindimensionale Daten, ein Feature/Merkmal
multivariat mehrdimensionale Daten, mehrere Features/Merkmale

Multivariate Explorative Analyse

Scatterplot --- Bubble Chart ---



Messung

Abhängigkeitsstrukturen zwischen Merkmalen

Charakterisierung

Richtung des Zusammenhangs

bivariate:

Pearson, Spearman - Korrelationskoeffizient
messen Zusammenhang X, Y
Pearson: linear
Spearman: monoton
Richtung wird nicht erkannt

Quantifizierung

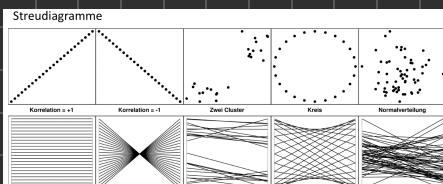
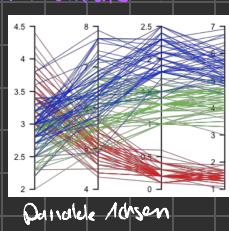
Direct / Indirektheit

PCP Parallel coordinate Plots

Visualisierung

multivariates Daten

Exploration



Descriptive Statistik

Lageparameter zentrale Tendenzen

• Modus, Median, Quantile, arith. Mittel

Stichprobe:	5	7	52	55	75
5 Number Summary:					
Kleiner Wert	5	7	52	55	75
Unteres Quartil					
Median					
Oberes Quartil					
Großer Wert					

Empirische p -Quantile

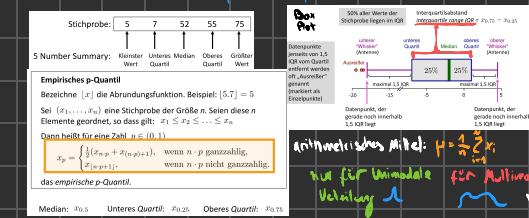
Berechnet: $|x|$ die Abrundungsfunktion. Beispiel: $[5,7] = 5$

Sei (x_1, \dots, x_n) eine Stichprobe der Größe n . Seien diese n Elemente geordnet, so dass gilt: $x_1 \leq x_2 \leq \dots \leq x_n$

Dann heißt F für eine Zahl $n \in [0, 1]$

$$x_F = \begin{cases} x_{\lceil n \cdot F \rceil} & \text{für } n \cdot F \text{ ganzzahlig,} \\ \frac{x_{\lfloor n \cdot F \rfloor + 1} + x_{\lfloor n \cdot F \rfloor}}{2} & \text{für } n \cdot F \text{ nicht ganzzahlig.} \end{cases}$$

Median: $x_{0.5}$ Unteres Quartil: $x_{0.25}$ Oberes Quartil: $x_{0.75}$



Streuungsparameter

• Varianz, Standardabweichung, Interquartilsabstand

$$\text{Standardabweichung } \sigma = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

$$\text{Varianz } S^2 = \frac{1}{n-1} \sum (x_i - \bar{x})^2$$

1. Maximieren Daten-Durchsichtbarkeit
2. Minimieren Lügenfaktor
3. Minimieren "Uniquität"
4. Angemessene Skalen (Residualität!)

Regeln nach Edward Tufte

Historgramm: graphische Darstellung einer Häufigkeitsverteilung durch Einheiten der Daten in Klassen (englisch: bins).

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Diagramm: Eine hierarchische Struktur von Kategorien und Unterkategorien.

Flowchart: Eine Abfolge von Schritten und Entscheidungen.

Infografik: Eine Kombination aus verschiedenen Darstellungsformen.

Timeline: Eine zeitliche Abfolge von Ereignissen.

Map: Eine geographische Darstellung von Orten und Daten.

Tableau: Eine Tabelle mit Daten.

Diagramm: Eine hierarchische Struktur von Kategorien und Unterkategorien.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Scatterplot: Eine Punktwolke mit Achsen und Legende.

Boxplot: Eine Box mit Whiskers und Medianlinie.

Mutual Information

Definition (Informationsgehalt)

Informationsgehalt I_k für das Eintreffen eines Ereignisses k mit Wahrscheinlichkeit p_k sei:

$$I_k := -\log_2 p_k$$

Maßeinheit: bit

Je seltener Ereignis $k \Rightarrow$ Je höher Informationsgehalt

Definition (Entropie)¹: Erwartungswert Informationsgehalt für Zufallsexperiment

Der mittlere Informationsgehalt eines Ereignisses (Ausgangs) eines Zufallsexperimentes mit Zufallsvariable X heißt **Entropie**

$H(X)$.

$$H(X) := \mathbb{E}[I] = -\sum_{k=1}^C p_k \log_2 p_k \quad \text{mit } 0 \log_2 0 = 0$$

↓
Erwartungswert

Mittelung gewichtet bzgl der
Wahrscheinlichkeiten der
Ereignisse (bzw. der Klassen)

(entsprechend dem Grenzwert:
 $\lim_{x \rightarrow 0} x \log_2 x = 0$)

Maß für Unordnung / (Un-)Vorhersagbarkeit

- $p(x_i, y_j)$ bezeichne die gemeinsame Wahrscheinlichkeit (Verbundwahrscheinlichkeit) für das gemeinsame Auftreten von x_i, y_j
- $p(x_i|y_j) = \frac{p(x_i, y_j)}{p(y_j)}$ bezeichne die bedingte Wahrscheinlichkeit für x_i unter der Bedingung, dass y_j vorgegeben ist.

Entropie von X unter der Bedingung des Auftretens eines Wertes y_j :

$$H(X|Y = y_j) = -\sum_i p(x_i|y_j) \log_2 p(x_i|y_j) \quad (1)$$

Definition (bedingte Entropie)

Der mittlere Informationsgehalt eines Ergebnisses einer Zufallsvariablen X unter der Bedingung, dass der Wert einer Zufallsvariablen Y bekannt ist, heißt **bedingte Entropie** $H(X|Y)$.

$$\begin{aligned} H(X|Y) &= \sum_j p(y_j) H(X|Y = y_j) \\ &= -\sum_{i,j} p(y_j) p(x_i|y_j) \log_2 p(x_i|y_j) \\ &= -\sum_{i,j} p(x_i, y_j) \log_2 \frac{p(x_i, y_j)}{p(y_j)} \end{aligned}$$

Dies ist die gewichtete Summe
des Ausdrucks (1) für jeden
möglichen Wert von Y .
weil $p(x_i|y_j) = \frac{p(x_i, y_j)}{p(y_j)}$

Definition (Mutual Information)

Die Abnahme des mittleren Informationsgehalts eines Ergebnisses der Zufallsvariablen X durch Kenntnis des Ergebnisses einer Zufallsvariablen Y heißt **Mutual Information**.

$$I(X; Y) = H(X) - H(X|Y)$$

Auf Deutsch

Die Mutual Information misst, wieviel Information Y über X offenbart.

X Roger } wenn ich \Rightarrow keine kann
Y Regen } ich das andere besser
vorhersage

Mutual Information hoch \Rightarrow Informationsgewinn

$$\begin{aligned} p(x, y) &= \text{Eintreten von } x \text{ UND } y \\ p(x)p(y) &= \text{Wahrscheinlichkeit } X, Y \text{ unabhängig} \\ p(x|y_i) &= \frac{p(x_i, y_i)}{p(x_i)p(y_i)} \quad \text{Wahrscheinlichkeit } x_i \text{, wenn } y_i \text{ gegeben} \end{aligned}$$

Entropie $H(x)$: Unsicherheit über X

bedingte Entropie $H(x|y)$: Unsicherheit X , wenn y bekannt

$$MI = I(x; y) = H(x) - H(x|y)$$

Eigenschaften

symmetrisch $I(x; y) = I(y; x)$

nicht-negativ $I(x; y) \geq 0$

Multivariate Explorative Analyse

Korrelation & Kausalität

• hohe Korrelation $\not\Rightarrow$ Kausalität

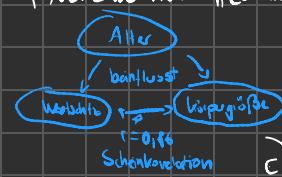
 nur Hinweis

↳ Nur durch Experimente

Zusammenhangsmaße | Interpretationsfehler

1. Scheinkausalität

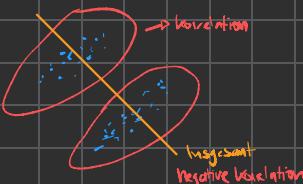
a) Netzwerkativ „common driver“ b) Netzwerkativ „indirekte Beziehung“



c) Zweifellige Korrelation

2. Verdeckte Korrelation

bsp Population zerfällt in Teilpop.



Dimensionsreduktion

Ziel: Reduzieren Merkmale (Dimensionen)
OHNE großen Informationsverlust

PCA Hauptkomponentenzerlegung

Idee: Neue Achsen finden mit möglichst großer Varianz
(durch Orthogonale Projektion)
 \Rightarrow andere weglassen

ges.

x_n : N Datenpunkte

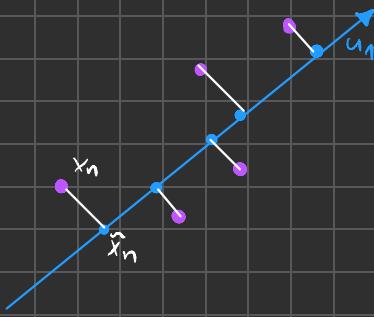
D Merkmale

$$\bar{x} = \frac{1}{N} \sum_{n=1}^N x_n \text{ Mittelpunkt}$$

$$u_1: \text{Einheitsvektor } u_1^\top \cdot u_1 = 1$$

ges.

Richtung u_1 mit Varianz des Projektions
auf u_1 maximiert



$$\text{Projektion } x_n \text{ auf } u_1 = u_1^\top x_n$$

$$\text{Kovarianzmatrix: } S = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})(x_n - \bar{x})^\top$$

$$\text{Var}(u_1^\top x_n) = \frac{1}{N} \sum_{n=1}^N (u_1^\top x_n - u_1^\top \bar{x})^2$$

$$= u_1^\top S u_1 \leftarrow \text{soll maximiert werden! NB: } u_1^\top \cdot u_1 = 1$$

Optimierungsproblem ↴

Lösung

Richtung der größten Varianz
 \Rightarrow meiste Information

1. Zentrieren der Daten (Mittelwert 0): $\tilde{x}_i = x_i - \bar{x}$

Notwendig: Mittelpunkt beeinflusst PCA-Ergebnisse (Offset in nachfolgenden Schritten)

2. (Optional) Standardisierung (Z-Skoring) (Varianz 1)

\bar{x} Mittelpunkt σ Standardabweichung

$$z = \frac{x - \bar{x}}{\sigma} \quad (\text{größere Zahlen = mehr Gewicht}) \quad (\text{Invariant gegenüber Verschiebung, aber nicht Skalierung})$$

Empfohlen: wenn Skalen der Merkmale unterschiedlich

in Matrix: jede Spalte (Merchmal) separat auf Mittelpunkt 0, Varianz 1 transf.

3. Berechnung Kovarianzmatrix S

$$\begin{bmatrix} \text{Var(Größe)} & \text{cov(Größe, Geschlecht)} \\ \text{cov(Geschlecht, Größe)} & \text{Var(Größe)} \end{bmatrix} \quad x_{\text{centred}} = \text{data} - \text{np.mean(data, axis=0)}$$

$$S = (x_{\text{centred}}^\top \otimes x_{\text{centred}}) / N$$

4. Eigenwertzerlegung von S eigvals, eigvecs = np.linalg.eigh(S) idx = np.argsort(eigvals)
größte Varianz = größtes Eigenwert idx[-1] ↓ sort

5. 4 größte Eigenwerte (4 = gewünschte Hauptkomponenten)

$$W = \text{eigvecs[:, :4]}$$

$$reduced = (x - \text{np.mean(x, axis=0)}) @ W$$

PVE

$$\frac{\text{Var(eig-val)}}{\text{Var total}} = \frac{\text{Var(eig-val)}}{\sum \text{Var(eig-val)}}$$

$$\text{pve} = \text{eig-val.sum() / np.sum(eig-val)}$$

$$\text{cum-pve} = \text{np.cumsum(pve)}$$

PCA

Datensatz $X \in \mathbb{R}^{n \times p}$ hat n : Datenpunkte p : Dimensionen (Features)

Empfehlung

- Zentrieren Sie jedes Merkmal auf Mittelwert 0 vor Anwendung der PCA:

$$\tilde{x}_i = x_i - \bar{x}$$

PCA:

1. Kovarianzmatrix

$$\begin{bmatrix} \text{Var(Größe)} & \text{(ou (Größe, Gewicht))} \\ (\text{ou (gewicht, Größe)}) & \text{Var(Größe)} \end{bmatrix}$$

$\{\mathbf{x}_n\}$: N Datenpunkte mit $n=1, \dots, N$
mit D Merkmalen

Mittlerer Vektor
der Daten $\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n$

$$S = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T$$

$$\mathbf{x}_{\text{centred}} = \mathbf{data} - \mathbf{data}.mean$$

- ### 2. Eigenvektoren der Kovarianzmatrix
- $S = (\mathbf{x}_{\text{centred}}.T @ \mathbf{x}_{\text{centred}}) / N$
 = PCA-Komponenten nplmngs `eigh(cov_matrix)`

`iidx = np.argsort(eig_vals)`

`iidx[-1:-1]` ↓ sort

`eigvals[iidx]`

`eigvecs[:, iidx]`

PVE

eig_vals_system / sum(eig_vals_system)

$$\text{Var}(z_i) = \lambda_i$$

$$\text{Var}_{\text{Total}} = \sum_{i=1}^D \text{Var}(z_i) = \sum_{i=1}^D \lambda_i$$

$$PVE(i) = \frac{\text{Var}(z_i)}{\text{Var}_{\text{Total}}} = \frac{\lambda_i}{\sum_{i=1}^D \lambda_i}$$

= Bruchteil der Gesamtvarianz der Merkmale über PCA-Komponente

Eigenvektor \rightarrow Richtung der Hauptkomponente

Eigenwert \rightarrow Varianz jeder Komponente

3. Coeffizienten erstellen

$$f_{-k} = \text{eig_vals_system}[:, :k]$$

$$g = \mathbf{data}[:, i]$$

$$m = \text{np.mean}(\mathbf{data}, axis=0)$$

$$c = f_{-k}.T @ (g - m)$$

$$g_{\text{rekonstrukt}} = f_{-k} @ c_{-k} + m$$

PCA Nachträge

X : ($N \times P$) Matrix mit N Datenpunkten zu je P Merkmalen (Features)
Featurevektoren sind die Zeilen dieser Matrix

$$\bar{\mathbf{x}} = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n = \frac{1}{N} X^T \vec{1} \quad \text{mittlerer Vektor der Daten (Mittelwertsvektor)}$$

Annahme für diese Vorlesungseinheit: Daten X sind schon zentriert,
das heißt: $\bar{\mathbf{x}} = \mathbf{0}$

Kovarianzmatrix

$$S = \frac{1}{N} \sum_{n=1}^N (\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T = \frac{1}{N} \sum_{n=1}^N \mathbf{x}_n \mathbf{x}_n^T = \frac{1}{N} X^T X$$

Idee

Algorithmen zur Eigenwertzerlegung skalieren typischerweise mit $O(P^3)$.

Typische Fälle in der Praxis:

$P \ll N$

- viele Datenpunkte mit „wenig“ Dimensionen

Eigenwertzerlegung der ($P \times P$) Kovarianzmatrix:
 $S = \frac{1}{N} X^T X$
 (kennen Sie schon)

$N \ll P$

- wenige Datenpunkte mit „vielen“ Dimensionen

Eigenwertzerlegung der ($N \times N$) Matrix:
 $\frac{1}{N} X X^T$
 (betrachten wir jetzt)

Falls $N \ll P$ ist es rechnerisch meist günstiger, die PCA über die Eigenwertzerlegung der $N \times N$ Matrix $\frac{1}{N} X X^T$ durchzuführen:

- Bestimme Eigenwerte λ_i und Eigenvektoren \mathbf{v}_i von $\frac{1}{N} X X^T$. Eigenwerte λ_i sind auch Eigenwerte der Kovarianzmatrix S .
- Wegen $\underbrace{\frac{1}{N} X^T X}_{S} (\mathbf{v}_i^T \mathbf{v}_i) = \lambda_i \mathbf{v}_i^T \mathbf{v}_i$ (siehe vorherige Folie)
- Setze als Eigenvektoren der Kovarianzmatrix S : $\mathbf{u}_i \propto X^T \mathbf{v}_i$
- Normiere die Eigenvektoren auf Länge 1: $\mathbf{u}_i = \frac{1}{\|\mathbf{X}^T \mathbf{v}_i\|} X^T \mathbf{v}_i$

MDS

Multidimensional Scaling

= Verfahren Visualisierung von Punkten / Dimensionsreduktion

(Classical Multidimensional) Scaling = Torgerson-Scaling

geg.

$D = d_{ij}$, $N \times N$ Matrix der quadratischen paarweisen Distanzen

ges.

Datenmatrix X + unbekannt!

Koordinaten (Features) jenseits der N Objekte (Approximation)

\Rightarrow Visualisierung

\Rightarrow nicht eindeutig Transformationen (Spiegelung, Rotation), Koordinatenursprungswahl

$$B = -\frac{1}{2} H D H^T \quad H = I - \frac{1}{N} \mathbf{1} \mathbf{1}^T$$

Eigenwertzerlegung B

($B^T = B$)

$BV = V\Lambda$ (V : Matrix Eigenvekt., Λ : Diagonalmatrix Eigenwerte)

$\Leftrightarrow B = V\Lambda V^T$ (V=U $^{-1}$)

$\Leftrightarrow B = V\Lambda V^T$ spektrale Zerlegung

$$\Rightarrow X = VA^{\frac{1}{2}} = [\underbrace{\sqrt{\lambda_1} \cdot v_1, \dots, \sqrt{\lambda_p} \cdot v_p}_{\text{Ersten } p}] \quad \begin{array}{l} \text{Eigenwerte \& Vektoren} \\ \text{nach Größe absteigend} \end{array}$$

Wenn $x^T x = 0$ MDS

- Datenmatrix X zentriert ist (d.h. Mittelwertvektor ist 0)
- Distanzmatrix D aus euklidischen Distanzen besteht dann sind
- MDS-Koordinaten der Daten = PCA-Koordinaten der Daten

Warum benötigen wir dann überhaupt MDS, wenn wir PCA schon kennen? Weil:

- Andere Ausgangssituation (Distanzmatrix bekannt, Koordinaten unbekannt) nicht known
- MDS ist Ausgangspunkt für verschiedene Verfahren, z.B. Isomap

MDS Als Dimensionenreduktionsverfahren

1. $x: (N \times P)$ geg. N : Datenpunkte P : Dimensionen

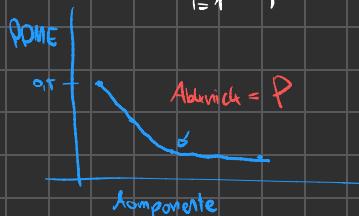
2. D bestimmen (z.B. euklidische Distanz)

3. MDR mit $Q \subset P = x^T (N \times Q)$

P bestimmen = Dimensionenwahl

Proportion of Distance Matrix Explained

$$PDME(i) = \frac{\lambda_i}{\sum_{i=1}^N |\lambda_i|} \quad \text{mit } \lambda_i \geq 0$$



PDME

- Visualisierung komplexe Beziehungen
- Dimensionsreduktion

Isomap

Algorithmus

- Ermittle Nachbarschaftsgraph mit euklidischen Distanzen:
 - Verbinde einen Punkt mit seinen k -nächsten Nachbarn
 - Verbinde einen Punkt mit allen Nachbarn innerhalb eines Epsilonballs
- Ermittle paarweise Distanzen zwischen Punkten auf dem Graph; quadriere diese Distanzen \rightarrow Distanzmatrix D
- Wende Multidimensional Scaling auf Distanzmatrix D an und erhalte neue Datenmatrix X .

p_1, p_2, p_3

p_1

p_2

p_3

08 Clustering

= Identifikation von Gruppen (**Clustern**) ähnlicher Datenpunkte in einem Datensatz
& Einteilung

Daten \rightarrow Cluster-Vorfahren \rightarrow Cluster
interpretieren die Daten
es werden immer Cluster zuordnen

\rightarrow Muß bewertet werden ob Ergebnis sinnvoll

Welches Clustering ist richtig?

Hypoparameter (mittel zur Bestimmung):

- Anzahl der Cluster

K-Means Clustering

K: muss vorher angegeben werden
findet K - nicht überlappende Cluster

Seien C_1, \dots, C_K die Mengen der Indizes der Datenpunkte jedes Clusters.

Die Mengen sollen zwei Eigenschaften genügen:

- $C_1 \cup C_2 \cup \dots \cup C_K = 1, \dots, N$
Auf Deutsch: Jeder Datenpunkt gehört zu **mindestens einem Cluster**. \Rightarrow keine **nicht zugehörigen Datenpunkte**
- $C_k \cap C_{k'} = \emptyset \quad \forall k \neq k'$
Auf Deutsch: Die **Cluster überlappen nicht**. Jeder Datenpunkt gehört nicht mehr als einem Cluster an.

Hauptidee

$$\text{minimiere}_{C_1, \dots, C_K} \left\{ \sum_{k=1}^K W(C_k) \right\}$$

Intra-Cluster-Variation
• wird meist definiert als Summe der quadratischen euklidischen Distanzen zwischen allen Datenpunkten eines Clusters

$$W(C_k) = \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2$$

Anzahl Datenpunkte im Cluster k | für Komponente des Datenpunktes (Featurevektor)

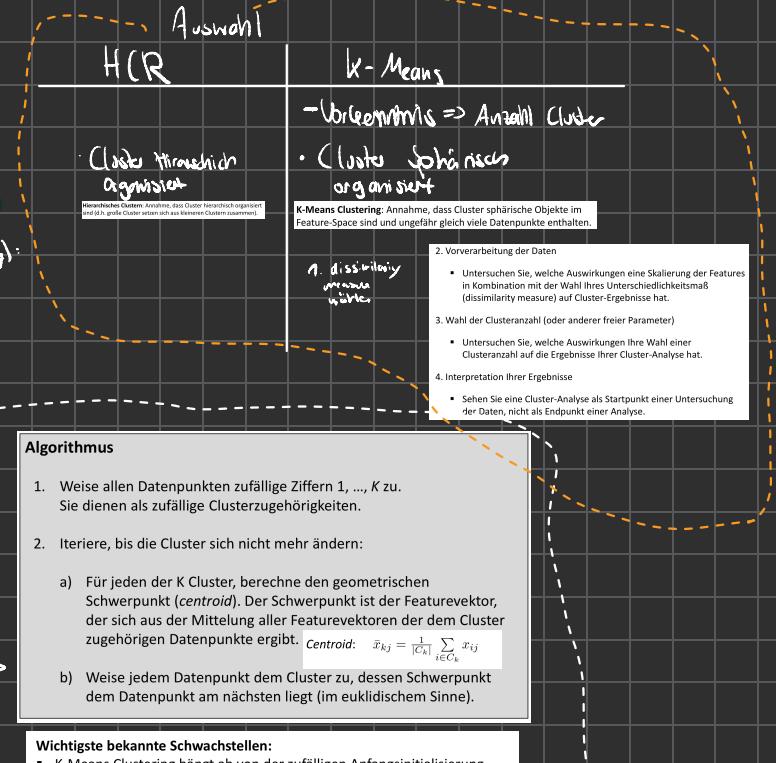
Optimierungsproblem – k-means Clustering
(Clustering durch Varianzminimierung)

$$\text{minimiere}_{C_1, \dots, C_K} \left\{ \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \right\}$$

kein outlier

(crispy (nicht-fuzzy))

sehr schwere Ränder



Wichtigste bekannte Schwachstellen:

- K-Means Clustering hängt ab von der zufälligen Anfangsinitialisierung der Clusterzugehörigkeiten (Schritt 1) \rightarrow Findet lokale Minima unserer Optimierungsfunktion, nicht notwendigerweise das globale Minimum.
- Clusteranzahl K muss zu Beginn festgelegt werden. Verschiedene Werte für K können ganz unterschiedliche Cluster liefern.
- Keine Möglichkeit, Ausreißer in den Daten zu erkennen und getrennt zu behandeln.

Empfehlung:

- Starten Sie K-Means mehrfach mit zufälligen initialen Clusterzugehörigkeiten. Wählen Sie das Ergebnis aus, das den kleinsten Wert für die summierte Intra-Cluster Variation aufweist:
 $\sum_k W(C_k)$ (vgl. Folien zum Optimierungskriterium)

HCA Hierarchische Clusteranalyse

- einfache (und oft genutzte) Methode zum Finden von Clustern in Daten
- erzeugt eine Menge verschiedener Partitionierungen (=Aufteilung der Daten in Cluster), aus denen eine Partition (Clustering) ausgewählt wird.
- Zentrales Konzept: Aufbau eines Dendrogramms

Dendrogramm: Diagramm, das einen Baum repräsentiert.

Baum: Ungerichteter Graph, in dem jedes Paar von Knoten genau durch nur einen Pfad miteinander verbunden ist.

- Bäume können aufgebaut werden
- von den Blättern hin zur Wurzel (agglomeratives Clustern, Bottom-Up-Ansatz)
 - von der Wurzel hin zu den Blättern (divisives Clustern, Top-Down-Ansatz)
- Vorlesung behandelt diesen Ansatz.

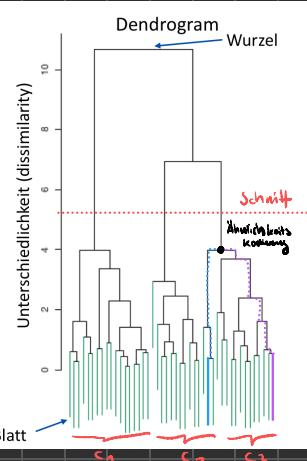
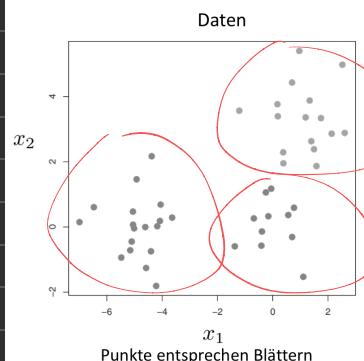
W x 201
divi. agglomerativ
Blätter

Algorithmus (agglomeratives hierarchisches Clustern)

- Betrachte alle N Datenpunkte und interpretiere sie als N **separate Cluster**. Wähle ein Maß für die Unterschiedlichkeit (dissimilarity measure) aus.
- Für $i = N, N-1, \dots, 2$:
 - Bestimme die paarweisen **Unterschiedlichkeiten für alle Cluster**. \rightarrow Linkage
 - Fusioniere das Clusterpaar, das sich am ähnlichsten ist, zu einem neuen Cluster. Die Unterschiedlichkeit zwischen den beiden soeben fusionierten Clustern entspricht der Höhe im Dendrogramm, in der der Zusammenschluss stattfindet.

Beispiel

(später wird Algorithmus behandelt)



y-Achse = hohe Unähnlichkeit
x-Achse = Bedeutigkeit

- Jeder Datenpunkt entspricht einem Blatt im Dendrogramm.
- Je höher wir im Baum wandern, desto mehr Blätter werden zu Zweigen verbunden.
- Je früher (= weiter unten) Zweige/Blätter verbunden werden, desto ähnlicher sind die entsprechenden Datenpunkte.
- Für jedes Paar von Datenpunkten zeigt die Höhe (y-Achse) der Verbindung der beiden ihre Unterschiedlichkeit (dissimilarity) an.

Complete Linkage
= maximale Inter-Cluster-Unterschiedlichkeit

- Alle paarweisen Unterschiedlichkeiten zwischen Punkten aus dem ersten und Punkten aus dem zweiten Cluster bestimmen.
- Unterschiedlichkeit zwischen beiden Clustern entspricht der **größten** Unterschiedlichkeit aus Schritt (1).

Beispiel **komplexe gleichfarbige Cluster**



Single Linkage **schwierige Skalierung**
= minimale Inter-Cluster-Unterschiedlichkeit

- Alle paarweisen Unterschiedlichkeiten zwischen Punkten aus dem ersten und Punkten aus dem zweiten Cluster bestimmen.
- Unterschiedlichkeit zwischen beiden Clustern entspricht der **kleinsten** Unterschiedlichkeit aus Schritt (1).

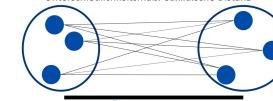
Beispiel **schlauchartige Struktur**



Average Linkage
= mittlere Inter-Cluster-Unterschiedlichkeit

- Alle paarweisen Unterschiedlichkeiten zwischen Punkten aus dem ersten und Punkten aus dem zweiten Cluster bestimmen.
- Unterschiedlichkeit zwischen beiden Clustern entspricht dem **Mittelwert** der Unterschiedlichkeiten aus Schritt (1).

Beispiel **viel kleinere Strukturen**



Unterschiedlichkeit auf Ebene der Datenpunkten
 \rightarrow dissimilarity measure (Unterschiedlichkeitsmaß)

- hat wichtigen Einfluss auf das Clustering-Ergebnis
- oft wird euklidische Distanz gewählt, **hierarchisch kohärent**

hierarchisch kohärent

Muster Wahl

	K-Means	PCA
Kategorie	Partitionierendes Verfahren 1. Zufällig k Start-Centroids (Datenpunkte) \hookrightarrow 2. weise jedem Punkt nächstgelegenes Centroid zu (euklidische Distanz) \hookrightarrow 3. für jedes Cluster neue Centroids (Mittelwert aller Punkte) \hookrightarrow bis Zuweisung sich nicht mehr ändert	Hierarchisches Verfahren euklidisch von. Koeffizient. 1. Start Punkt = Cluster 2. Distanzmatrix zwischen Clustern (Dissimilarity measure) \hookrightarrow 3. 2 Cluster mit geringer Distanz zusammenführen \Rightarrow 1 neues Cluster \hookrightarrow 4. Distanzmatrix aktualisieren [Linkage-Methode] \hookrightarrow Abbrechen nur wenn 1 Cluster \hookrightarrow bestimmt Anzahl Cluster
(Cluster-Anzahl nötig?)	JA	NEIN erzeugt Dendrogramm \Rightarrow (Clusterzahl) kann später gewählt werden
(Clusterform)	Konvex / Kreuzförmig (am besten)	Pbeliebige Formen nicht gut bei starken Überlappungen
Zuweisung	Hart (Punkt \mapsto Cluster)	Hart (aber Rauen Struktur)
Ergebniss	Clusterzuweisung	Hierarchischer Baum (Dendrogramm)
Robustheit gg. Ausreißer	empfindlich	Single: sehr empfindlich Complete: relativ robust Average: mittelmäßig robust
Rechenaufwand	Schnell (linear)	langsam (quadratisch)
Typische Anwendung	Große Datensätze \hookrightarrow bekannt oder simuliert	Explorative Datenanalyse Clusteranzahl unbekannt
Einfussfaktoren	Schlechte Startpunkte \hookrightarrow konzentriert lokale Minima \Rightarrow mehrfach ausführen, Cluster = $\min \{ \sum W(c_u) \}$	Dissimilarity Measure \Rightarrow passendes wählen 1. Wahl Dissimilarity Measure 2. Strategie der Features + Proz. Meth. \Rightarrow Auswählen 3. Wahl (Clusteranzahl) \hookrightarrow Startpunkt Analyse

Intra-Cluster-Variation
 wird meist definiert als Summe der quadratischen euklidischen Distanzen zwischen allen Datenpunkten eines Clusters

$$W(C_k) = \frac{1}{|C_k|} \sum_{i \in C_k} \sum_{j=1}^p (x_{ij} - \bar{x}_{kj})^2$$

Anzahl Datenpunkte im Cluster k
 selektiert Datenpunkte aus Cluster k

Clustervalidierung

Beobachtung

- Cluster-Analyseverfahren finden immer Cluster in den Daten.
- Welches Clustering (Partitionierung) ist sinnvoll, welches nicht?

Clustervalidierungsverfahren

- bewerten Qualität von Clusterings (Partitionen)
 - ermöglichen das Einstellen von Hyperparametern (Beispiel: Anzahl K der Cluster bei K-Means)

keine feste Lösung

Werkzeuge

① Test auf Abwesenheit von Cluster

Uniformitätsprinzip a) liegen zufällig verteilt (uniform) im Merkmalsraum

Unimodale Multimodale Hypothese b) Daten wurden aus unimodalem Verhalten gegen meist nur im Fehlgründen Kontakt (fast trivial im Praxi vorkommen)

② Externe Clustervalidierung

externe Informationen werden genutzt

Sanity (Richtigkeit von Clustern)

C_1, \dots, C_K Mengen der Indizes der Datenpunkte jedes Clusters.

T_1, \dots, T_K Mengen der Indizes der Datenpunkte der wahren Cluster gemäß der externen Information (*ground truth*)

$N_i = |C_i|$ Anzahl Datenpunkte in Cluster i (kommen wir mehrfach nicht)

$\text{purity}_i = \frac{1}{N_i} \max_{j=1}^K |C_i \cap T_j|$ Reinheit (Purity) des Clusters C_i wird 1, wenn Cluster C_i nur Punkte eines wahren Clusters der Partition T enthält

$\text{purity} = \sum_{i=1}^K \frac{N_i}{N} \text{purity}_i = \frac{1}{N} \sum_{i=1}^K \max_{j=1}^K |C_i \cap T_j|$ wird 1, wenn alle Cluster C_i nur Punkte eines wahren Clusters enthalten

Randfälle: Jeder Cluster enthält nur ein Datensymbol. Es gibt nur ein Cluster

Mutual Information

Mutual information (misst Ähnlichkeit zwischen Partitionen C und T)

$$I(C, T) = \sum_{i,j} p_{ij} \log_2 \left(\frac{p_{ij}}{p_{Ci} p_{Tj}} \right) \quad \text{mit} \quad p_{ij} = \frac{|C_i \cap T_j|}{N}$$

→ kennen Sie bereits aus Vorlesung 5. und $p_{Ci} = \frac{|C_i|}{N}$ Wahrscheinlichkeit, dass ein Punkt aus Cluster i zur Partition T aus T gehört

Je größer die Mutual Information, desto besser das Clustering.

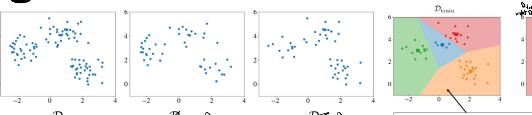
$$p_{Tj} = \frac{|T_j|}{N} \quad \text{Wahrscheinlichkeit für Cluster } C_j \text{ bzw. für Cluster } T_j$$

Prediction Strength

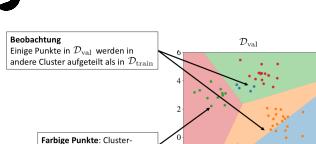
Idee: Clusterqualität ist hoch, wenn Clusterzugehörigkeiten auf anderen Realisation der Daten zuverlässig vorhergesagt werden können.

Vorgehen:

1 Teilen Sie die Daten (D) zufällig auf in zwei Mengen: Training- und Validierungsset ($D_{\text{train}}, D_{\text{val}}$)



3 Ermitteln Sie für alle Daten in D_{val} die Clusterzugehörigkeit gemäß des Clustering in D_{train} .



Beobachtung: Einige Punkte in D_{val} werden in andere Cluster aufgeteilt als in D_{train}

Farbige Punkte: Clusterzugehörigkeit gemäß D_{train}

Farbige Flächen: Cluster gefunden über K-Means in D_{val}

2 Bestimmen Sie für jeden Cluster in D_{val} den Bruchteil p aller Paare von Datenpunkten, die sich auch im selben Cluster in D_{train} befinden würden.

Frage: Welchen Wert p hat der - rote Cluster in D_{val} ? - blaue Cluster in D_{val} ? - grüne Cluster in D_{val} ?

Antwort: Der kleinste Wert p über allen Clustern heißt prediction strength.

Farbige Punkte: Clusterzugehörigkeit gemäß D_{train}

Farbige Flächen: Cluster gefunden über K-Means in D_{val}

Prediction Strength – formale Behandlung

Sei $C = \{C_1, \dots, C_K\}$ ein Clustering von D_{train} mit Indexmengen C_i (Cluster in D_{train}), deren Datenpunkte in den jeweiligen Regionen R_{C_i} des Merkmalsraum liegen.

Sei $A = \{A_1, \dots, A_K\}$ ein Clustering von D_{val} mit Indexmengen A_i (Cluster in D_{val}).

Sei M eine $N_{\text{val}} \times N_{\text{val}}$ Matrix (sog. Ko-Mitgliedschaftsmatrix), wobei N_{val} die Anzahl der Datenpunkte in D_{val} bezeichnet:

$$M_{ii'} := \begin{cases} 1, & \exists k \text{ mit } (i, i') \in R_{C_k} \\ 0, & \text{sonst} \end{cases} \quad \text{d.h.: Matrixeintrag ist 1, wenn die jeweiligen zwei Datenpunkte aus dem Validierungsset zur selben Region eines Clusters k im Trainingsset gehören}$$

Prediction strength: $ps(K) = \min_{j=1, \dots, K} \left[\frac{1}{|A_j|(|A_j|-1)} \sum_{i, i' \in A_j, i \neq i'} M_{ii'} \right]$

Summe über Paare im Cluster A des Validierungssets

9 - 24

Wie werden die Regionen R_{C_i} ermittelt?

Ermittlung der Regionen

- hängt vom jeweiligen Clusterverfahren ab
- kodiert Vorstellung davon, was ein Cluster ist

K-Means

interpretiert Cluster als sphärische Objekte im Merkmalsraum

Regionen sind Polygone: Centroids der Cluster C_i definieren Regionen und damit Clusterzugehörigkeit!

Farbige Punkte: Clusterzugehörigkeit gemäß D_{train}

Farbige Flächen: Cluster gefunden über K-Means in D_{val}

Single Linkage

betrachtete Abstände sind kleinste euklidische Distanzen

Datenpunkt i aus D_{val} wird dem Cluster aus D_{train} zugeordnet, zu dem er den kleinsten euklidischen Abstand hat

- Für gegebenen Datenpunkt i aus D_{val} suche nächsten Nachbarn j in D_{train}
- Übernehme Clustermitgliedschaft von j für den Punkt i

Werte 1 Jahr stabil

2018 instabil

⇒ u. leichteren Verlustes nehmen bei den PS > 80 ->

$ps(K)$

$$= \frac{1}{N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} \frac{1}{N_{\text{val}}} \sum_{i'=1, i' \neq i}^{N_{\text{val}}} M_{ii'} = \frac{1}{N_{\text{val}}^2 - N_{\text{val}}} \sum_{i=1}^{N_{\text{val}}} \sum_{i'=1, i' \neq i}^{N_{\text{val}}} M_{ii'}$$

⇒ dieses Maß

- betrachtete Abstände sind mittlere euklidische Distanzen
- Weise Datenpunkt i aus D_{val} dem Cluster aus D_{train} zu, zu dem er den kleinsten euklidischen Abstand hat.

Die Daten kamen aus einer Hypothese (keine gewisse Wahrheit)

Was zeigt der Silhouetten Score (rechts): Nicht die richtige Clusteraufteilung

Problem: Funktioniert nur bei separaten Daten

entweder sehr gut oder sehr schlecht

Problem: Funktioniert nur bei separaten Daten

entweder sehr gut oder sehr schlecht

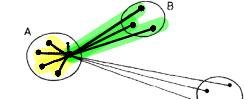
Problem: Funktioniert nur bei separaten Daten

entweder sehr gut oder sehr schlecht

Problem: Funktioniert nur bei separaten Daten

entweder sehr gut oder sehr schlecht

Silhouettenindex



Sei $i \in C_A$ ein Datenpunkt des Clusters A. Sei d_{ij} die (euklidische) Distanz zwischen Punkten i und j .

$a(i) := \frac{1}{|C_A|-1} \sum_{j \in C_A, j \neq i} d_{ij}$ mittlere Distanz zwischen i und allen anderen Punkten in Cluster A

weil wir nicht über d_i summieren

$b(i) := \min_{X \neq A} \frac{1}{|X|} \sum_{j \in X} d_{ij}$ mittlere Distanz zwischen i und allen anderen Punkten im benachbarten Cluster (hier: B)

$s(i) := \begin{cases} 1 - \frac{a(i)}{b(i)}, & \text{falls } a(i) < b(i) \\ 0, & \text{falls } a(i) = b(i) \\ 1 - \frac{b(i) - a(i)}{\max(a(i), b(i))}, & \text{falls } a(i) > b(i) \end{cases}$

i ist gut簇化: Unähnlichkeiten im Cluster kleiner als zum benachbarten Cluster

Uneindeutig: sollte i zum Cluster A oder B gehören?

i ist falsch簇化: liegt im Mittel näher zu Punkten in Cluster B als in Cluster A.

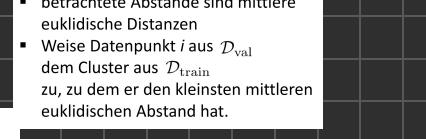
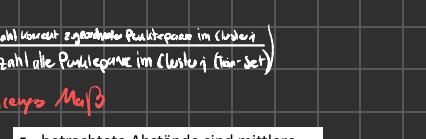
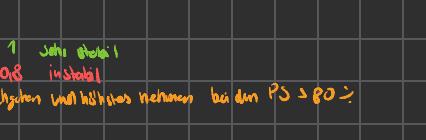
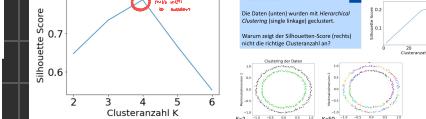
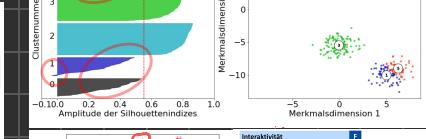
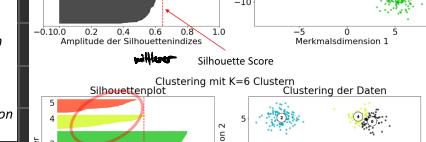
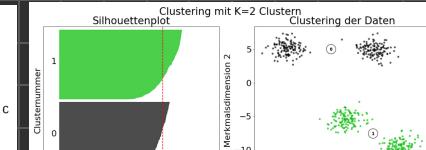
für $|C_A| > 1$, sonst 0.

Wertebereich: $-1 \leq s(i) \leq 1$

Mittlere Silhouettenindizes

... für Cluster X: $\bar{s}_X = \frac{1}{|C_X|} \sum_{i \in C_X} s(i)$ „average silhouette width“ bewertet einzelne Cluster

... für alle Daten: $\bar{s} = \frac{1}{N} \sum_{i=1}^N s(i)$ „silhouette score“ bewertet komplettes Clustering



EDA wissenschaftlich erforschen, Fragestellungen

Feature Engineering Zielgrößen

Tab

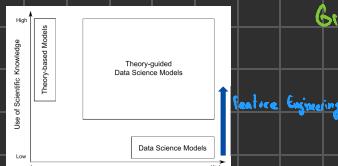
Shift-Tab - auf die Funktion

Feature Engineering

domainspezifisch, zeitaufwändig

Domänenwissen \Rightarrow Merkmale aus Daten erzeugen \Rightarrow Daten-getriebene Merkmale (Vorhersagen)

Erstellen einer mathematischen Beschreibung mithilfe von Daten



vs. Theoriebasierte Modellierung

Grundprinzipien \Rightarrow mathematische Beschreibung

Vorgehen

- Klassische Explorative Datenanalyse
- Domänenexperten aufsuchen
- selbst zu Domänenexperten werden
speziellisch

- Modelle ermöglichen Vorhersagen

Typen von Vorhersagen daten-getriebener Modelle:



Ergebnis der Vorhersage:

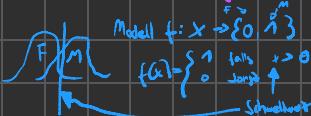
Modell wird auch genannt:

Klassifikator (classifier)

Klassifikation

- Binäre Klassifikation (zwei Klassen)

Schwellenbasierte Klassifikation

P: Anzahl der Datenpunkte der vorherzusagenden Klasse (Positiv)
(in unserem Beispiel: Klasse „Mann“)

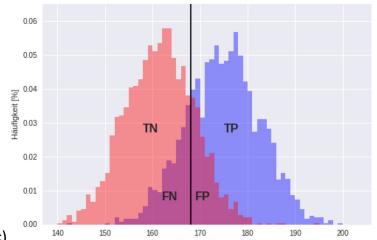
N: Anzahl Datenpunkte, die nicht der vorherzusagenden Klasse entsprechen (Negativ)

TP: Anzahl der korrekt vorhergesagten Positiven (True Positives)

TN: Anzahl der korrekt vorhergesagten Negativen (True Negatives)

FP: Anzahl der Falsch-Positiven (False Positives)

FN: Anzahl der Falsch-Negativen (False Negatives)



Gütemaß

• Accuracy $ACC = \frac{TP + TN}{TP + TN + FN + FP}$

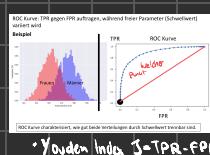
Problem: oft für neue (Class imbalance)
Modelle zu niedrige ACC

• Precision $PPV = \frac{TP}{TP + FP}$
Positive predictive value
braucht falsch-positiv

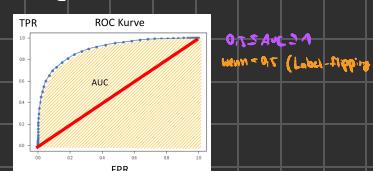
• Recall $TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$
True Positive Rate
braucht falsch-negativ

• F1-Score: harmonisches Mittel aus PPV & TPR
 $F_1 = \left(\frac{1}{PPV} + \frac{1}{TPR} \right)^{-1} = 2 \cdot \frac{PPV \cdot TPR}{PPV + TPR}$

• Receiver Operating Characteristic (ROC)
 True Positive Rate $TPR = \frac{TP}{P}$
 False Positive Rate $FPR = \frac{FP}{N}$



• AUC Area Under ROC Curve



Baseline Models

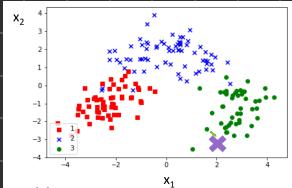
- einfache Modelle für Einschätzung (Baseline)
Als Vergleich komplexes Modell

Anwendungsfälle:

Was ist wichtiger?
Große Zahl Richtig-Positiv (TP)
Kleine Anzahl Fehler-Positiv (FP)
=> Abwägungstrag

Multiklassen-Klassifikation

• Nearest Neighbour NN



Ähnlichkeitsmaß

- Euklidische Distanz (Unähnlichkeitsmaß): $d(\mathbf{x}, \mathbf{x}') = \|\mathbf{x} - \mathbf{x}'\|$
- Kosinus-Ähnlichkeit: $\text{CosSim}(\mathbf{x}, \mathbf{x}') = \frac{\mathbf{x}^T \mathbf{x}'}{\|\mathbf{x}\| \|\mathbf{x}'\|} \in [-1, 1]$
Entspricht dem Kosinus des Winkels zwischen den Richtungsvektoren von \mathbf{x} und \mathbf{x}' :
 $O = \text{Ähnlich}$
 $A = \text{Unehnlich}$
- Jaccard-Koeffizient: $J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|} \in [0, 1]$
(zur Charakterisierung der Ähnlichkeit der Mengen S_1 und S_2)

Notation:

Trainingsset $D = (\mathbf{x}_1, y_1) \dots (\mathbf{x}_N, y_N)$

Distanzmaß $d(\mathbf{x}, \mathbf{x}')$ (eukl. $\|\mathbf{x} - \mathbf{x}'\|$)

$x_{[1]}$ nächster Nachbar

$$d(\mathbf{x}, \mathbf{x}_{[1]}) \leq d(\mathbf{x}, \mathbf{x}_{[2]}) \leq \dots \leq d(\mathbf{x}, \mathbf{x}_{[N]})$$

Nearest Neighbor (NN) Modell

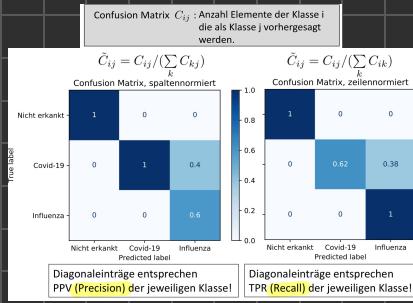
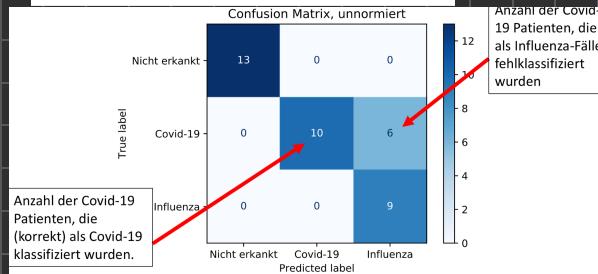
Das Modell lautet: $g(\mathbf{x}) = y_{[1]}(\mathbf{x})$

Auf Deutsch: Für einen beliebigen Datenpunkt \mathbf{x} , suche den zu \mathbf{x} nächsten Punkt aus dem Trainingset $(\mathbf{x}_{[1]})$ und gebe das Label $y_{[1]}$ dieses Punktes zurück.

Reduzierung aufpassen! NN verändert sich!

Beurteilung Klassifikator

- Top-K Accuracy (oft: K=5)
Vorhersage wird als True Positive TP gewertet, sofern die wahre Klasse unter den Top-K der wahrscheinlichsten Klasse liegt.
(Grenzfall K=1 entspricht der normalen Accuracy)
- One-vs-All Ansatz
Ermitteln des F1-Scores für jede Klasse (z.B. Covid-19 vs „Nicht-covid-19“)
Mittelwert über die F1-Scores aller Klassen bilden.
- Confusion Matrix (Wahrheitsmatrix) → auf der nächsten Folie



↳ Verallgemeinerung k-Nearst Neighbour (kNN)

Eine Verallgemeinerung des NN Modells ist kNN (k-Nächste Nachbarn):

Sei $k \geq 1$ eine ganze Zahl.

Signumfunktion (gibt das Vorzeichen des Arguments zurück)

k-Nearest Neighbor (kNN) Modell (binäre Klassifikation)

Das Modell lautet:

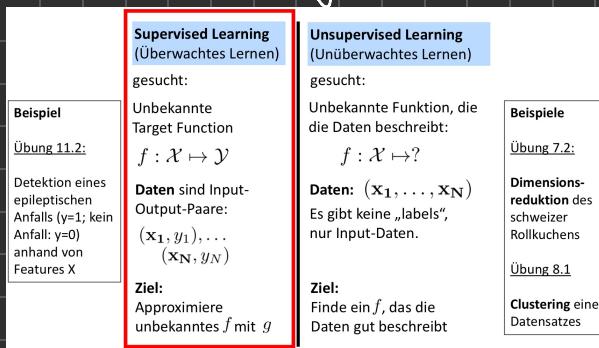
$$g(\mathbf{x}) = \text{sign} \left(\sum_{i=1}^k y_{[i]}(\mathbf{x}) \right); y \in \{-1, 1\}$$

$k=1 \rightarrow$ NN Modell (= 1NN Modell)

Falls dieser Ausdruck 0 ist, wird meist zufällig $g(\mathbf{x}) = 1$ oder -1 ausgegeben.

Auf Deutsch: Für einen beliebigen Datenpunkt \mathbf{x} , suche die k zu \mathbf{x} nächsten Punkte aus dem Trainingset und gebe das Label der Mehrheit dieser nächsten Punkte zurück.

Machine Learning Grundlagen



Menge der Kandidatenfunktionen (Hypothesen)

- Schwellwerte parametrisieren die Menge der möglichen schwellwertbasierten Modelle. Verschiedene Schwellwerte → verschiedene Modelle.

Lernalgorithmus:

- Mithilfe der ROC Statistik und des Youden-Index einen Schwellwert finden. Das dazugehörige Modell nennen wir *finale Hypothese* g .

Training- und Testset

Daten \mathcal{D} (N Datenpunkte)

Trainingsset $\mathcal{D}_{\text{train}}$ ($N-K$ Datenpunkte)

Testset $\mathcal{D}_{\text{test}}$ (K Datenpunkte)

Modell finden

- Teile Daten \mathcal{D} zufällig in Trainingsset und Testset auf (wir wollen keine Verzerrungen bei der Einteilung erzeugen)

In der Praxis: Typische Wahl ist $K = N/5$

- Erhalte das finale Modell auf dem Trainingsset: $g \in \mathcal{H}$

- Berechne den Vorhersagefehler¹ des Modells mithilfe von g auf dem Testset.

Daten-getriebene Modellierung | Vorgehen

- Sie haben Daten mit Features und Labels vorliegen.
Klassifikationsproblem: Labels sind diskret (Kategorien)
Regressionsproblem: Labels sind reellwertig (reelle Zahlen)
- Sie teilen die Daten in Trainings- und Testdaten auf.
- Sie finden ein Modell g (finale Hypothese) mithilfe der Trainingsdaten.
- Sie evaluieren den Fehler Ihres Modells auf den Testdaten, um die Qualität Ihres Modells für bisher „ungesehene Daten“ einschätzen zu können.

