Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen

Clusteranalyse

Tobias Scheffer
Thomas Vanck

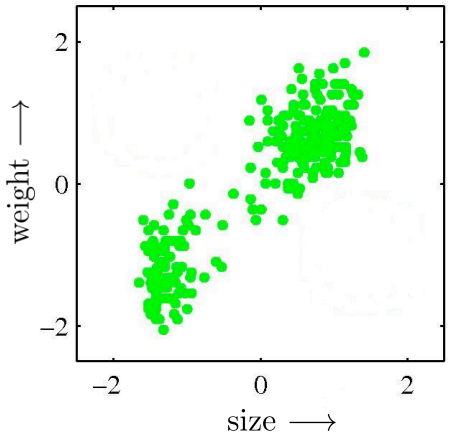
Überblick

- Problemstellung/Motivation
- Deterministischer Ansatz: K-Means
- Probabilistischer Ansatz: Gaußsche Mischmodelle
- Bayesscher Ansatz: Gaußsche Mischmodelle + Priors

Clusteranalyse: Was ist Clustern?

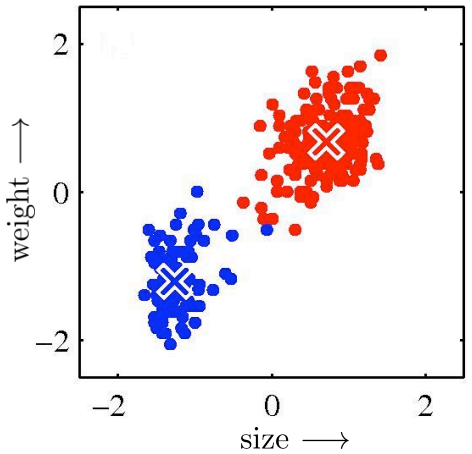
- Wir haben Datenpunkte $\mathbf{x}_1, ..., \mathbf{x}_N$
  - Z.B. $\mathbf{x}_n \in \mathbb{R}^D$

Beispiel $\mathbb{R}^2$, 272 Datenpunkte



- Wir wollen Einteilung der Datenpunkte in „Cluster“

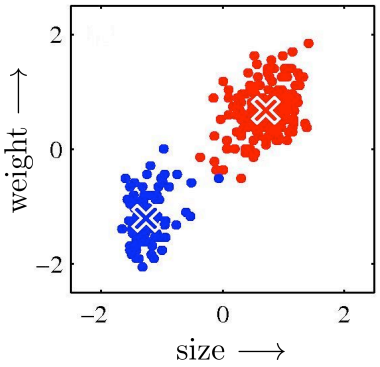
Jeder Punkt wird entweder Cluster 1
oder Cluster 2 zugewiesen
(im Allgemeinen: $K \geq 2$ Cluster)



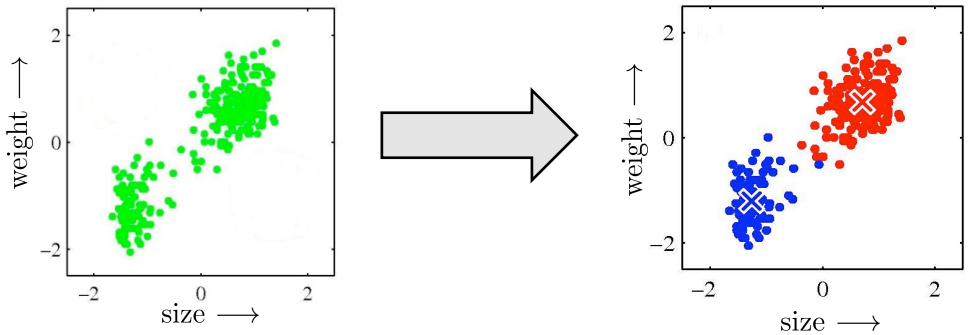
Clusteranalyse: Was ist Clustern?

- Annahme oft, dass Datenpunkte zu verschiedenen Klassen gehören

Beobachtungen: Haustiere
Klasse I: Katzen
Klasse II: Hunde

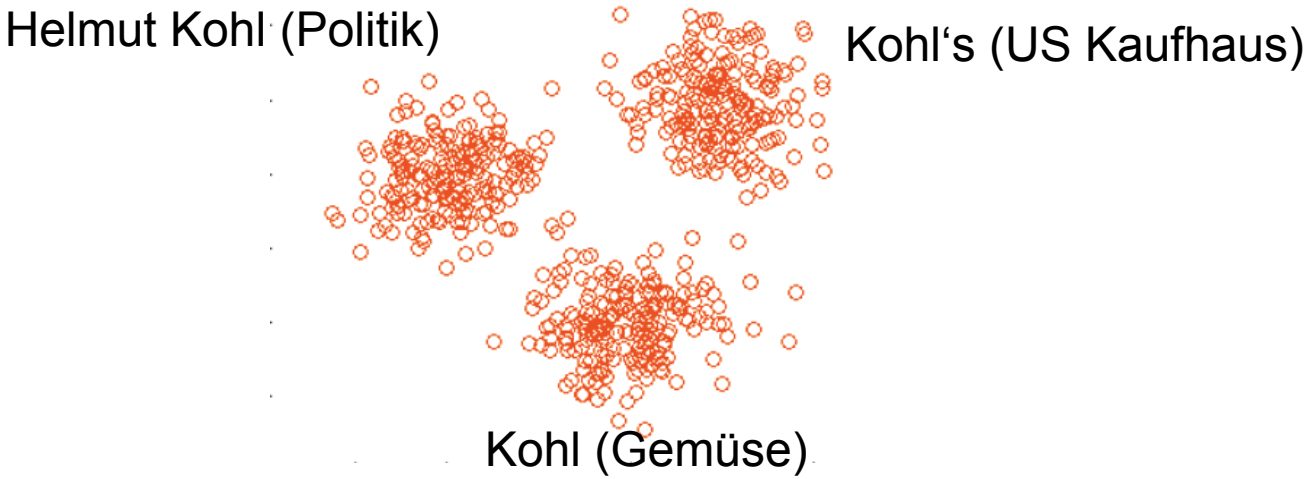


- …aber wir sehen keine Klassenlabels!
- Nicht-überwachtes Lernen: rekonstruiere Klassen ohne Labels



Clusteranalyse: Anwendungen

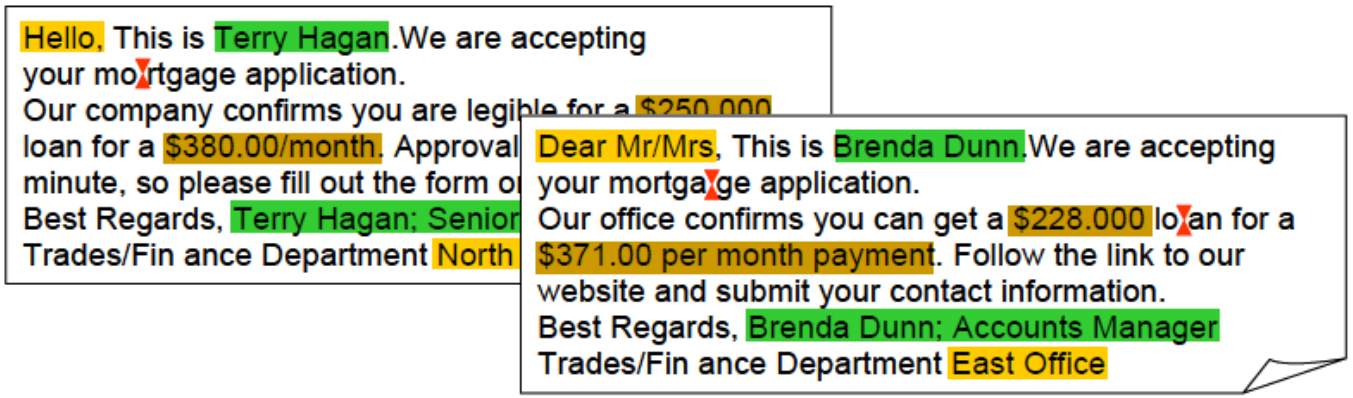
- Überblick über eine Dokumentenkollektion
  - Z.B. Suchmaschine: Suchwort „Kohl“
  - Liefert grosse Menge von Dokumenten



  - Idee: zeige dem Nutzer die Cluster, um genauere Auswahl des Themas zu ermöglichen

Clusteranalyse: Anwendungen

- Spam Kampagnen identifizieren
  - Spam-Kampagne: große Menge ähnlicher (aber nicht gleicher) e-mails



  - Eine Kampagne ist ein deutlicher Cluster ähnlicher e-mails
  - Attribute z.B. Worte die im e-mail Text auftauchen

Clusteranalyse: Anwendungen

- Erstellen der Wortklassen für n-gram Klassenmodelle
  - Erinnerung Vorlesung 3: n-gram Klassenmodell

$$p(w_n \mid w_{n-1}, ..., w_1) = p(w_n \mid c_n)p(c_n \mid c_{n-1}, ..., c_1)$$

„Wir sehen uns [Montag | Dienstag | Mittwoch | …] Nachmittag“

  - Wortklassen entsprechen Clustern
  - Attribute z.B. Auftreten der Worte in bestimmten Kontexten…

Überblick

- Problemstellung/Motivation
- Deterministischer Ansatz: K-Means
- Probabilistischer Ansatz: Gaußsches Mischmodell
- Bayesscher Ansatz: Gaußsches Mischmodell + Priors

Problemstellung Clustering (Deterministisch)

- Gegeben
  - Daten $\mathbf{x}_1, ..., \mathbf{x}_N$ mit $\mathbf{x}_n \in \mathbb{R}^D$
  - Anzahl K vermuteter Cluster

Andere Attributtypen (binär, nominal) möglich

Oft problematisch (woher wissen wir K?)

- Gesucht
  - Zuweisung der Daten zu Clustern $1, ..., K$

$$\mathbf{r}_n \in \{0,1\}^K \qquad r_{nk} = \begin{cases} 1: & \mathbf{x}_n \text{ in Cluster } k \\ 0: & sonst \end{cases}$$

z.B. $\mathbf{r}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$

  - Clusterzentren

$$\mu_1, ..., \mu_K \in \mathbb{R}^D$$

  - So dass „Abstand zwischen Punkten im selben Cluster **klein** und der Abstand zwischen Punkten in verschiedenen Clustern **groß** ist“

Problemstellung Clustering (Deterministisch)

- Gegeben
  - Daten $\mathbf{x}_1, ..., \mathbf{x}_N$ mit $\mathbf{x}_n \in \mathbb{R}^D$
  - Anzahl K vermuteter Cluster

Andere Attributtypen (binär, nominal) möglich

Oft problematisch (woher wissen wir K?)

- Gesucht
  - Zuweisung der Daten zu Clustern $1, ..., K$

$$\mathbf{r}_n \in \{0,1\}^K \qquad r_{nk} = \begin{cases} 1: & \mathbf{x}_n \text{ in Cluster } k \\ 0: & sonst \end{cases} \qquad \text{z.B. } \mathbf{r}_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

  - Clusterzentren

$$\mu_1, ..., \mu_K \in \mathbb{R}^K$$

  - So dass der quadratische Abstand zum Clusterzentrum minimiert wird:

$$\text{minimiere } J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2$$

K-Means Algorithmus

- Gleichzeitiges Min. über $\mu_1, ..., \mu_K$ und $\mathbf{r}_1, ..., \mathbf{r}_N$ schwierig

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2$$

- Iterativer Algorithmus: Abwechselnde Minimierung
  - Starte mit zufälligen $\mu_1, ..., \mu_K$
  - Update

$$\mathbf{r}_1^{neu}, ..., \mathbf{r}_N^{neu} = \arg\min_{\mathbf{r}_1,...,\mathbf{r}_N} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2 \qquad \text{„Expectation“}$$

$$\mu_1^{neu}, ..., \mu_K^{neu} = \arg\min_{\mu_1,...,\mu_N} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}^{neu} \|\mathbf{x}_n - \mu_k\|^2 \qquad \text{„Maximization“}$$

  - Iteriere bis Konvergenz

- Konvergenz sicher, weil *J* immer sinkt – aber im Allgemeinen nur lokales Optimum

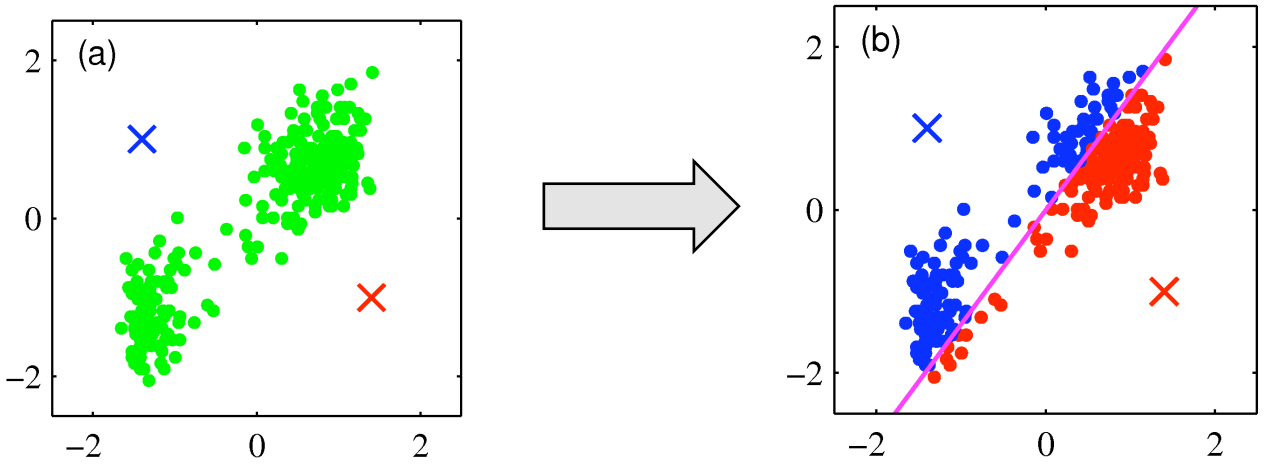
K-Means Algorithmus

- Expectation Schritt

$$\mathbf{r}_1^{neu}, ..., \mathbf{r}_N^{neu} = \underset{\mathbf{r}_1,...,\mathbf{r}_N}{\arg\min} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2$$

  - Einfach: ordne jeden Punkt dem ihm nächsten Cluster(zentrum) zu

$$r_{nk}^{neu} = \begin{cases} 1: & \text{if } k = \arg\min_j \|\mathbf{x}_n - \mu_j\| \\ 0: & sonst \end{cases}$$



K-Means Algorithmus

- Maximization Schritt

$$\mu_1^{neu}, ..., \mu_K^{neu} = \arg\min_{\mu_1,...,\mu_K} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}^{neu} \|\mathbf{x}_n - \mu_k\|^2$$

  - Ableitungen Null setzen

$$\begin{aligned}
\frac{\partial J}{\partial \mu_{kd}} &= \frac{\partial}{\partial \mu_{kd}} \sum_{k'=1}^{K} \sum_{n=1}^{N} r_{nk'}^{neu} \|\mathbf{x}_n - \mu_{k'}\|^2 \\
&= \sum_{n=1}^{N} r_{nk}^{neu} \frac{\partial}{\partial \mu_{kd}} \|\mathbf{x}_n - \mu_k\|^2 \\
&= \sum_{n=1}^{N} r_{nk}^{neu} \frac{\partial}{\partial \mu_{kd}} \sum_{d'=1}^{D} (x_{nd'} - \mu_{kd'})^2 \\
&= -2 \sum_{n=1}^{N} r_{nk}^{neu} (x_{nd} - \mu_{kd})
\end{aligned}$$

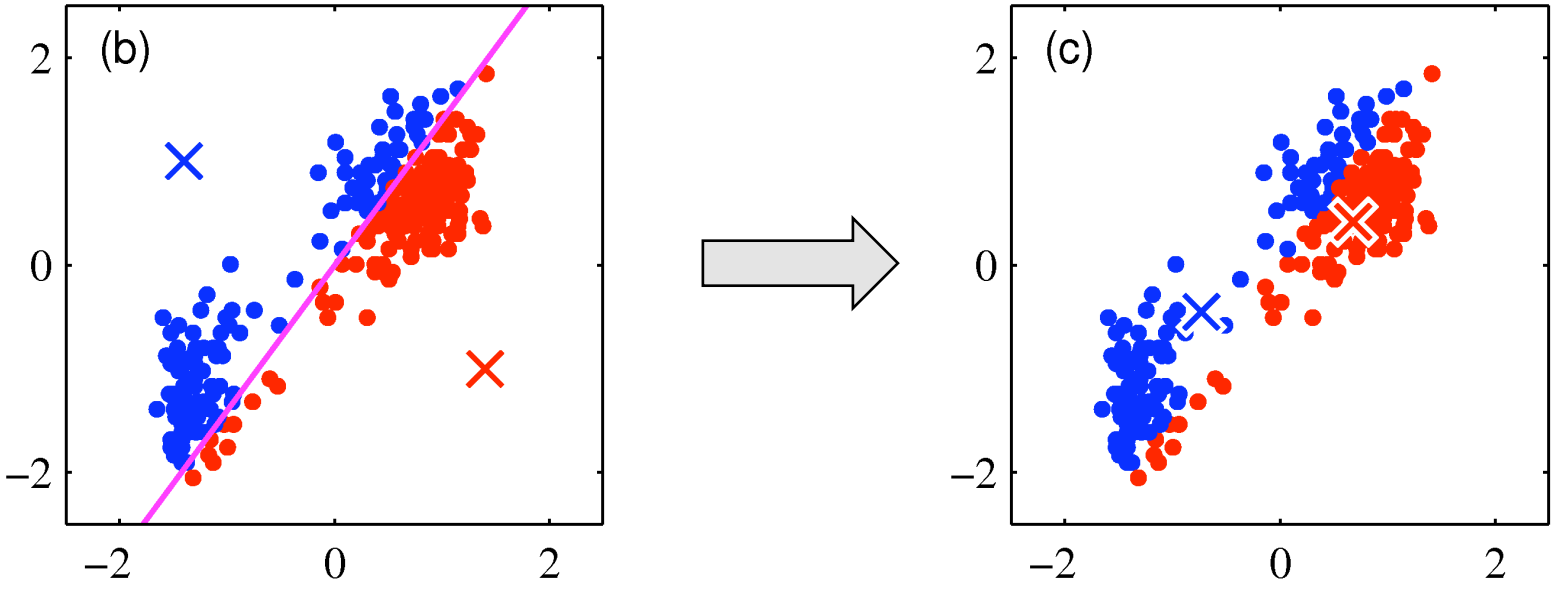
$$\forall k : 2 \sum_{n=1}^{N} r_{nk}^{neu} (\mathbf{x}_n - \mu_k) = 0 \quad \Rightarrow \quad \mu_k = \frac{\sum_n r_{nk}^{neu} \mathbf{x}_n}{\sum_n r_{nk}^{neu}}$$

K-Means Algorithmus

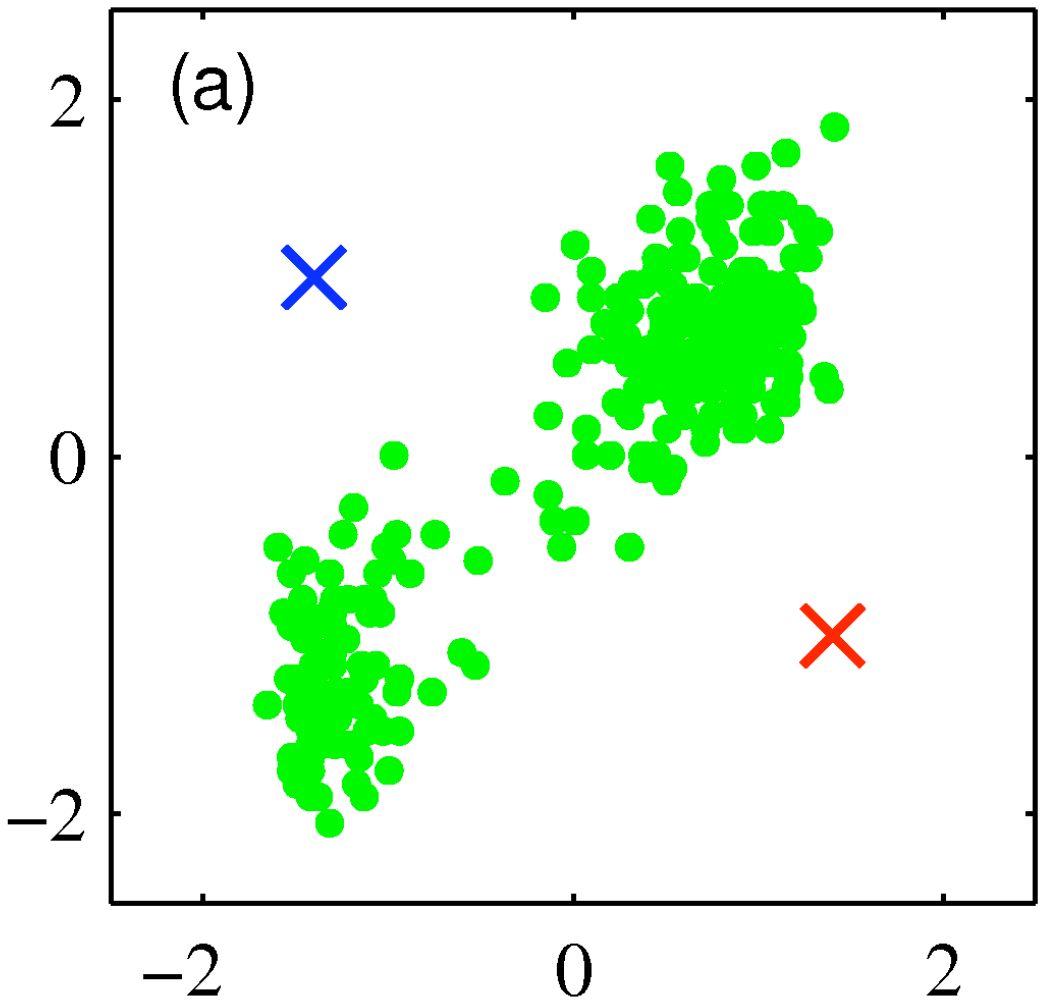
- Maximization Schritt

$$\mu_1^{neu}, ..., \mu_K^{neu} = \arg\min_{\mu_1,...,\mu_K} \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk}^{neu} \|\mathbf{x}_n - \mu_k\|^2$$

$$\mu_k = \frac{\sum_n r_{nk} \mathbf{x}_n}{\sum_n r_{nk}}$$

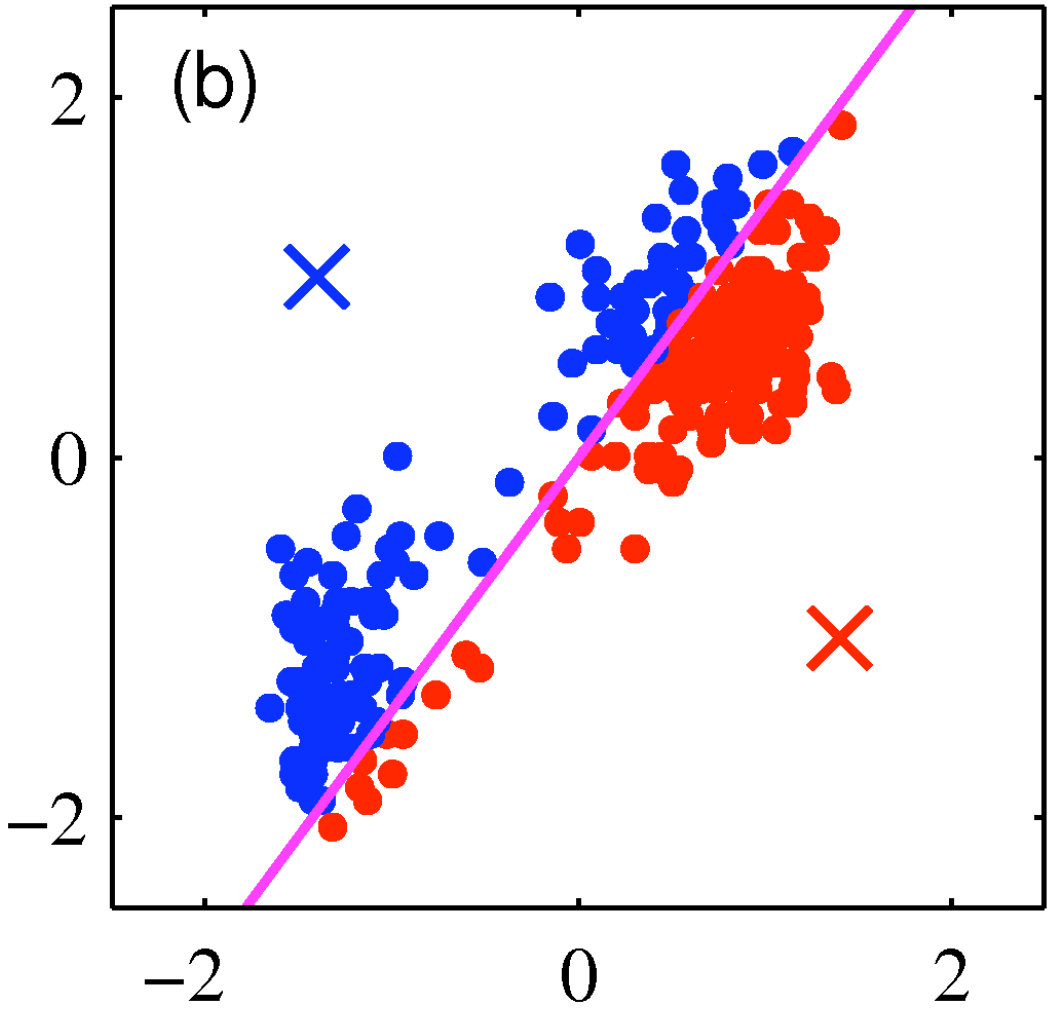


K-Means: Beispiel K = 2



Start:
Zufällige Initialisierung von μ_1, μ_2

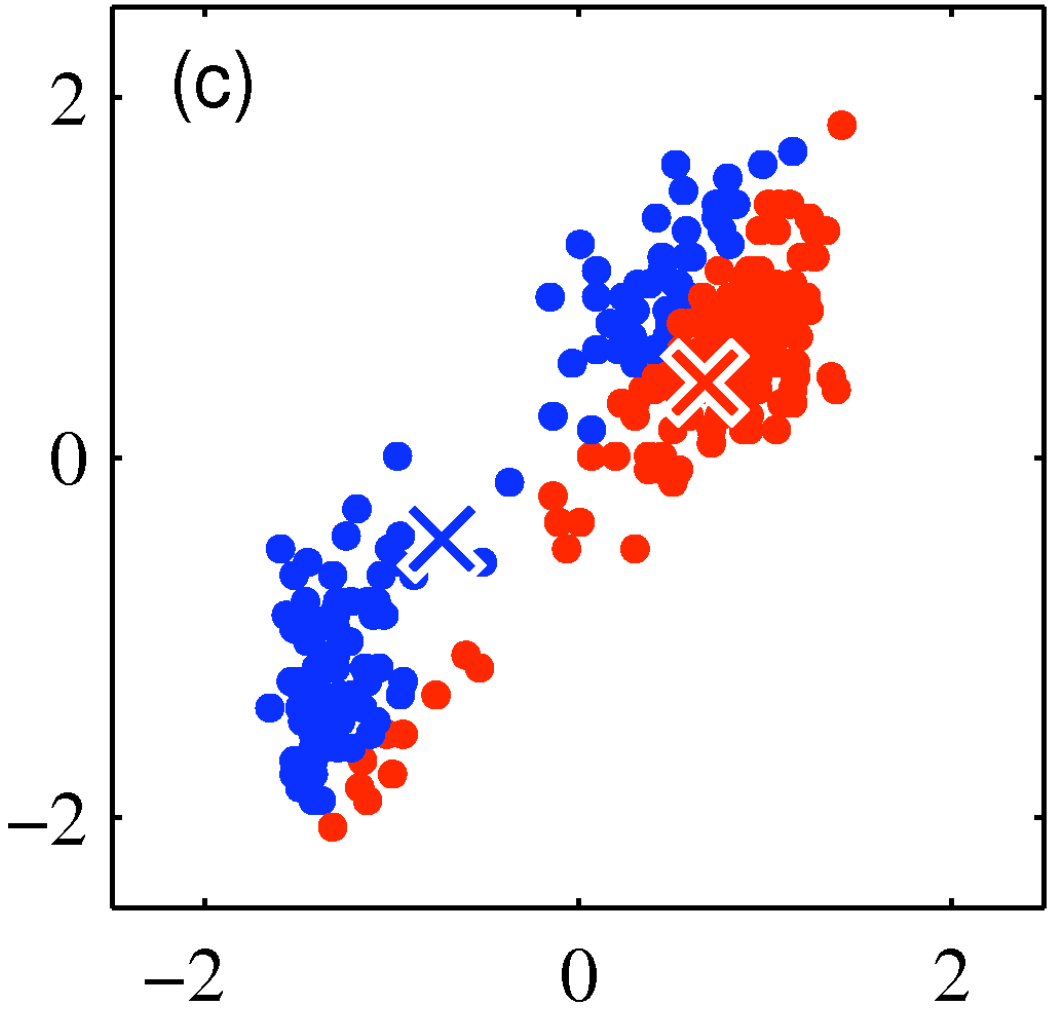
K-Means: Beispiel K = 2



Expectation:

$$r_{nk}^{neu} = \begin{cases} 1: & \text{if } k = \arg\min_j \|\mathbf{x}_n - \mu_j\| \\ 0: & sonst \end{cases}$$

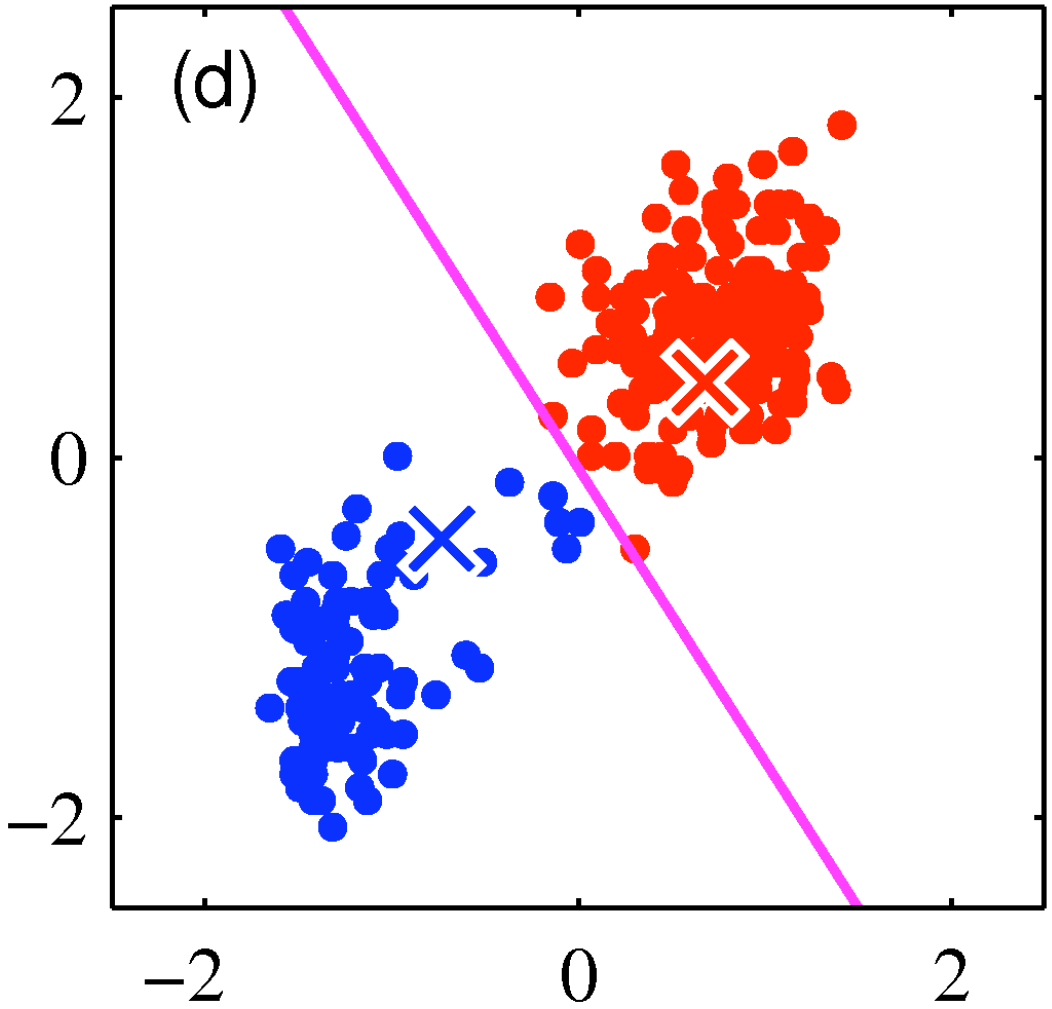
K-Means: Beispiel K = 2



Maximization:

$$\mu_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

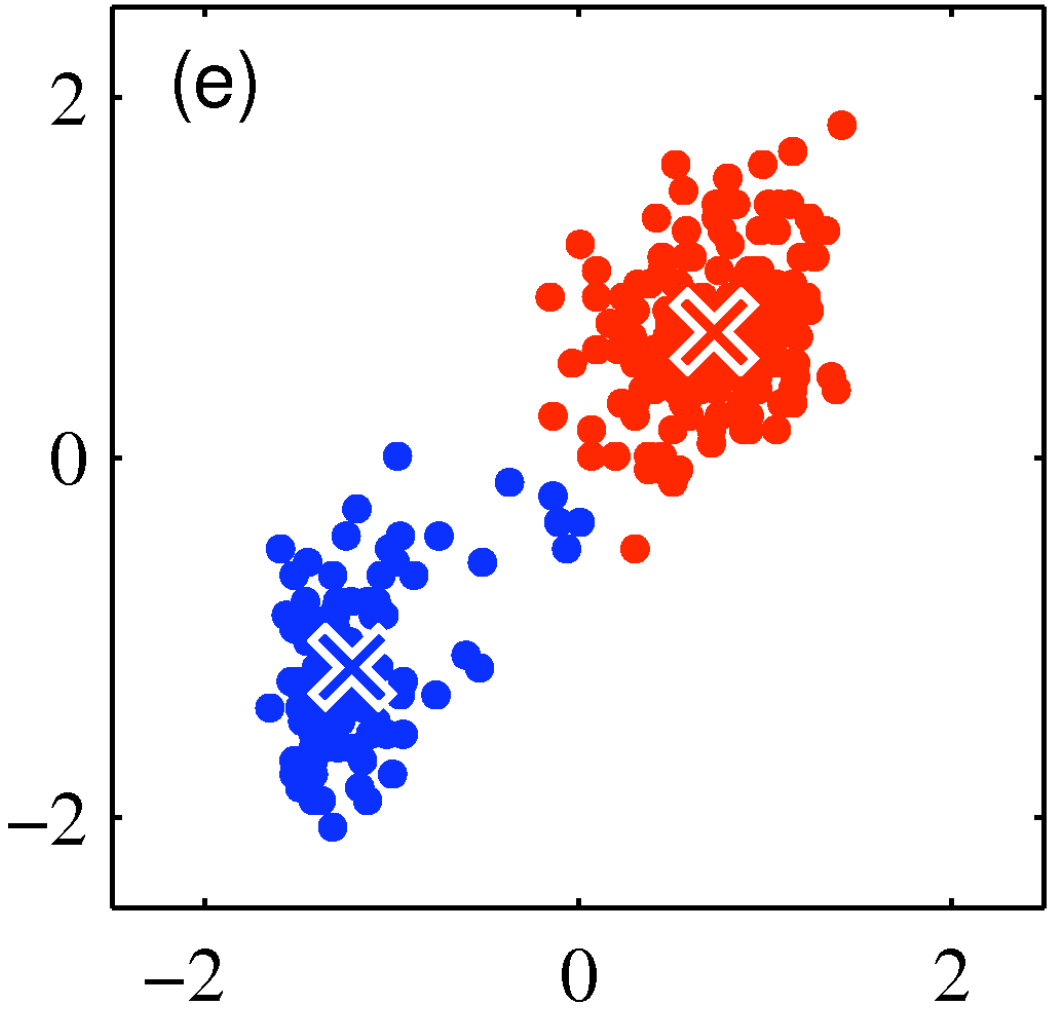
K-Means: Beispiel K = 2



Expectation:

$$r_{nk}^{neu} = \begin{cases} 1: & \text{if } k = \arg\min_j \|\mathbf{x}_n - \mu_j\| \\ 0: & sonst \end{cases}$$

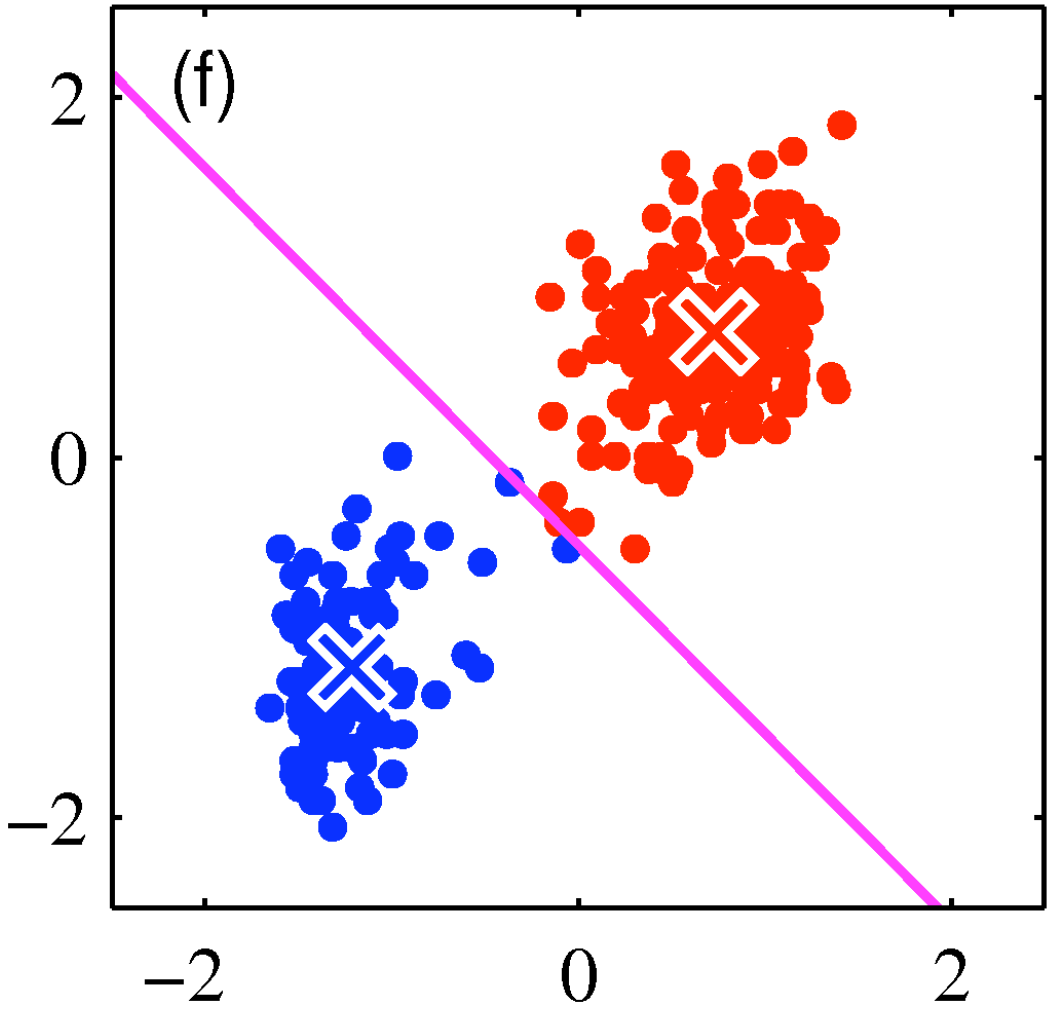
K-Means: Beispiel K = 2



Maximization:

$$\mu_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

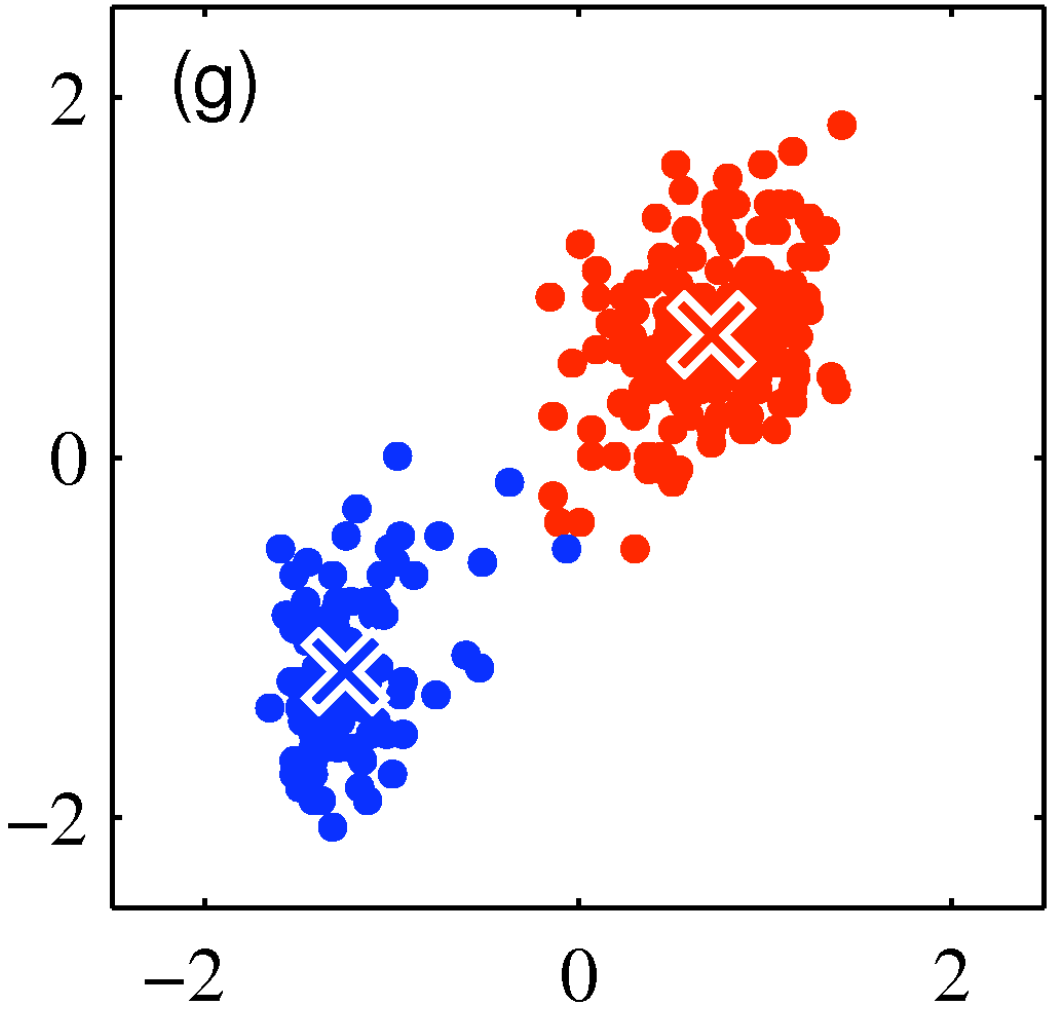
K-Means: Beispiel K = 2



Expectation:

$$r_{nk}^{neu} = \begin{cases} 1: & \text{if } k = \arg\min_j \|\mathbf{x}_n - \mu_j\| \\ 0: & sonst \end{cases}$$

K-Means: Beispiel K = 2



Maximization:

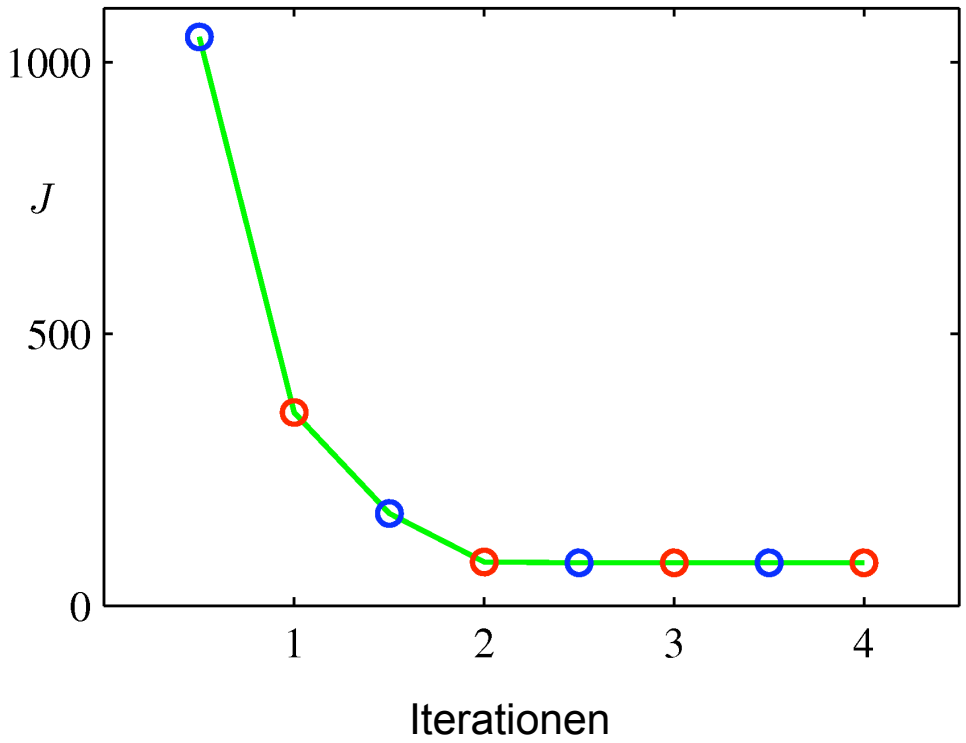
$$\mu_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

usw. (Konvergenz in nächster Iteration)

K-Means: Beispiel K = 2

- Kostenfunktion *J* fällt kontinuierlich

$$J = \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mu_k\|^2$$

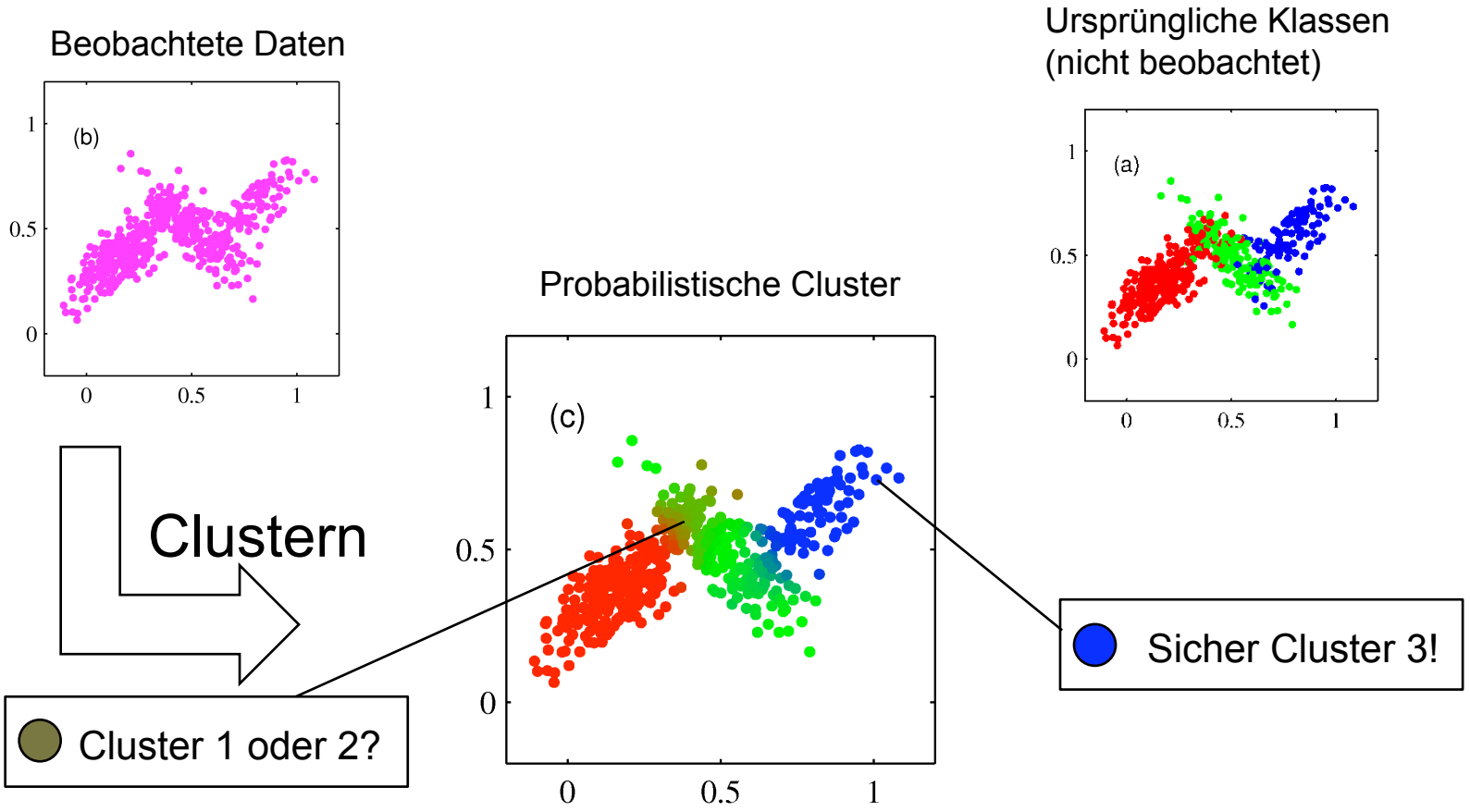


Kommentare K-Means

- ☺ Einfach zu implementieren
- ☺ Relativ schnell:
  - O(NK) per Iteration
  - Beschleunigung durch Datenstrukturen zur effizienten Berechnung des nächsten Clusterzentrums möglich
- ☹ Nur lokales Optimum garantiert
  - unterschiedliche Startwerte = unterschiedliche Lösungen
- ☹ Nicht probabilistisch
- ☹ Muss Anzahl Cluster vorgeben

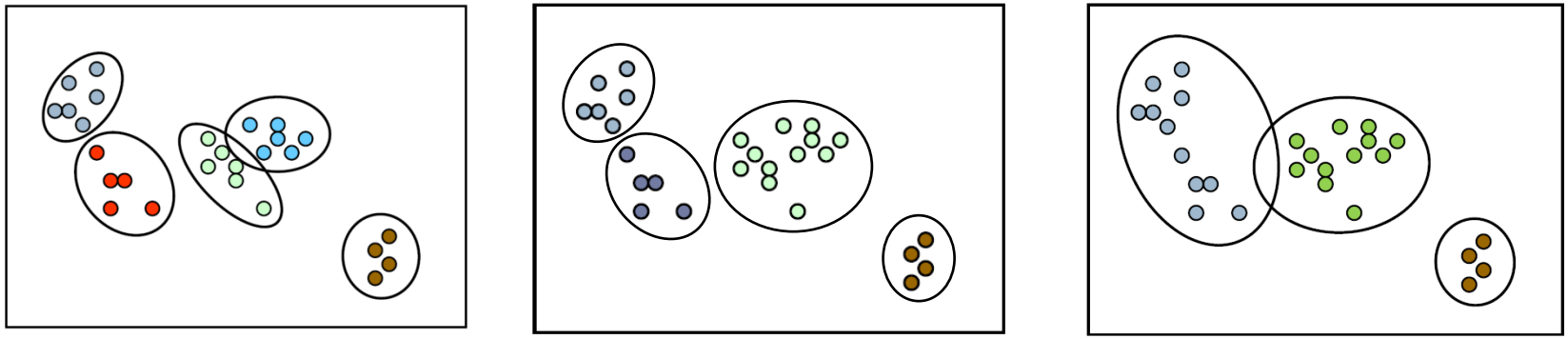
Probabilistisches Clustern besser

- Clustern sollte Konfidenz liefern: für einige Datenpunkte können wir keine sichere Entscheidung treffen!



Vorgegebene Anzahl von Clustern?

- Woher wissen wir, wie viele Cluster in Daten?
  - Manchmal klar aus der Anwendungsdomäne
  - Oftmals aber auch unklar



- Anzahl Cluster sollte vom Clustering Algorithmus (soweit möglich) mit bestimmt werden

Überblick

- Problemstellung/Motivation
- Deterministischer Ansatz: k-Means
- Probabilistischer Ansatz: Gaußsches Mischmodell
- Bayesscher Ansatz: Gaußsches Mischmodell + Priors

Probabilistisches Clustern mit Generativem Modell

- Idee: Generatives Modell, dass die Daten erzeugt haben könnte
- Clusterzugehörigkeit ist eine versteckte Variable in diesem Modell
- Modell hat Parameter Θ (Familie von Modellen)

- Clustering:
  - Gegeben die Daten
  - Suche Parameter Θ, so dass die Wahrscheinlichkeit dass Modell mit Parametern Θ Daten erzeugt hat:

$$\Theta^* = \arg\max_{\Theta} p(\mathbf{X} \mid \Theta)$$ „Likelihood“

Probabilistisches Clustern: Gaußsches Mischmodell

- Ersetze feste Clusterzuweisungen $\mathbf{r}_1, ..., \mathbf{r}_N$ durch entsprechende Zufallsvariablen $\mathbf{z}_1, ..., \mathbf{z}_N$
- Zufallsvariable Clusterzugehörigkeit

$$\mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ ... \\ z_K \end{pmatrix} \qquad z_k = \begin{cases} 1: & \mathbf{x} \text{ in Cluster } k \\ 0: & sonst \end{cases} \qquad \mathbf{z} = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}$$

- Generatives probabilistisches Modell
  - Wähle erst einen Cluster $k \in \{1, ..., K\}$
  - Generiere dann Datenpunkt aus Cluster *k*

Probabilistisches Clustern: Gaußsches Mischmodell

- Verteilung über Cluster: Mischgewichte $\pi_1, \ldots, \pi_K$

$p(z_k = 1) = \pi_k$ Wahrscheinlichkeit, Daten aus Cluster *k* zu sehen

$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$ Resultierende Verteilung über **z**

- Gaußverteilung der Daten gegeben Cluster

Gaußverteilung

Clusterzentrum

$$p(\mathbf{x} \mid z_k = 1) = \mathcal{N}(\mathbf{x} \mid \mu_k, \mathbf{\Sigma}_k)$$

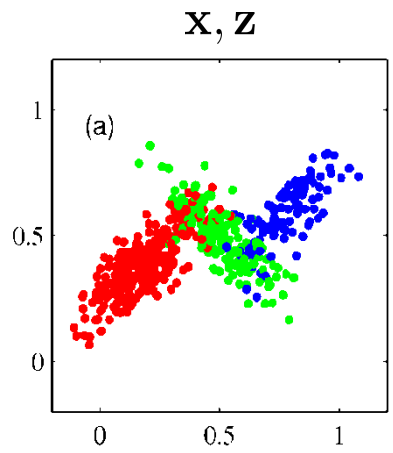
Clusterkovarianz

$$\mathcal{N}(\mathbf{x} \mid \mu, \mathbf{\Sigma}) = \frac{1}{Z} exp\left(-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma^{-1}} (\mathbf{x}-\mu)\right) \qquad Z = 2\pi^{D/2} \mid \mathbf{\Sigma} \mid^{1/2}$$

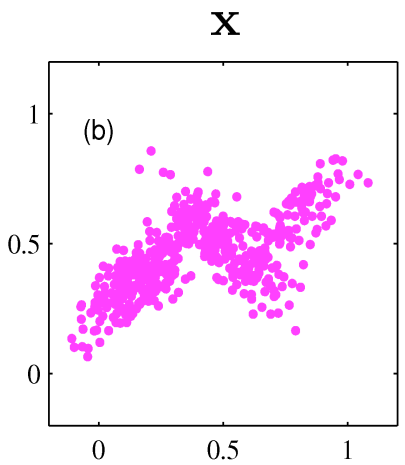
- Gesamtmodell: $p(\mathbf{x} \mid \pi, \mu, \mathbf{\Sigma}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \mathbf{\Sigma}_k)$

Beispiel Gaußsches Mischmodell

- Gaußsches Mischmodell,
  - K = 3, 500 Datenpunkte gezogen



ZV $\mathbf{z}$ nicht beobachtet



Mischgewichte

$$\pi_1 \approx 0.5$$
$$\pi_2 \approx 0.3$$
$$\pi_3 \approx 0.2$$

Clusterzentren

$$\mu_1 \approx \begin{pmatrix} 0.2 \\ 0.4 \end{pmatrix}$$
$$\mu_2 \approx \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}$$
$$\mu_3 \approx \begin{pmatrix} 0.8 \\ 0.6 \end{pmatrix}$$

Clusterkovarianzen

Σ_1
Σ_2
Σ_3

Geben an, wie die Punkte um das Clusterzentrum streuen

Multivariate Gaußverteilung

- Dichtefunktion

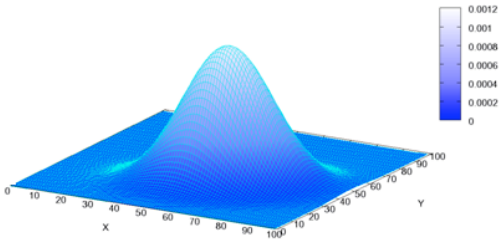
Kovarianz

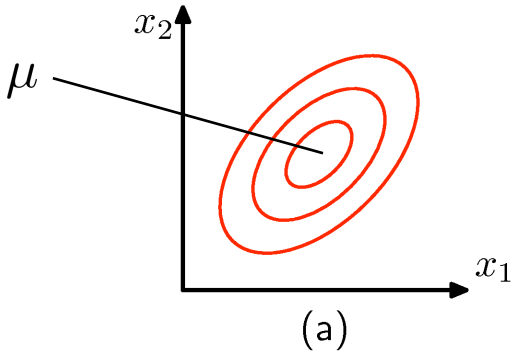
Mittelwert (Mean)

$$\mathcal{N}(\mathbf{x} \mid \mu, \mathbf{\Sigma}) = \frac{1}{Z} exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \mathbf{\Sigma^{-1}} (\mathbf{x}-\mu)\right)$$

$$Z = 2\pi^{D/2} \mid \mathbf{\Sigma} \mid^{1/2}$$

Beispiel D=2





Multivariate Gaußverteilung

- Dichtefunktion

Kovarianz

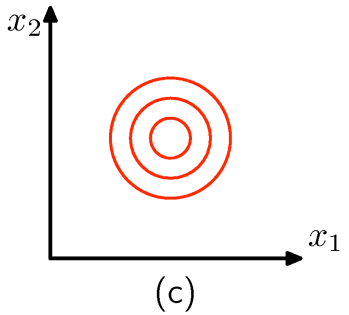
$$\mathcal{N}(x \mid \mu, \mathbf{\Sigma}) = \frac{1}{Z} exp\left(-\frac{1}{2}(x-\mu)^T \mathbf{\Sigma^{-1}}(\mathbf{x}-\mu)\right)$$

Mittelwert (Mean)

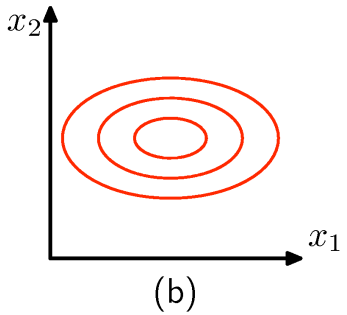
$$Z = 2\pi^{D/2} \mid \mathbf{\Sigma} \mid^{1/2}$$

- Kovarianzmatrix beschreibt wie Dichte um den Mittenwert streut

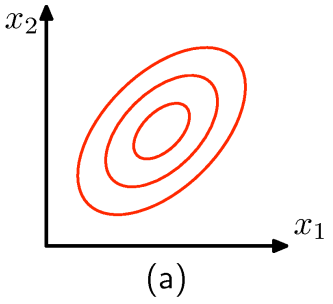
$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \qquad \mathbf{\Sigma} = \begin{pmatrix} 4 & 0 \\ 0 & 1 \end{pmatrix} \qquad \mathbf{\Sigma} = \begin{pmatrix} 2.5 & 1.5 \\ 1.5 & 2.5 \end{pmatrix}$$



(c)



(b)

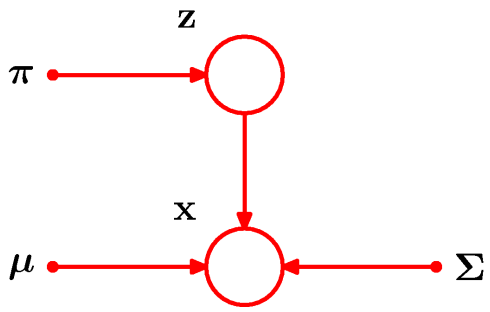


(a)

Gaußsches Mischmodell: Darstellung als Graphisches Modell

- Abhängigkeit zwischen Variablen $\mathbf{z}$ und $\mathbf{x}$ graphisch darstellbar

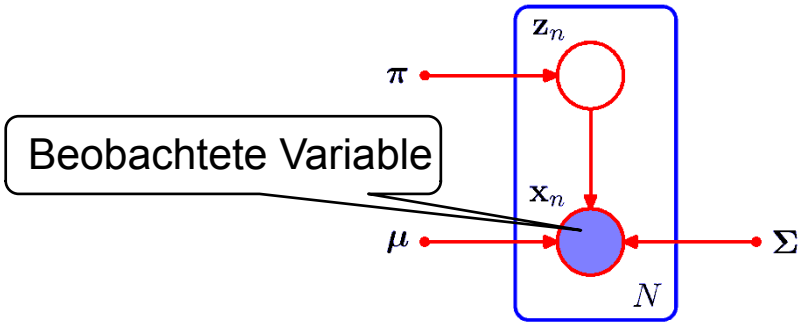
„graphisches Modell“



$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}$$

$$p(\mathbf{x} \mid z_k = 1) = \mathcal{N}(\mathbf{x} \mid \mu_k, \mathbf{\Sigma}_k)$$

- Gegeben Daten $\mathbf{x}_1, \ldots, \mathbf{x}_N$, haben wir auch N Zufallsvariablen $\mathbf{z}_1, \ldots, \mathbf{z}_N$ für Clusterzugehörigkeit



„Plate Notation“

Problemstellung Gaußsches Mischmodell Clustering (Maximum Likelihood)

- Gegeben Daten $\mathbf{X} = \mathbf{x}_1, ..., \mathbf{x}_N$
- Gesucht
  - Mischgewichte $\pi = \pi_1, ..., \pi_K$
  - Clusterzentren $\mu = \mu_1, ..., \mu_K$
  - Clusterkovarianzen $\mathbf{\Sigma} = \mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_K$

- So dass (logarithmische) Likelihood maximal:

$$\text{maximiere} \quad p(\mathbf{X} \mid \pi, \mu, \mathbf{\Sigma}) = \prod_{n=1}^{N} \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \mu_k, \mathbf{\Sigma}_k) \right)$$

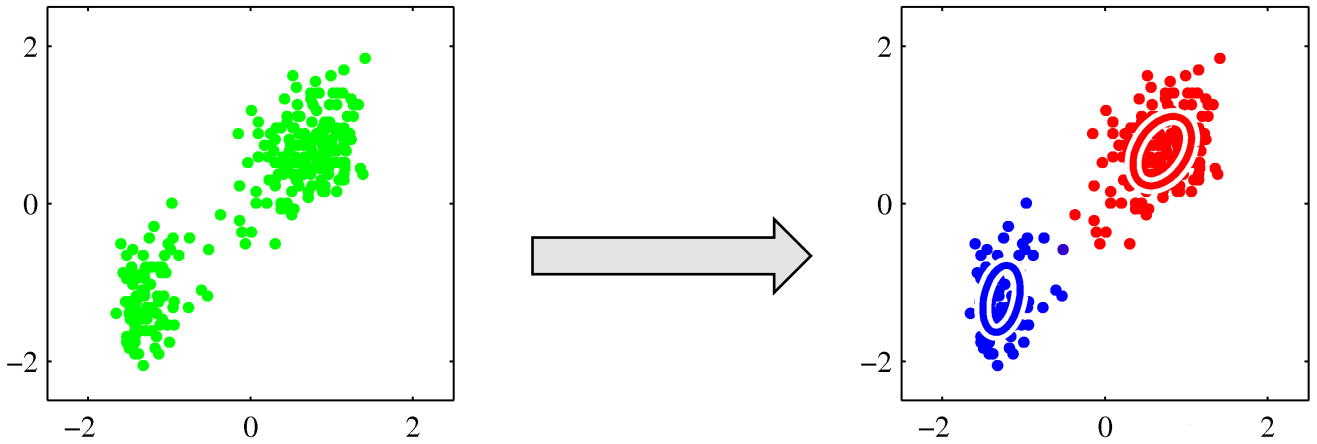
$$\text{maximiere} \log p(\mathbf{X} \mid \pi, \mu, \mathbf{\Sigma}) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \mu_k, \mathbf{\Sigma}_k) \right)$$

- Daraus lassen sich Clusterzugehörigkeiten berechnen

$$\gamma(z_{nk}) = p(z_{nk} = 1 \mid \mathbf{x}_n, \pi, \mu, \mathbf{\Sigma})$$

Problemstellung Gaußsches Mischmodell Clustering (Maximum Likelihood)

- Gegeben Daten $\mathbf{X} = \mathbf{x}_1, ..., \mathbf{x}_N$
- Gesucht
  - Mischgewichte $\pi = \pi_1, ..., \pi_K$
  - Clusterzentren $\mu = \mu_1, ..., \mu_K$
  - Clusterkovarianzen $\mathbf{\Sigma} = \mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_K$



EM Algorithmus für Maximum Likelihood

- Erinnerung (Vorlesung 4): Expectation-Maximization Algorithmus zur Maximierung der Likelihood

$$\text{Suche } \Theta^* = \arg\max_\Theta p(\mathbf{X} \mid \Theta)$$

- Daten bestehen aus ***X*** (sichtbar) und ***Z*** (versteckt, hier Clusterzugehörigkeit)

- Idee: Likelihood Maximierung wäre leicht, wenn wir ***X*** und ***Z*** hätten

$$\text{Suche } \Theta^* = \arg\max_{\boldsymbol{\Theta}} p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta})$$

Werte aller ZV in Modell beobachtet!

Erinnerung: EM Algorithmus

- Aber ***Z*** nicht beobachtet!
- Idee: maximiere Q-Funktion

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}_t) = \mathbb{E}_{\mathbf{Z}}[\log\, p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta}) \mid \mathbf{X}, \boldsymbol{\Theta}_t]$$

Parameterwert im letzten Schritt

$p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta}_t)$

- Iteriere
  - Expectation:

$$\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}_t) = \mathbb{E}[\log\, p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta}) \mid \mathbf{X}, \boldsymbol{\Theta}_t]$$

Berechnen als Funktion von $\boldsymbol{\Theta}$

  - Maximization:

$$\boldsymbol{\Theta}_{t+1} = \arg\max_{\boldsymbol{\Theta}} \mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}_t)$$

  - Theorem (Konvergenz): $p(\mathbf{X} \mid \boldsymbol{\Theta}_{t+1}) \geq p(\mathbf{X} \mid \boldsymbol{\Theta}_t)$
  - Allerdings nur lokales Maximum

EM für Gaußsches Mischmodell

- Likelihood für vollständige Daten ***X,Z***

$$
\begin{aligned}
p(\mathbf{X}, \mathbf{Z} \mid \mu, \mathbf{\Sigma}, \pi) &= p(\mathbf{Z} \mid \pi) p(\mathbf{X} \mid \mathbf{Z}, \mu, \mathbf{\Sigma}) \\
&= \prod_{n=1}^{N} \prod_{k=1}^{K} \pi_k^{z_{nk}} \mathcal{N}(\mathbf{x}_n \mid \mu_k, \mathbf{\Sigma}_k)^{z_{nk}}
\end{aligned}
$$

$$
\log\, p(\mathbf{X}, \mathbf{Z} \mid \mu, \mathbf{\Sigma}, \pi) = \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left(\log\, \pi_k + \log\, \mathcal{N}(\mathbf{x}_n \mid \mu_k, \mathbf{\Sigma}_k) \right)
$$

- … ***Z*** ist allerdings unbekannt!

EM für Gaußsches Mischmodell

- Q-Funktion für Gaußsches Mischmodell

$$
\begin{aligned}
\mathcal{Q}(\boldsymbol{\Theta}, \boldsymbol{\Theta}_t) &= \mathbb{E}[\log\, p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta}) \mid \mathbf{X}, \boldsymbol{\Theta}_t] \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta}_t) \log\, p(\mathbf{X}, \mathbf{Z} \mid \boldsymbol{\Theta}) \\
&= \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta}_t) \sum_{n=1}^{N} \sum_{k=1}^{K} z_{nk} \left(\log\, \pi_k + \log\, \mathcal{N}(\mathbf{x}_n \mid \mu_k, \boldsymbol{\Sigma}_k)\right) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \sum_{\mathbf{Z}} p(\mathbf{Z} \mid \mathbf{X}, \boldsymbol{\Theta}_t) z_{nk} \left(\log\, \pi_k + \log\, \mathcal{N}(\mathbf{x}_n \mid \mu_k, \boldsymbol{\Sigma}_k)\right) \\
&= \sum_{n=1}^{N} \sum_{k=1}^{K} \mathbb{E}[z_{nk}] \left(\log\, \pi_k + \log\, \mathcal{N}(\mathbf{x}_n \mid \mu_k, \boldsymbol{\Sigma}_k)\right)
\end{aligned}
$$

$$
\mathbb{E}[z_{nk}] = p(z_{nk} = 1 \mid \mathbf{x}_n, \boldsymbol{\Theta}_\mathbf{t})
$$

Q-Funktion = Likelihood, in der die z_{nk} durch ihre Erwartungswerte („Responsibilities“) ersetzt wurden! (Expectation Schritt)

EM für Gaußsches Mischmodell

- Expectation Schritt: Berechnung der „Responsibilities“

$$\mathbb{E}[z_{nk}] = p(z_{nk} = 1 \mid \mathbf{x}_n, \mathbf{\Theta_t})$$

Bayes Regel

$$= \frac{p(z_{nk} = 1 \mid \mathbf{\Theta}_t) p(\mathbf{x}_n \mid z_{nk} = 1, \mathbf{\Theta}_t)}{\sum_{j=1}^{K} p(z_{nj} = 1 \mid \mathbf{\Theta}_t) p(\mathbf{x} \mid z_{nj} = 1, \mathbf{\Theta}_t)}$$

$$= \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \mu_k, \mathbf{\Sigma}_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n \mid \mu_j, \mathbf{\Sigma}_j)} = \gamma(z_{nk})$$

EM für Gaußsches Mischmodell

- Maximization Schritt: maximiere

$$\mathcal{Q}(\mathbf{\Theta}, \mathbf{\Theta}_t) = \mathbb{E}[\log\, p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Theta}) \mid \mathbf{X}, \mathbf{\Theta}_t]$$

in $\mathbf{\Theta} = (\pi, \mu, \mathbf{\Sigma})$

- Ergebnis (ohne Beweis):
  - Wir definieren

$N_k := \sum_{n=1}^{N} \gamma(z_{nk})$ „Erwartete Anzahl von Punkten in Cluster *k*“

  - Dann

$\pi_k = \frac{N_k}{N}$ „Erwarteter Anteil von Punkten in Cluster *k*“

$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$ „Gewichteter Mittelwert für Cluster *k*“

EM für Gaußsches Mischmodell

- Maximization Schritt: maximiere

$$\mathcal{Q}(\mathbf{\Theta}, \mathbf{\Theta}_t) = \mathbb{E}[\log\, p(\mathbf{X}, \mathbf{Z} \mid \mathbf{\Theta}) \mid \mathbf{X}, \mathbf{\Theta}_t]$$

in $\mathbf{\Theta} = (\pi, \mu, \mathbf{\Sigma})$

- Ergebnis (ohne Beweis):
  - Definieren

$N_k := \sum_{n=1}^{N} \gamma(z_{nk})$ „Erwartete Anzahl von Punkten in Cluster *k*“

  - Dann

$$\mathbf{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nl})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

„Gewichtete Kovarianz für Cluster *k*“

Vergleich mit K-Means

- EM Zusammenfassung:
  - Starte mit zufälligen μ, Σ, π
  - Expectation: berechne „Responsibilities“

$$\gamma(z_{nk}) = p(z_{nk} = 1 \mid x_n)$$

„weiche“ Clusterzugehörigkeiten

  - Maximization:

$$\pi_k = \frac{N_k}{N}$$

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

„Weiche“ Berechnung der neuen Clusterzentren

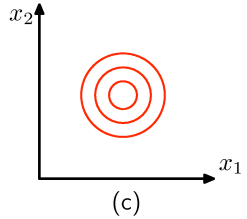
$$\mathbf{\Sigma}_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nl})(\mathbf{x}_n - \mu_k)(\mathbf{x}_n - \mu_k)^T$$

- Gaußsches Mischmodell + EM ≈ „Weicher“ K-Means

Vergleich mit K-Means (formal)

- Gaußsches Mischmodell mit festen Clusterkovarianzen

$$\mathbf{\Sigma}_k = \epsilon \mathbf{I} \qquad \mathbf{I} \text{ Einheitsmatrix}$$



  - Expectation:

$$\gamma(z_{nk}) = \frac{\pi_k exp(-\|\mathbf{x}_n - \mu_k\|^2/2\epsilon)}{\sum_j \pi_j exp(-\|\mathbf{x}_n - \mu\|^2/2\epsilon)}$$

$$\text{Für } \epsilon \to 0,\ \gamma(z_{nk}) \to 1 \text{ für } k = \arg\min_k \|\mathbf{x}_n - \mu_k\|^2$$

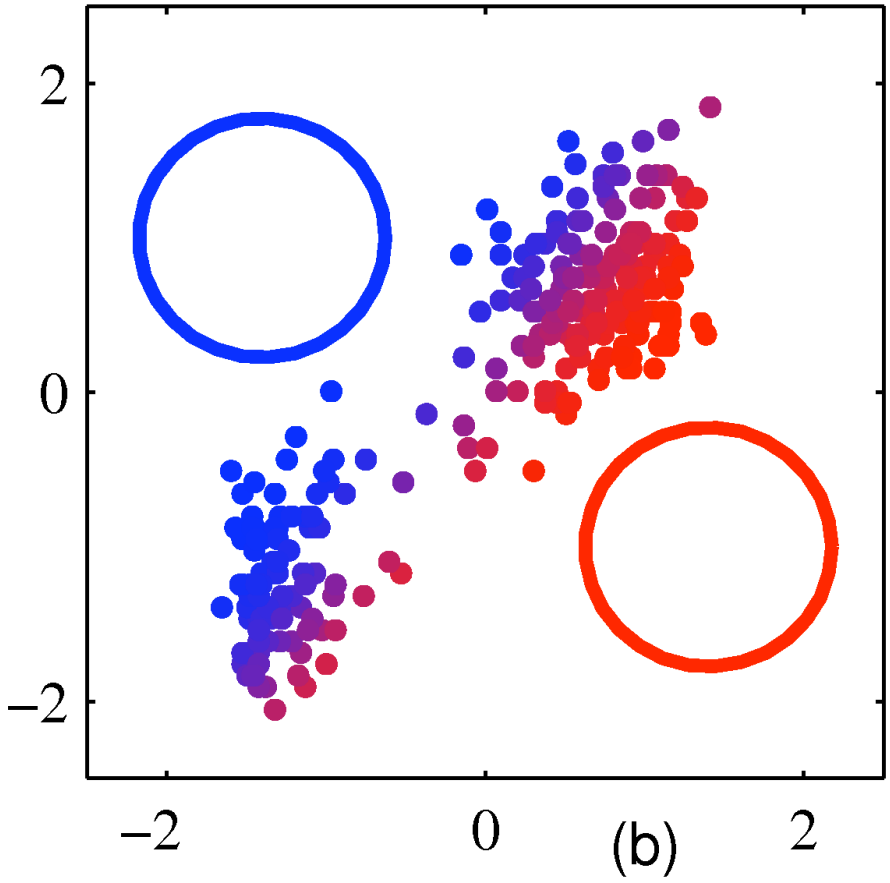
  - Maximization:

$$\mu_k = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

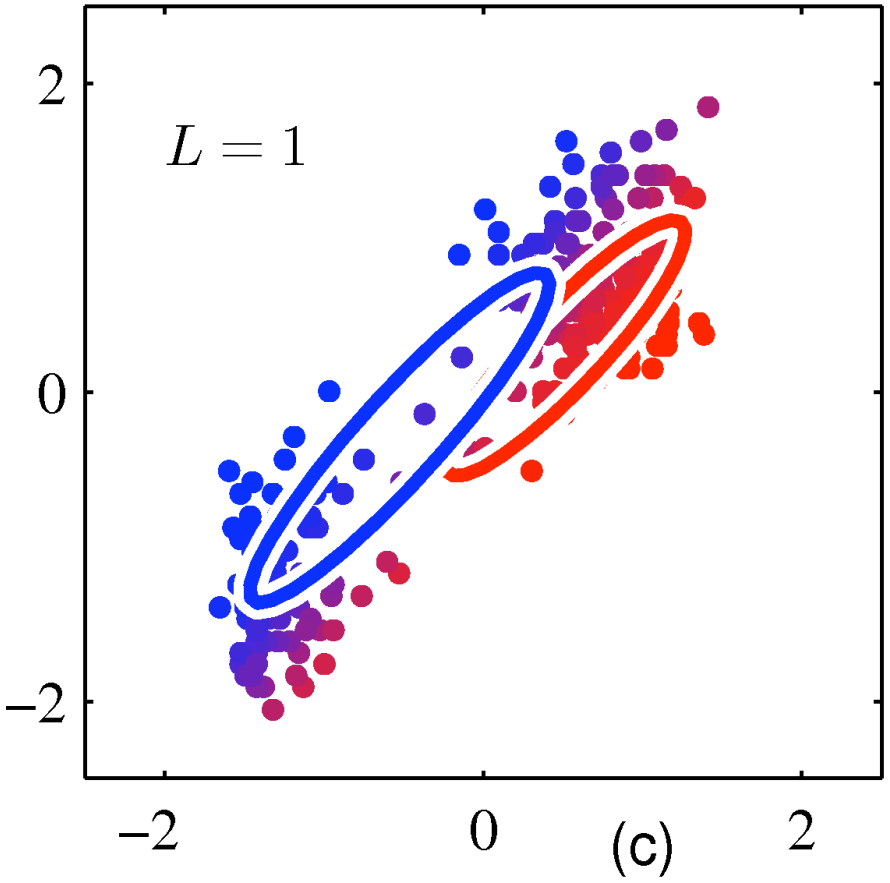
Für ε → 0 „harte“ Berechnung der neuen Clusterzentren

- Im Grenzfall ε → 0 wird Gaußsches Mischmodell zu K-Means

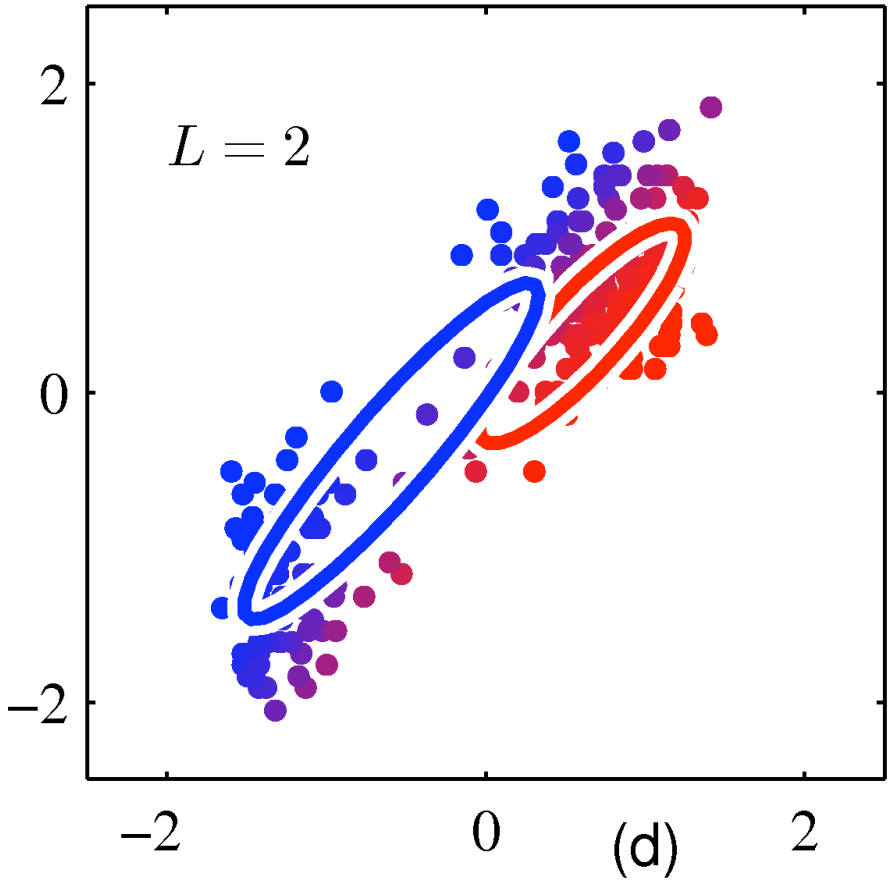
Beispiel Gaußsches Mischmodell Clustering



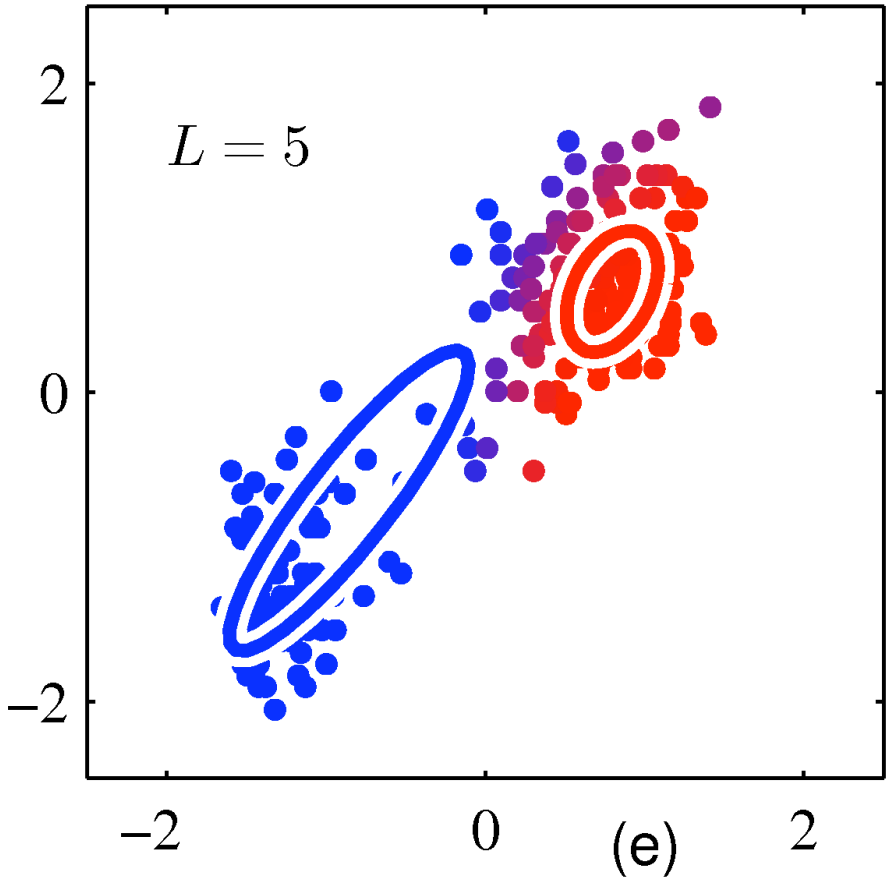
Beispiel Gaußsches Mischmodell Clustering



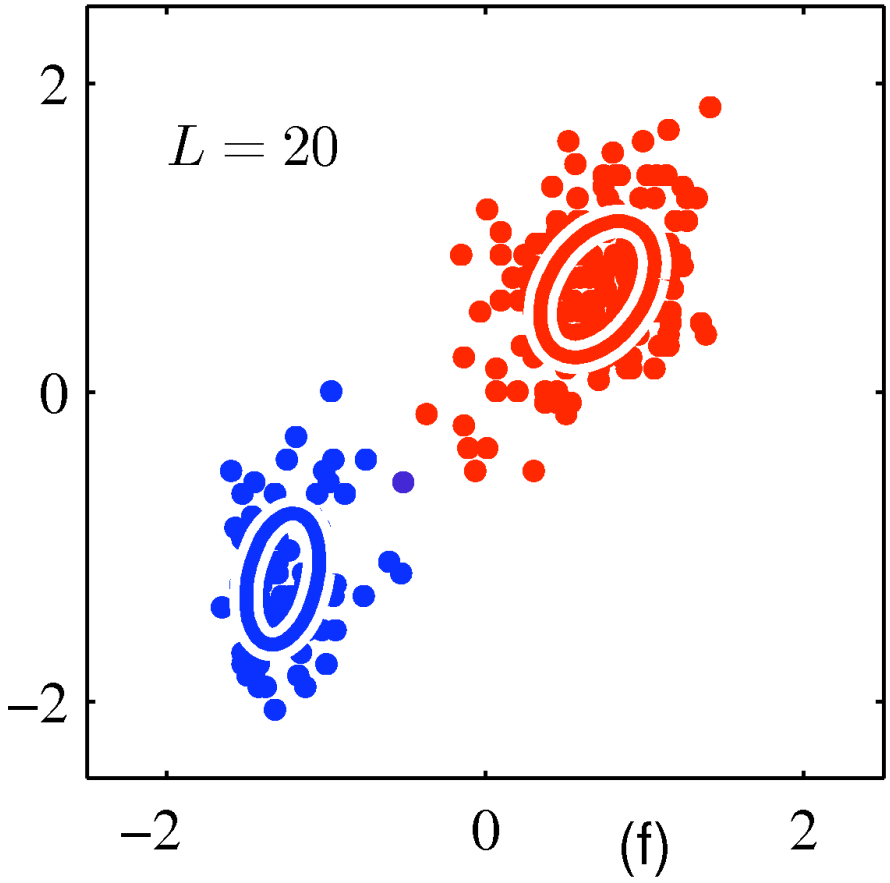
Beispiel Gaußsches Mischmodell Clustering



Beispiel Gaußsches Mischmodell Clustering



Beispiel Gaußsches Mischmodell Clustering



Problem I: Singularitäten

- EM maximiert Likelihood

$$p(\mathbf{X} \mid \pi, \mu, \mathbf{\Sigma}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \mathbf{\Sigma}_k)$$

- Problem: Singularität für

$$\mu_{\mathbf{k}} = \mathbf{x}_n, \quad \mathbf{x}_n \in \mathbf{X} \qquad \mathbf{\Sigma}_k \to \mathbf{0}$$

$$\mu_k = \mathbf{x}_n, \ \mathbf{\Sigma}_k = \epsilon^2 \mathbf{I}: \quad \mathcal{N}(\mathbf{x}_n \mid \mu_k, \epsilon^2 \mathbf{I}) = \frac{1}{(2\pi)^{1/2}} \frac{1}{\epsilon}$$

- Likelihood wird unendlich für $\mathbf{\Sigma}_k \to \mathbf{0}$
- „Overfitting“: Modell zu sehr an Daten angepasst
- In der Praxis: Während EM diesen Fall detektieren und entsprechende Clusterkomponente neu initialisieren

Problem II: Wie bestimmt man Anzahl der Cluster?

- Likelihood Funktion

$$p(\mathbf{X} \mid \pi, \mu, \mathbf{\Sigma}) = \prod_{n=1}^{N} \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \mu_k, \mathbf{\Sigma}_k)$$

- Je mehr Komponenten wir zulassen, desto größer wird

$$\arg\max_{\pi,\mu,\mathbf{\Sigma}} p(\mathbf{X} \mid \pi, \mu, \mathbf{\Sigma})$$

(bis Anzahl Komponenten = N)

- Modell mit N Clustern nutzlos!
- Likelihood kann nicht über Anzahl Cluster entscheiden
- Ebenfalls Overfitting Phänomen

Diskussion Gaußsches Mischmodell

- Probabilistisches Verfahren ☺
- Singularitäten ☹
- Anzahl Cluster muss vorgegeben werden ☹
- Problem ist der „Maximum Likelihood“ Ansatz
  - ML Ansatz erlaubt, Parameter zu sehr an den Datensatz anzupassen (Overfitting)
  - Lösung: Regularisierung durch Prior