

Universität Potsdam
Institut für Informatik
Lehrstuhl Maschinelles Lernen



Maschinelles Lernen II

Clustering 2

Christoph Sawade/Niels Landwehr
Tobias Scheffer

Überblick

- Zuletzt:
 - ◆ K-means
 - ◆ Mixture of Gaussians
- Hierarchisches Clustern
 - ◆ Bottom Up
 - ◆ Top Down
- Graphen-basiertes Clustern
 - ◆ Ähnlichkeitsgraph
 - ◆ Minimaler Schnitt

Clustern

- Geg.

- ◆ Objekte $V = \{x_1, \dots, x_n\}$
- ◆ Distanzfunktion $\text{dist}(x_i, x_j) \geq 0$ oder Ähnlichkeitsfunktion $w_{ij} = \text{sim}(x_i, x_j) \geq 0$
- ◆ Erwartete Clusteranzahl k

- Ziel: Partition P_1, \dots, P_k , wobei $P_i \cap P_j = \emptyset$, $\bigcup_{i=1 \dots k} P_i = V$ mit...
 - ◆ ... hoher intra-cluster-Ähnlichkeit
 - ◆ ... niedriger inter-cluster-Ähnlichkeit

Inter-Cluster Metriken

- Einfacher Abstand

$$d_{\min}(P_i, P_j) = \min_{v \in P_i, w \in P_j} \text{dist}(v, w)$$

- Kompletter Abstand

$$d_{\max}(P_i, P_j) = \max_{v \in P_i, w \in P_j} \text{dist}(v, w)$$

- Durchschnittsabstand

$$d_{\text{mean}}(P_i, P_j) = \frac{1}{|P_i| |P_j|} \sum_{v \in P_i} \sum_{w \in P_j} \text{dist}(v, w)$$

- Abstand der Zentroide

$$d_{\text{cent}}(P_i, P_j) = \text{dist} \left(\frac{1}{|P_i|} \sum_{v \in P_i} v, \frac{1}{|P_j|} \sum_{v \in P_j} v \right)$$

Optimales Clustering

- Berechnung des globalen Optimum bzgl. inter- und intra-cluster-Ähnlichkeit ist nicht effizient
 - ◆ Vgl. k-means:

$$\min_r \sum_{i=1}^n \sum_{j=1}^k r_{ij} \|x_i - \mu_j\|^2$$

- Bestimmung eines lokalen Optimums
 - ◆ EM-Algorithmus (siehe letzte VL)
 - ◆ Heuristik (Hierarchisches Clustering)
 - ◆ Relaxation (Spectral Clustering)

Überblick

- Hierarchisches Clustern
 - ◆ Bottom Up
 - ◆ Top Down
- Graphen-basiertes Clustern
 - ◆ Ähnlichkeitsgraph
 - ◆ Minimaler Schnitt

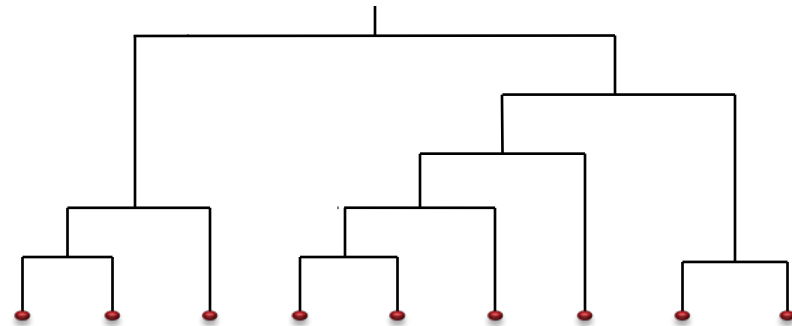
Hierarchisches Clustern

- Bottom Up
 - ◆ Agglomerative Nesting (Agnes)
- Top Down
 - ◆ Divisive Analysis (Diana)

Hierarchisches Clustern

Agnes (Algorithmus)

- Geg.: Objekte V , Inter-Cluster Metrik d
- Setze $C_0 = \{\{x\} \mid \forall x \in V\}$
- Solange unterschiedliche Cluster existieren
 - ◆ berechne min. Distanz über alle $c^v, c^w \in C_{i-1}$
 $(s, t) = \arg \min_{v, w} d(c^v, c^w)$; $D_i = \min_{v, w} d(c^v, c^w)$
 - ◆ Setze $C_i = \{c^v \mid \forall v \neq s, t\} \cup \{c^s \cup c^t\}$
- Liefere C_0, C_1, \dots zurück



Hierarchisches Clustern

Agglomerative Coefficient

- Sei $m_k = d(c^s, \{x_k\})$, wobei c^s das Cluster ist, mit dem x_k im i -ten Schritt verschmolzen wurde

$$C_i = \{c^v \mid \forall v \neq s, t\} \cup \{c^s \cup \{x_k\}\}$$

- Agglomerative Coefficient :

$$AC = \frac{1}{n} \sum_{i=1}^n \left(1 - \frac{m_i}{D_{\text{final}}} \right) \in [0, 1]$$

- Ein Maß für die Qualität eines Clusterings
- Nicht geeignet um Datensätze unterschiedlicher Größe zu vergleichen

Überblick

- Hierarchisches Clustern
 - ◆ Bottom Up
 - ◆ Top Down
- Graphen-basiertes Clustern
 - ◆ Ähnlichkeitsgraph
 - ◆ Minimaler Schnitt

Hierarchisches Clustern

Diana

- Bottom up: alle $\binom{n}{2}$ möglichen Fusionen werden betrachtet
- Top down: $2^{n-1} - 1$ mögliche Splits

Hierarchisches Clustern

Diana (Algorithmus)

- Geg.: Objekte V , Inter-Cluster Metrik d
- Setze $C_0 = \{V\}$
- Solange mehr-elementige Cluster existieren
 - ◆ Bestimme Cluster mit höchstem Durchmesser
$$c = \arg \max_{c \in C_{i-1}} \max_{s, t \in c} d(s, t)$$
 - ◆ Bestimme unähnlichstes Element
$$s = \arg \max_{v \in c} d(v, c \setminus \{v\})$$
 und setze $\bar{c} = \{s\}$
 - ◆ Solange $\max_{v \in c \setminus \bar{c}} D(v) > 0$, wobei $D(v) = d(v, c \setminus \bar{c}) - d(v, \bar{c})$
 - ★ $t = \arg \max_{v \in c \setminus \bar{c}} D(v)$
 - ★ $\bar{c} = \bar{c} \cup \{t\}$
 - ◆ Setze $C_i = (C_{i-1} \setminus \{c\}) \cup \{c \setminus \bar{c}\} \cup \{\bar{c}\}$
- Liefere C_0, C_1, \dots zurück

Hierarchisches Clustern

Divisive Coefficient

- Sei dia_i , der Durchmesser des Cluster aus dem das Objekt v_i zu letzt herausgelöst wurde (bis es einzeln war)

- Divisive Coefficient:

$$DC = \frac{1}{n} \sum_{i=1}^n dia_i$$

- Ein Maß für die Qualität eines Clusterings
- Nicht geeignet um Datensätze unterschiedlicher Größe zu vergleichen

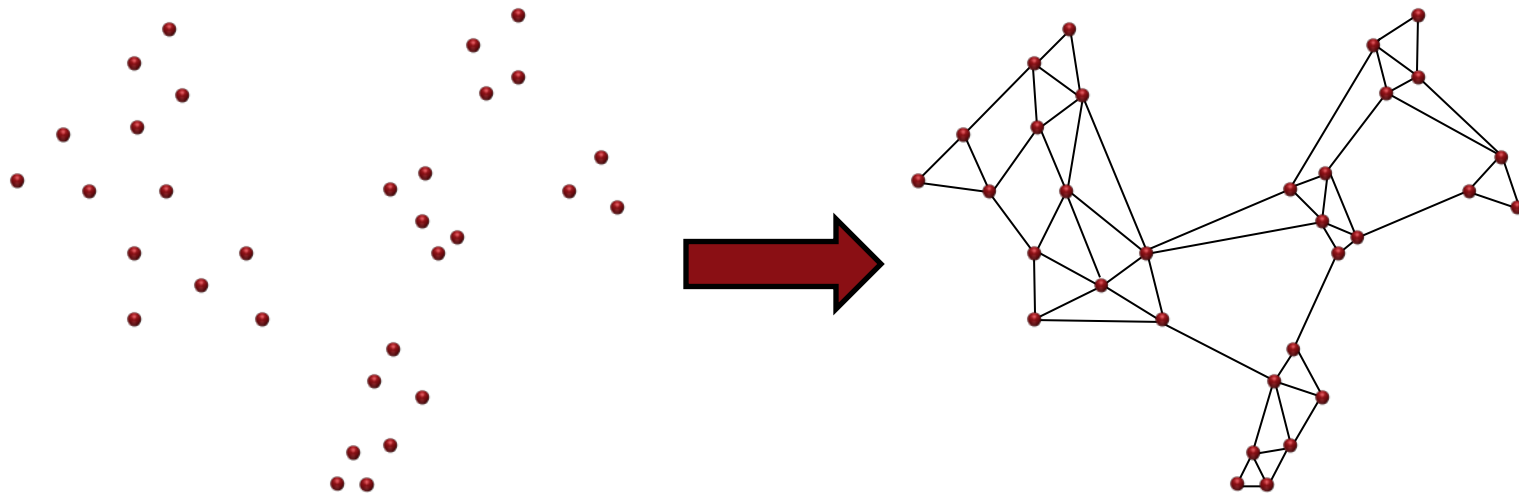
Überblick

- Hierarchisches Clustern
 - ◆ Bottom Up
 - ◆ Top Down
- Graphen-basiertes Clustern
 - ◆ Ähnlichkeitsgraph
 - ◆ Minimaler Schnitt

Graphen-basiertes Clustern

Ähnlichkeitsgraph

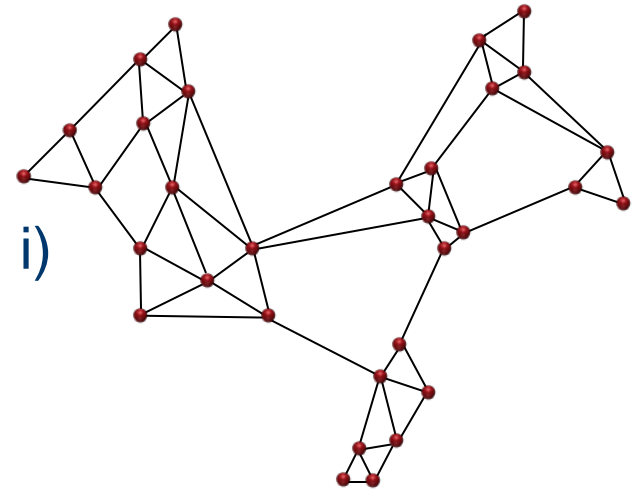
- Ähnlichkeit zwischen Datenpunkten v (Knoten) bilden gewichtete Kanten:



Graphen-basiertes Clustern

Ähnlichkeitsgraph

- Ähnlichkeit zwischen Datenpunkten v (Knoten) bilden gewichtete Kanten:
 - ◆ Vollständiger Graph: Kantengewichte = Ähnlichkeit
 - ◆ knn-Graph: Kante, wenn Knoten i (oder j) einer der k nächsten Nachbarn von j (bzw. i)
 - ◆ ε -Nachbarschaftsgraph: Kante, wenn $\text{dist}(v_i, v_j) < \varepsilon$



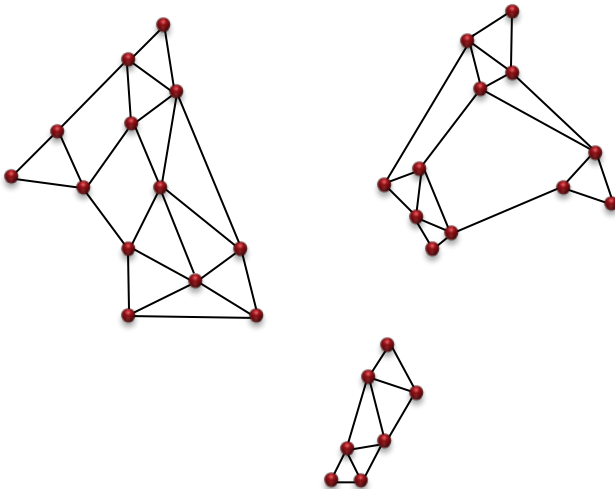
Graphen-basiertes Clustern

Definitionen

- Gewichtete Adjazenzmatrix $\mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & \cdots & \mathbf{W}_{1n} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{1n} & \cdots & \mathbf{W}_{nn} \end{pmatrix}$
- Knotengrad-Matrix $\mathbf{D} = \begin{pmatrix} d_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_n \end{pmatrix}$ $d_i = \sum_{j=1}^n w_{ij}$
- Laplace-Matrix
 - ◆ unnormalisiert $\mathbf{L}_{\text{un}} = \mathbf{D} - \mathbf{W}$
 - ◆ Symmetrisch normalisiert $\mathbf{L}_{\text{sym}} = \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{W} \mathbf{D}^{-1/2}$

Beobachtung

- Zusammenhängende Teilgraphen...
 - ◆ entspricht Anzahl Eigenwerte von \mathbf{L} mit Wert 0.
 - ◆ Zugehörige (unnormierte) Eigenvektoren enthalten Indikatorvektoren der Teilgraphen.
 - ◆ Erkenntnis für schwach zusammenhäng. Teilgraphen?



$$\lambda_1 = \lambda_2 = \lambda_3 = 0$$

$$u_1 = (1, \dots, 1, 0, \dots, 0, 0, \dots, 0)$$

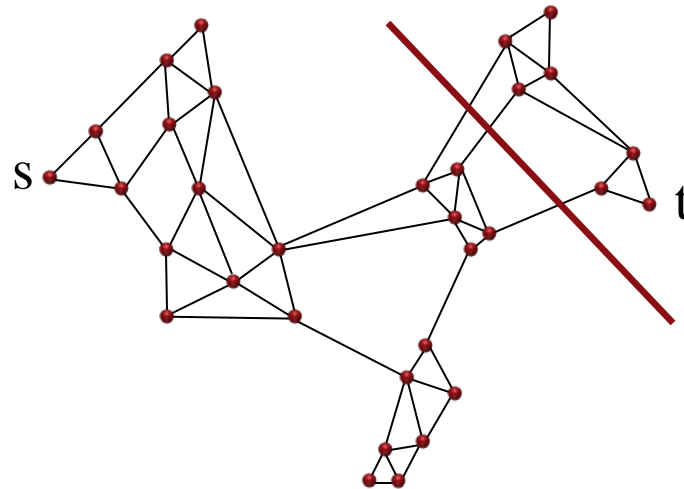
$$u_2 = (0, \dots, 0, 1, \dots, 1, 0, \dots, 0)$$

$$u_3 = (0, \dots, 0, 0, \dots, 0, 1, \dots, 1)$$

Minimaler Schnitt

Spezialfall $k=2$

- Betrachten Ähnlichkeitsgraphen mit zwei unterschiedlichen ausgezeichneten Knoten $s, t \in V$



- Ein s - t -Schnitt ist eine Partitionierung der Knoten, wobei $s \in P$ und $t \in \bar{P} = V \setminus P$ mit
$$\text{Cut}^{s,t}(P) = \sum_{\substack{v_i \in P, v_j \in \bar{P} \\ s \in P, t \in \bar{P}}} w_{ij}$$

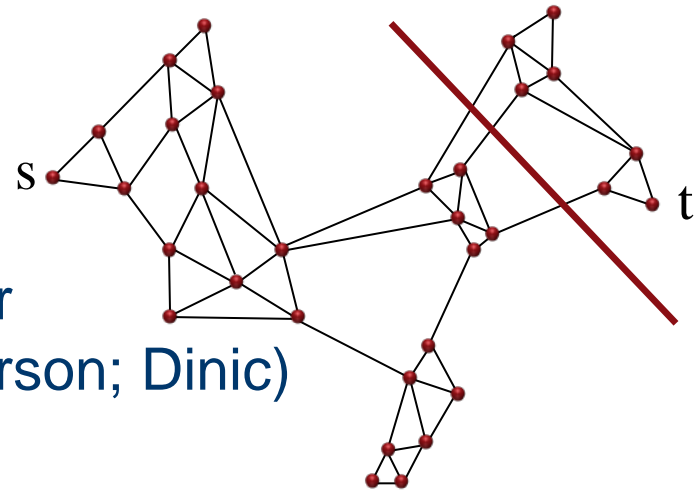
Minimaler Schnitt

Spezialfall $k=2$

- Der minimale s-t-Schnitt

ist $P^* = \arg \min_{P \subset V} \text{Cut}^{s,t}(P)$

- ◆ Problem ist in polynomieller Laufzeit lösbar (Ford/Fulkerson; Dinic)



- Der minimale Schnitt ist der minimale

s-t-Schnitt über alle s-t-Schnitte: $\text{Cut}(P) = \sum_{v_i \in P, v_j \in \bar{P}} w_{ij}$

- ◆ Problem ist in polynomieller Laufzeit lösbar

$$\mathcal{O}(nm + n^2 \log n)$$

Minimaler Schnitt

Balanzierung

- Problem: MinCut-Lösung separiert häufig einzelne Knoten

Minimaler Schnitt

Balanzierung

- Problem: MinCut-Lösung separiert häufig einzelne Knoten
- Balanzierung:
 - ◆ $\text{RatioCut}(P) = \frac{\text{Cut}(P)}{|P|}$ wobei $|P|$ Anzahl der Knoten in P
 - ◆ $\text{Ncut}(P) = \frac{\text{Cut}(P)}{\text{vol}(P)}$, wobei $\text{vol}(P) = \sum_{v_i \in P} d_i$
- Balanzierteres MinCut-Problem ist NP-hart

Minimaler Schnitt

Balanzierung

- **Lemma 1: Sei** $f_i^2 = \begin{cases} |\bar{P}| / |P| & , \text{wenn } v_i \in P \\ -|\bar{P}| / |P| & , \text{sonst} \end{cases}$
dann gilt

$$|V| \cdot \text{RatioCut}(P) = f^T L_{\text{un}} f$$

- **Lemma 2: Sei** $f_i^2 = \begin{cases} \text{vol}(\bar{P}) / \text{vol}(P) & , \text{wenn } v_i \in P \\ -\text{vol}(\bar{P}) / \text{vol}(P) & , \text{sonst} \end{cases}$
dann gilt

$$\text{vol}(V) \cdot \text{NCut}(P) = f^T L_{\text{sym}} f$$

Spectral-Clustering (unnormalisiert)

Relaxation

- RatioCut

$$\min_{P \subset V} f^T L f, \text{ wobei } \sum_{i=1}^n f_i = 0, \sum_{i=1}^n f_i^2 = n$$

Spectral-Clustering (unnormalisiert)

Relaxation

- RatioCut

f_i kann nur 2 Werte annehmen

$$\min_{P \subset V} f^T L f, \text{ wobei } \sum_{i=1}^n f_i = 0, \sum_{i=1}^n f_i^2 = n$$

$$f_i^2 = \begin{cases} |\bar{P}| / |P| & , \text{ wenn } v_i \in P \\ -|\bar{P}| / |P| & , \text{ sonst} \end{cases}$$

Spectral-Clustering (unnormalisiert)

Relaxation

NP-hart

- RatioCut

$$\min_{P \subset V} f^T L f, \text{ wobei } \sum_{i=1}^n f_i = 0, \sum_{i=1}^n f_i^2 = n$$

Eigenwertproblem

- (Unnormalisiertes) Spectral-Clustering

$$\min_{f \in \mathbb{R}^n} f^T L f, \text{ wobei } \sum_{i=1}^n f_i = 0, \sum_{i=1}^n f_i^2 = n$$

Spectral-Clustering (unnormalisiert)

Relaxation

NP-hart

- RatioCut

$$\min_{P \subset V} f^T L f, \text{ wobei } \sum_{i=1}^n f_i = 0, \sum_{i=1}^n f_i^2 = n$$

Eigenwertproblem

- (Unnormalisiertes) Spectral-Clustering

$$\min_{f \in \mathbb{R}^n} f^T L f, \text{ wobei } \sum_{i=1}^n f_i = 0, \sum_{i=1}^n f_i^2 = n$$

- Diskretisierung: $\text{sign}(f_i)$

Spectral-Clustering (unnormalisiert)

Verallgemeinerung auf $k > 2$

- $$\text{Cut}(P_1, \dots, P_k) = \frac{1}{2} \sum_{i=1 \dots k} \text{Cut}(P_i)$$

$$\text{RatioCut}(P_1, \dots, P_k) = \frac{1}{2} \sum_{i=1 \dots k} \text{RatioCut}(P_i)$$

$$\text{Ncut}(P_1, \dots, P_k) = \frac{1}{2} \sum_{i=1 \dots k} \text{Ncut}(P_i)$$

Spectral-Clustering (unnormalisiert)

Verallgemeinerung auf $k > 2$

- $$\text{Cut}(P_1, \dots, P_k) = \frac{1}{2} \sum_{i=1 \dots k} \text{Cut}(P_i)$$

$$\text{RatioCut}(P_1, \dots, P_k) = \frac{1}{2} \sum_{i=1 \dots k} \text{RatioCut}(P_i)$$

$$\text{Ncut}(P_1, \dots, P_k) = \frac{1}{2} \sum_{i=1 \dots k} \text{Ncut}(P_i)$$

- $$f_i^2 = \begin{cases} |\bar{P}| / |P| & , \text{ wenn } v_i \in P \\ -|\bar{P}| / |P| & , \text{ sonst} \end{cases} \quad \longrightarrow \quad F_{ij}^2 = \begin{cases} 1 / |P_j| & , \text{ wenn } v_i \in P_j \\ -1 / |P_j| & , \text{ sonst} \end{cases}$$

- $$\text{RatioCut}(P_1, \dots, P_k) = \text{Tr}(F^T L F)$$

Spectral-Clustering (unnormalisiert)

Relaxierung ($k > 2$)

NP-hart

- RatioCut

$$\min_{P_1, \dots, P_k} \text{Tr}(F^T L F), \text{ wobei } F^T F = I$$

Eigenwertproblem

- (Unnormalisiertes) Spectral-Clustering

$$\min_{F \in \mathbb{R}^{n \times k}} \text{Tr}(F^T L F), \text{ wobei } F^T F = I$$

Spectral-Clustering (unnormalisiert)

Relaxierung ($k > 2$)

NP-hart

- RatioCut

$$\min_{P_1, \dots, P_k} \text{Tr}(F^T L F), \text{ wobei } F^T F = I$$

Eigenwertproblem

- (Unnormalisiertes) Spectral-Clustering

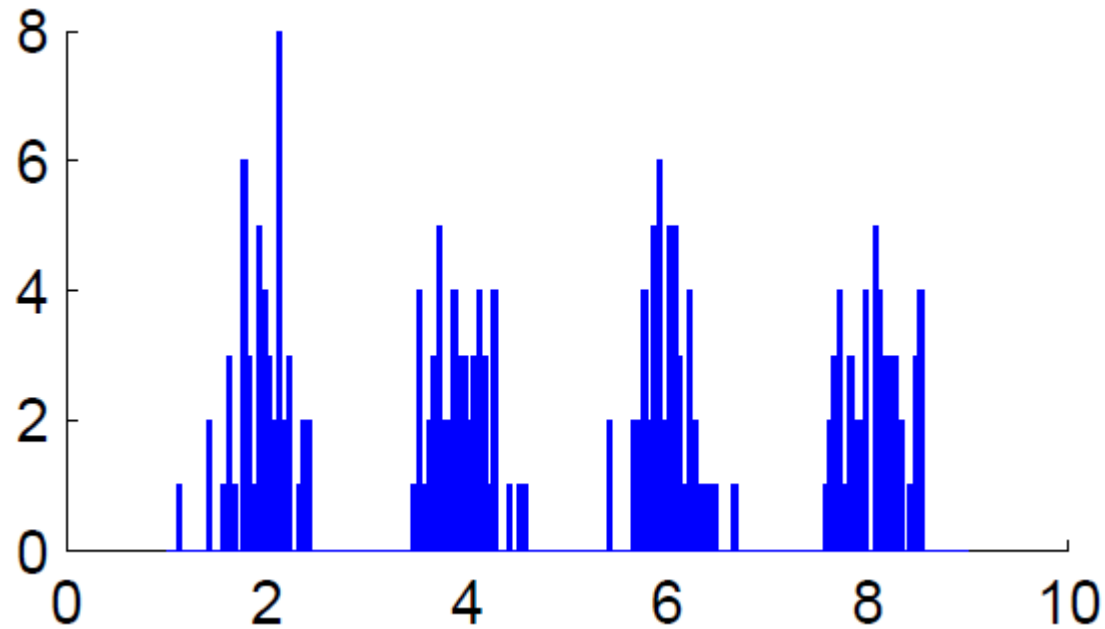
$$\min_{F \in \mathbb{R}^{n \times k}} \text{Tr}(F^T L F), \text{ wobei } F^T F = I$$

- Diskretisierung: k-means auf F_i

Spectral-Clustering

Beispiel

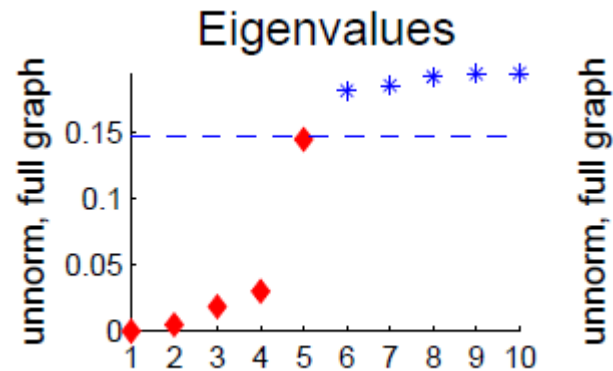
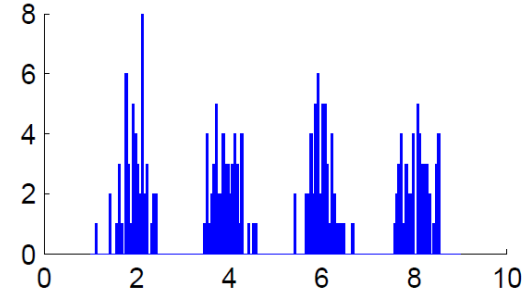
- Daten: Mixture of gaussian



Spectral-Clustering

Beispiel

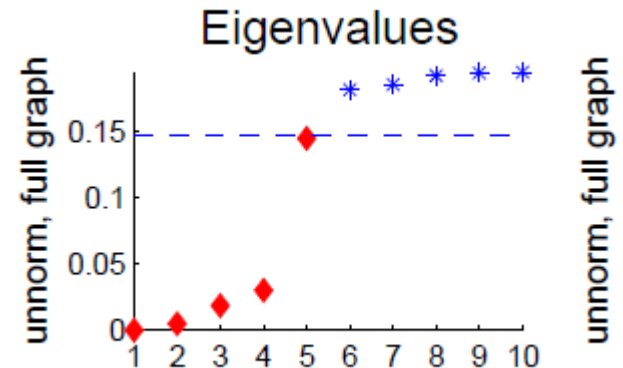
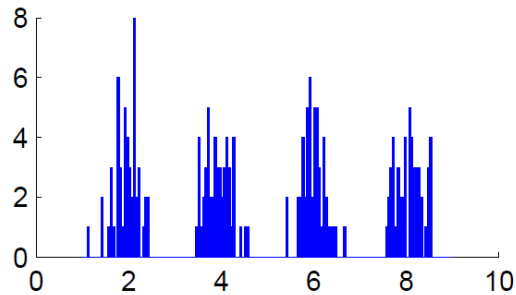
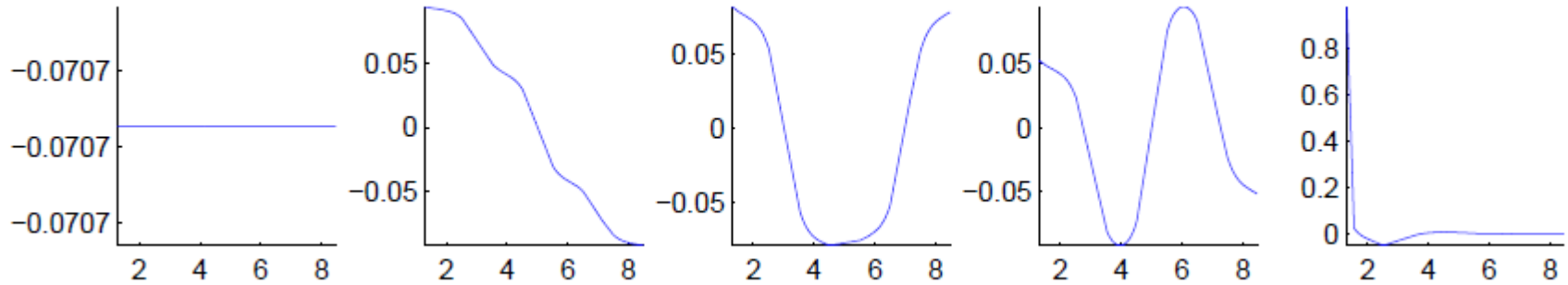
- sim: RBF mit $\sigma = 1$
- Eigenwerte der zugehörigen Laplacematrix (fully connected Graph)



Spectral-Clustering

Beispiel

Eigenvector 1 Eigenvector 2 Eigenvector 3 Eigenvector 4 Eigenvector 5



Spectral-Clustering (unnormalisiert)

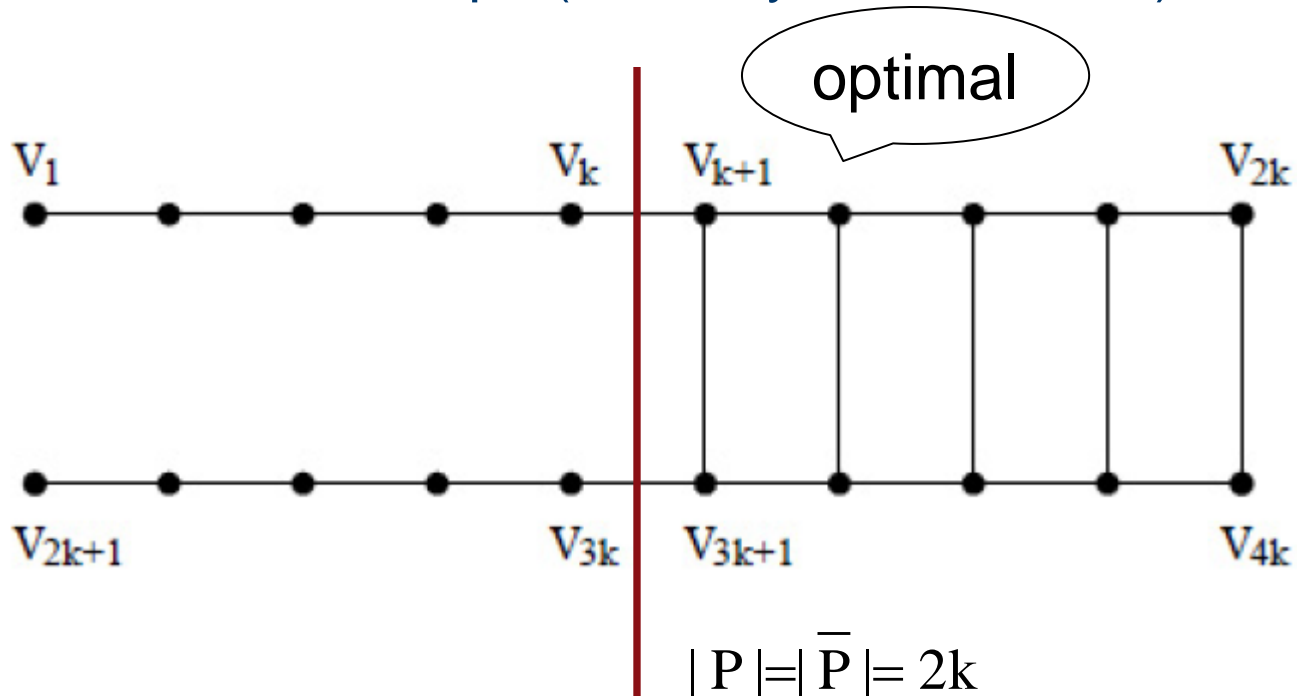
Algorithmus

- Geg.: Adjazenzmatrix $W \in \mathbb{R}_{\geq 0}^{n \times n}$, Clusteranzahl k
- Berechne zugehörige Laplacematrix L_{un}
- Berechne die kleinsten k Eigenvektoren $u_i \in \mathbb{R}^n$ von L_{un}
- Setze
$$\begin{pmatrix} - & x_1 & - \\ & \vdots & \\ - & x_n & - \end{pmatrix} = \begin{pmatrix} | & & | \\ u_1 & \dots & u_k \\ | & & | \end{pmatrix}$$
- Berechne Cluster C_j aus Datenpunkte x_i
- Liefere C_j zurück

Approximationsgüte

Balanzierte Schnitte

- Polynomieller Algorithmus mit konstanter Approximationsgüte existiert nicht
 - ◆ Cockroach Graph (Gatterly & Miller 1998)



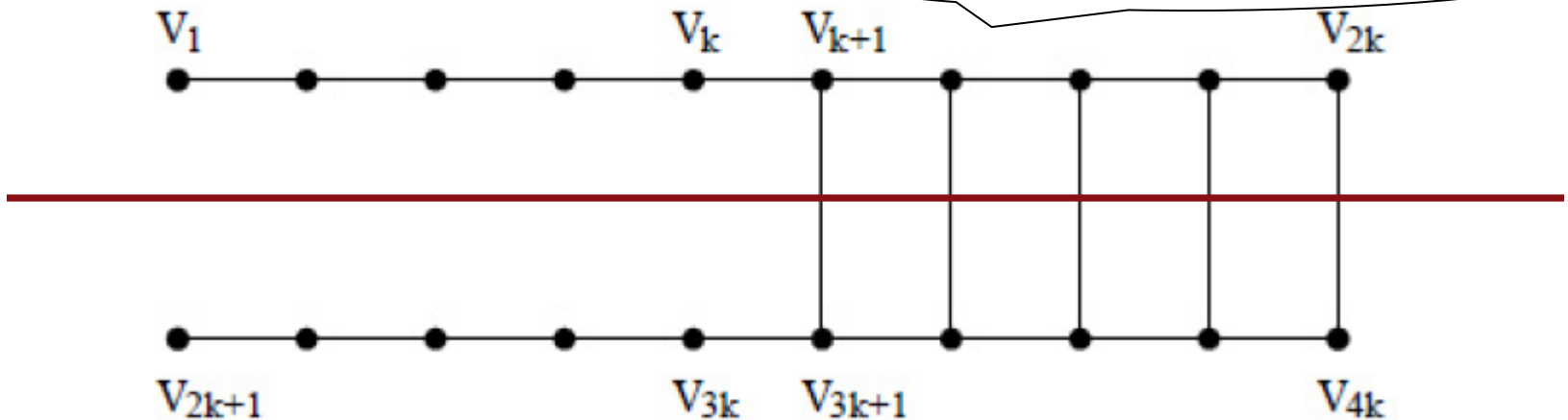
$$\text{cut}(P, \bar{P}) = 2$$

Approximationsgüte

Balanzierte Schnitte

- Polynomieller Algorithmus mit konstanter Approximationsgüte existiert nicht
 - ◆ Cockroach Graph (Gatterly & Miller 1998)

Un. Spectral Clustering



$$|P| = |\bar{P}| = 2k$$

$$\text{cut}(P, \bar{P}) = k$$

Anmerkungen

- Ncut führt zum verallgemeinerten Eigenvektorproblem (norm. Spectral clustering)
- Probabilistische Motivation...

Anmerkungen

- Ncut führt zum verallgemeinerten Eigenvektorproblem (norm. Spectral clustering)
- Probabilistische Motivation...
- Page-Rank: Sortieren der Beispiele nach entsprechendem Eintrag des größten Eigenvektors (bzgl. Eigenwert)
- Quelle:
U. von Luxburg: A Tutorial on Spectral Clustering; 2007

Evaluation von Clusterings

- Unüberwacht (nur heuristisch):
z.B. Agglomerativer Koeffizient, Divisive Coefficient
- Überwacht: Evaluationsdaten sind mit Clusterzugehörigkeit markiert
 - ◆ Relative Entropie

$$H = - \sum_{j=1}^k \frac{|P_j|}{n} \sum_i \delta_{ij} \log \delta_{ij}$$

δ_{ij} ... Anzahl Objekte markiert als i in Cluster j

- ◆ Durchschnittliches F-measure / Recall / Precision
- ◆ Rand-Index