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 $\operatorname{dist}(x_i, x_j) \geq 0$ oder Ähnlichkeitsfunktion $w_{ij} = \operatorname{sim}(x_i, x_j) \geq 0$
  - Erwartete Clusteranzahl k

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

Inter-Cluster Metriken

- Einfacher Abstand

$$d_{min}\left(P_i, P_j\right) = \min_{v \in P, w \in P_j} \mathrm{dist}\left(v, w\right)$$

- Kompletter Abstand

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

- Durchschnittsabstand

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

- Abstand der Zentroide

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

Optimales Clustering

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

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

- 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

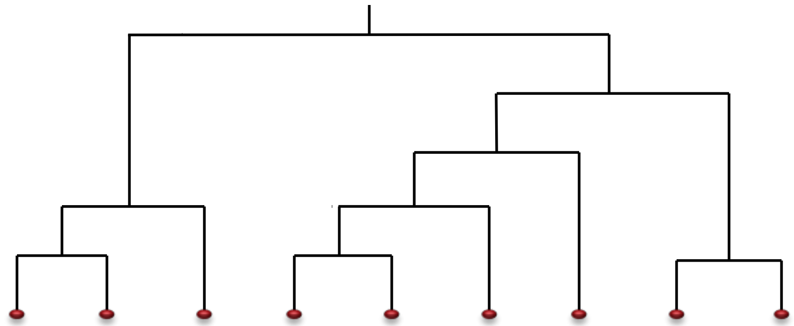
Hierarchisches Clustern

- 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 $c^v, c^w \in C_{i-1}$

    $$(s,t) = \arg\min_{v,w} d\left(c^v, c^w\right); \quad D_i = \min_{v,w} d\left(c^v, c^w\right)$$
  - Setze $C_i = \{c^v \mid \forall v \neq s,t\} \cup \{c^s \cup c^t\}$
- Liefere $C_0, C_1, \ldots$ zurück



Hierarchisches Clustern
Agglomerative Coefficient

- Sei $m_k = d\left(c^s, \{x_k\}\right)$, 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 :

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

- 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 höchsten 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 $C_0, C_1, \ldots$ 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:

$$DC = \frac{1}{n}\sum_{i=1}^{n} \mathrm{dia}_i$$

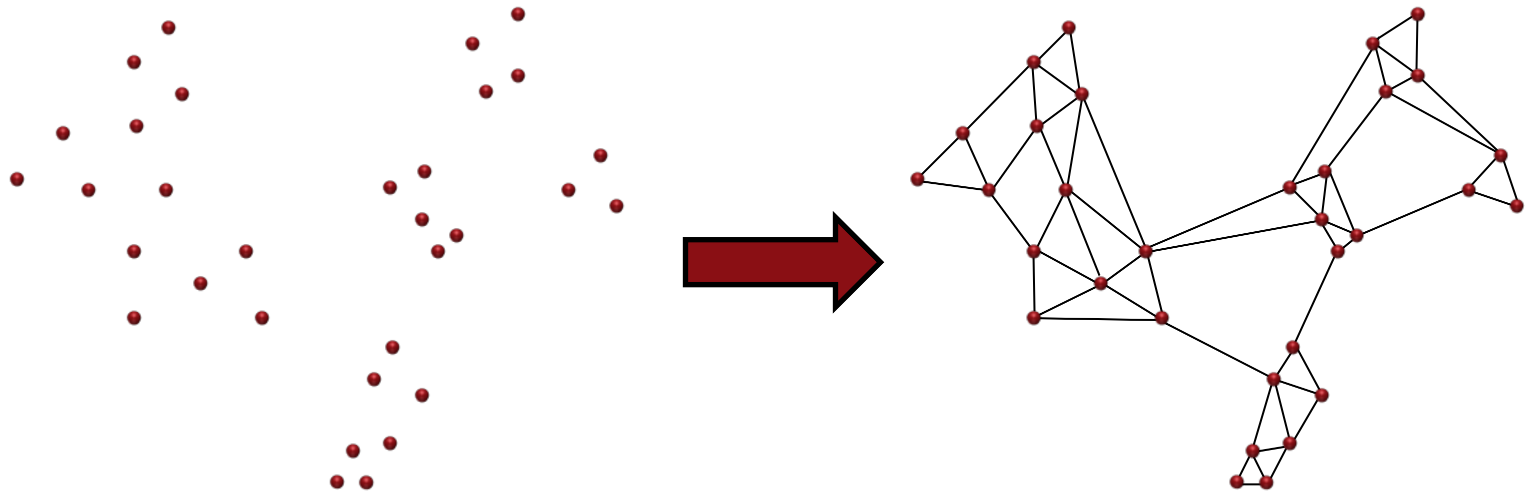
- Ein Maß für die Qualität eines Clusterings
- Nicht geeignet um Datensätze unterschiedlicher Größe zu vergleichen

Überblick

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

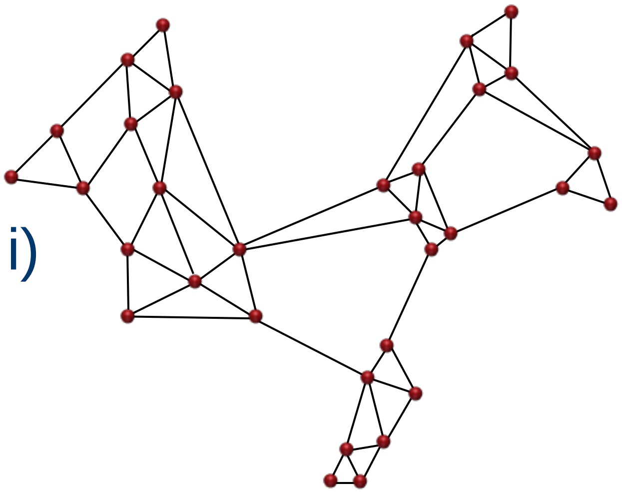
Graphen-basiertes Clustern
Ähnlichkeitsgraph

- 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 $\mathrm{dist}\left(v_i, v_j\right) < \varepsilon$

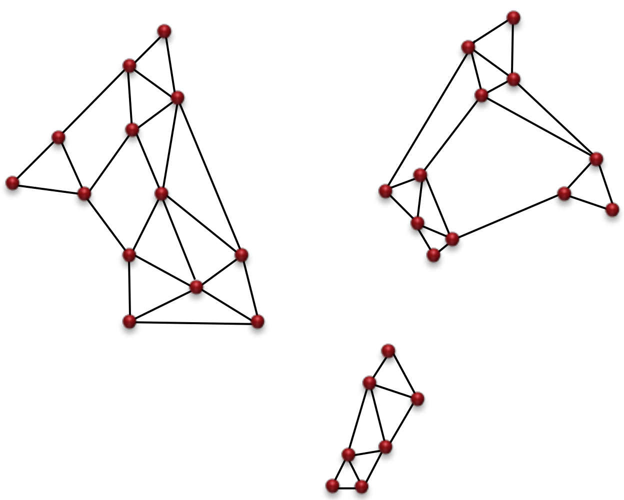


Graphen-basiertes Clustern
Definitionen

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

Beobachtung

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



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

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

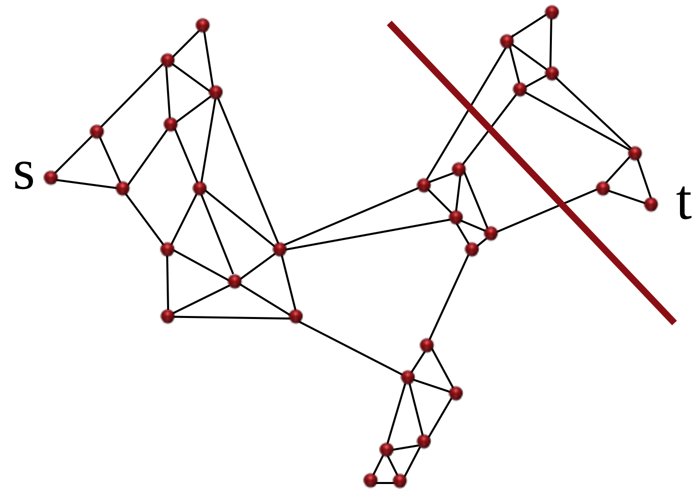
$$u_2 = (0,\ldots 0, 1, \ldots 1, 0, \ldots 0)$$

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

$$f^T L_{un} f = f^T D f - f^T W f = \frac{1}{2} \sum_{i,j=1}^{n} w_{i,j} \left(f_i - f_j\right)^2$$

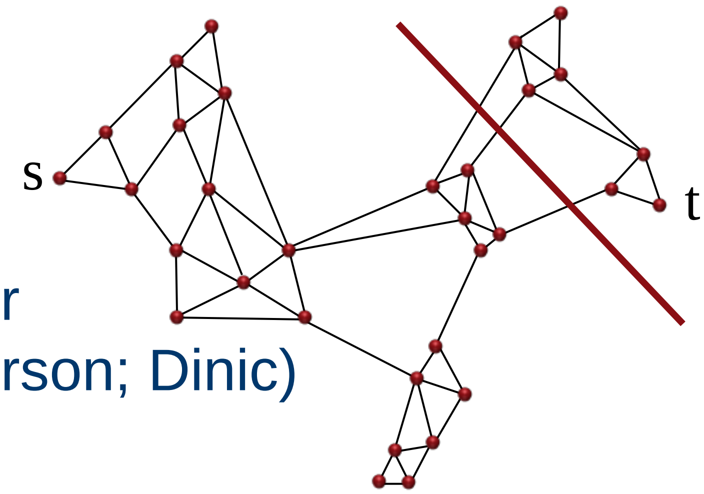
Minimaler Schnitt
Spezialfall k=2

- 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 $t \in \overline{P} = V \setminus P$ mit $\mathrm{Cut}^{s,t}(P) = \sum_{\substack{v_i \in P, v_j \in \overline{P} \\ s \in P, t \in \overline{P}}} w_{ij}$

Minimaler Schnitt
Spezialfall k=2



- Der minimale s-t-Schnitt ist $P^* = \arg\min_{P \subset V} \mathrm{Cut}^{s,t}(P)$
  - Problem ist in polynomieller Laufzeit lösbar (Ford/Fulkerson; Dinic)

- Der minimale Schnitt ist der minimale s-t-Schnitt über alle s-t-Schnitte: $\mathrm{Cut}(P) = \sum_{v_i \in P, v_j \in \bar{P}} w_{ij}$
  - Problem ist in polynomieller Laufzeit lösbar

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

Minimaler Schnitt
Balanzierung

- Problem: MinCut-Lösung separiert häufig einzelne Knoten

Minimaler Schnitt
Balanzierung

- Problem: MinCut-Lösung separiert häufig einzelne Knoten

- Balanzierung:
  - $\mathrm{RatioCut}(P) = \frac{\mathrm{Cut}(P)}{|P|}$ wobei $|P|$ Anzahl der Knoten in P
  - $\mathrm{Ncut}(P) = \frac{\mathrm{Cut}(P)}{\mathrm{vol}(P)}$, wobei $\mathrm{vol}(P) = \sum_{v_i \in P} d_i$

- Balanziertes MinCut-Problem ist NP-hart

Minimaler Schnitt
Balanzierung

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

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

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

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

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

f_i kann nur 2 Werte annehmen

- RatioCut

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

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

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: $\operatorname{sign}(f_i)$

Spectral-Clustering (unnormalisiert)

Verallgemeinerung auf k>2

- $$\mathrm{Cut}(P_1,\ldots P_k) = \frac{1}{2}\sum_{i=1\ldots k} \mathrm{Cut}\left(P_i\right)$$

$$\mathrm{RatioCut}(P_1,\ldots P_k) = \frac{1}{2}\sum_{i=1\ldots k} \mathrm{RatioCut}\left(P_i\right)$$

$$\mathrm{Ncut}(P_1,\ldots P_k) = \frac{1}{2}\sum_{i=1\ldots k} \mathrm{Ncut}\left(P_i\right)$$

Spectral-Clustering (unnormalisiert)

Verallgemeinerung auf k>2

- $$\mathrm{Cut}(P_1,\ldots P_k) = \frac{1}{2}\sum_{i=1\ldots k} \mathrm{Cut}\left(P_i\right)$$

  $$\mathrm{RatioCut}(P_1,\ldots P_k) = \frac{1}{2}\sum_{i=1\ldots k} \mathrm{RatioCut}\left(P_i\right)$$

  $$\mathrm{Ncut}(P_1,\ldots P_k) = \frac{1}{2}\sum_{i=1\ldots k} \mathrm{Ncut}\left(P_i\right)$$

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

- $$\mathrm{RatioCut}(P_1,\ldots P_k) = \mathrm{Tr}\left(F^T L F\right)$$

Spectral-Clustering (unnormalisiert)

Relaxierung (k>2)

NP-hart

- RatioCut

$$\min_{P_1,\ldots,P_k} \operatorname{Tr}\left(F^T L F\right), \text{ wobei } F^T F = I$$

Eigenwertproblem

- (Unnormalisiertes) Spectral-Clustering

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

Spectral-Clustering (unnormalisiert)

Relaxierung (k>2)

NP-hart

- RatioCut

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

Eigenwertproblem

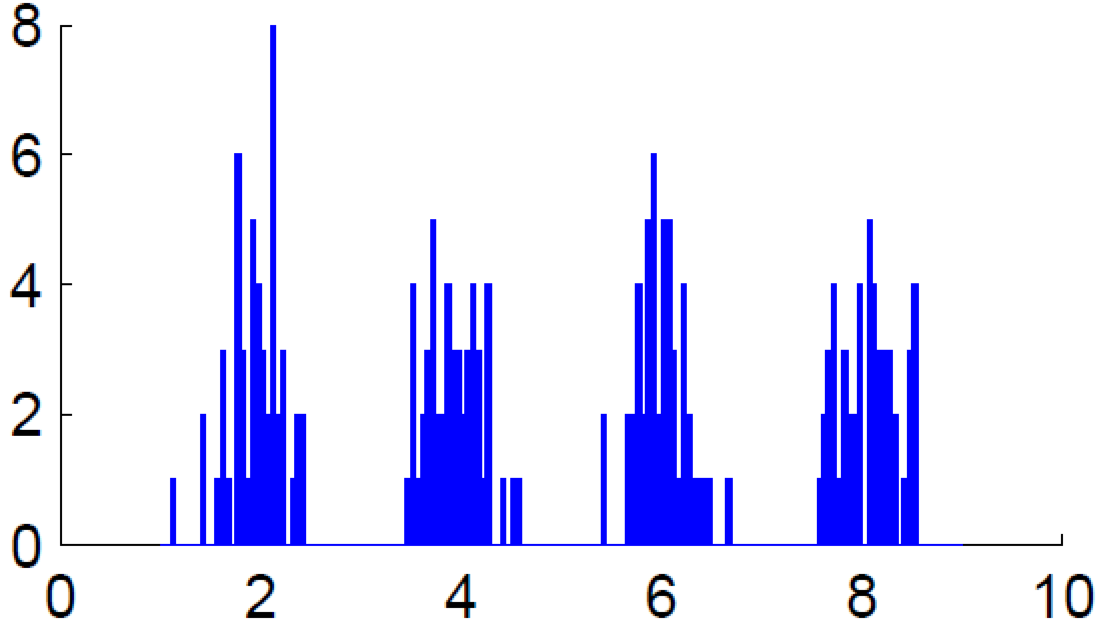
- (Unnormalisiertes) Spectral-Clustering

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

- Diskretisierung: k-means auf F_i

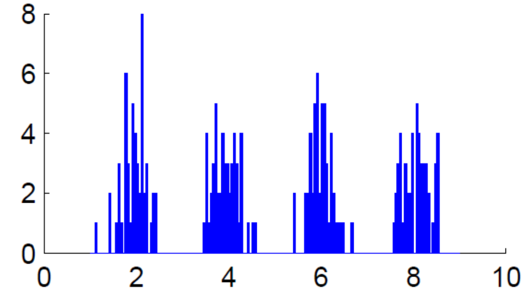
Spectral-Clustering
Beispiel

- Daten: Mixture of gaussian

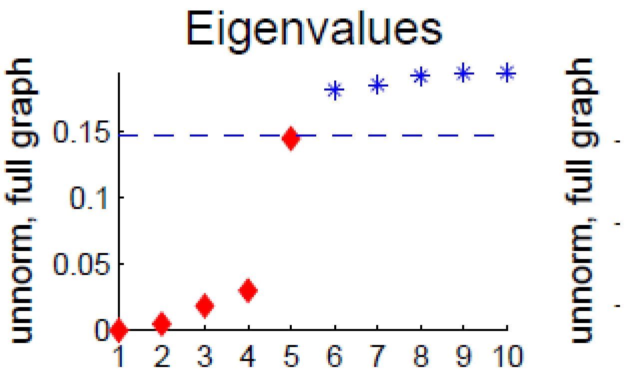


Spectral-Clustering
Beispiel

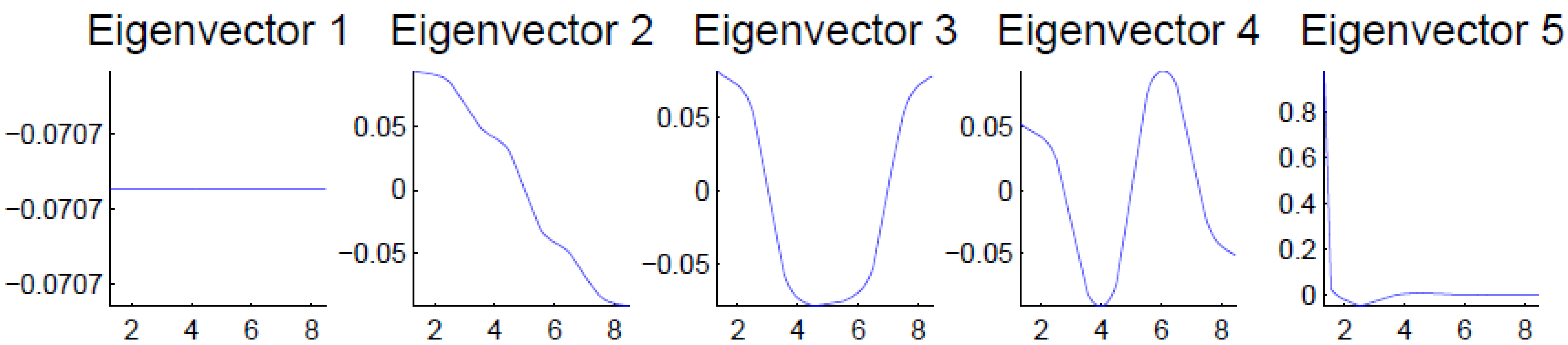
- sim: RBF mit $\sigma = 1$

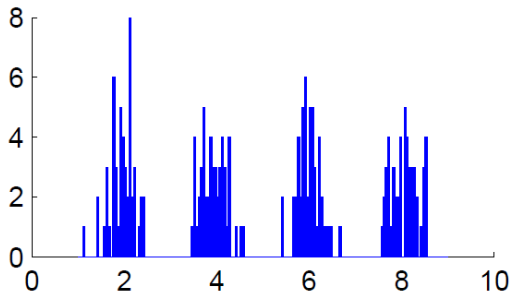


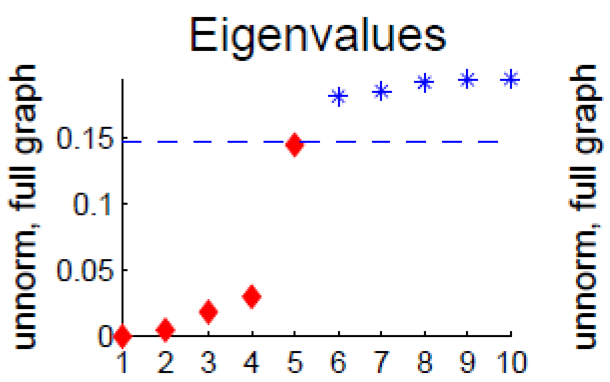
- Eigenwerte der zugehörigen Laplacematrix (fully connected Graph)



Spectral-Clustering
Beispiel





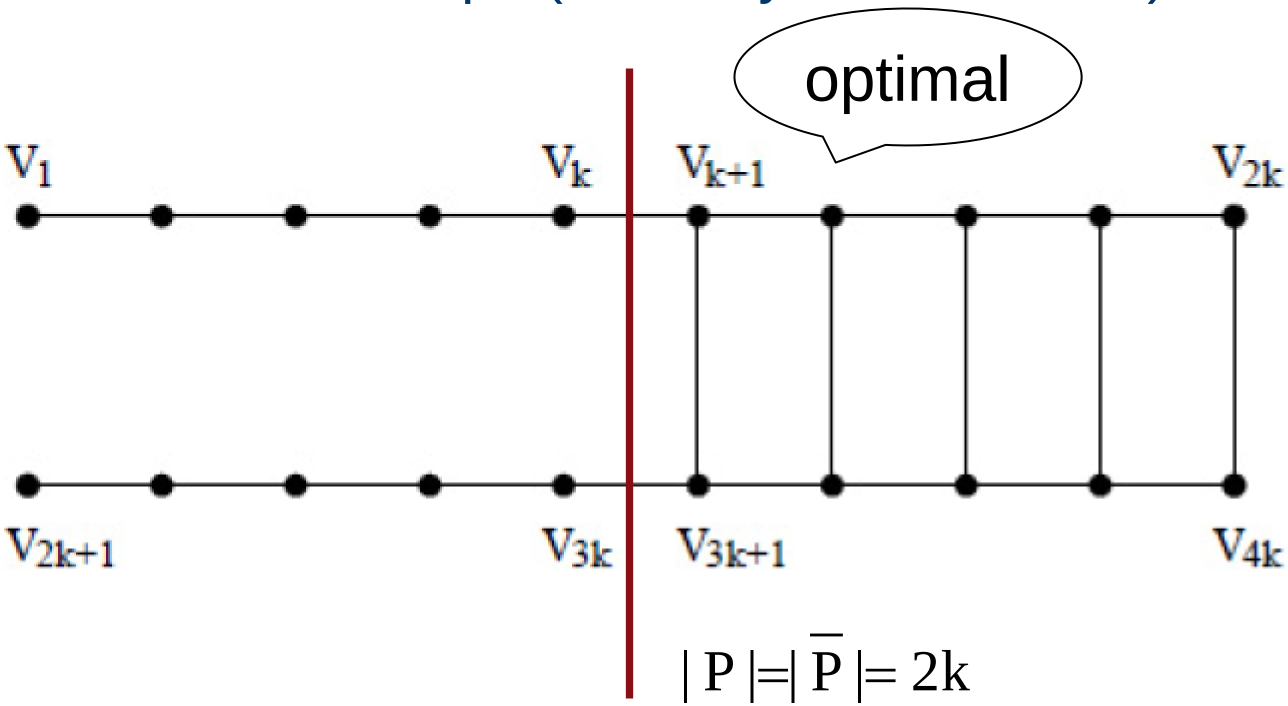


Spectral-Clustering (unnormalisiert)
Algorithmus

- Geg.: Adjazenzmatrix $W \in \mathbb{R}^{n\times n}_{\geq 0}$, 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 (Guattery & Miller 1998)



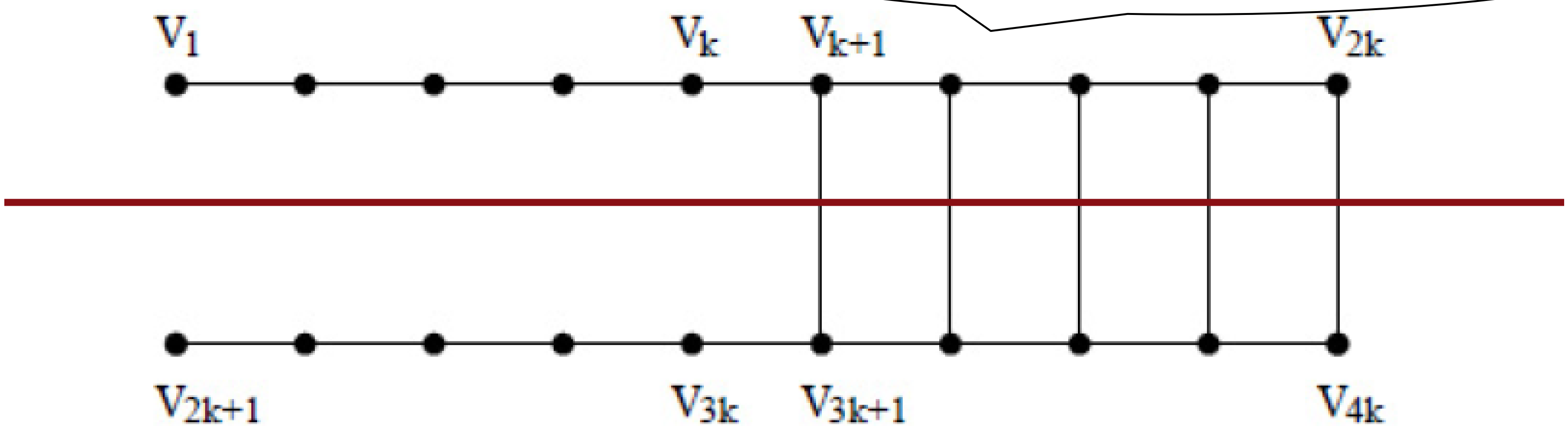
$$|P| = |\overline{P}| = 2k$$

$$\mathrm{cut}\left(P, \overline{P}\right) = 2$$

Approximationsgüte
Balanzierte Schnitte

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

Un. Spectral Clustering



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

$$\mathrm{cut}\left(P, \overline{P}\right) = k$$

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

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

    $\delta_{i|j}$ … Anzahl Objekte markiert als i in Cluster j
  - Durchschnittliches F-measure / Recall / Precision
  - Rand-Index

$$R\left(\{C_1,\ldots,C_k\},\{C_1^*,\ldots,C_l^*\}\right) = \frac{\left|\left\{\{x,\bar{x}\} \mid \exists i,j : \{x,\bar{x}\} \subseteq C_i \wedge \{x,\bar{x}\} \subseteq C_j^*\right\}\right| + \left|\left\{\{x,\bar{x}\} \mid \forall i,j : \{x,\bar{x}\} \subsetneq C_i \wedge \{x,\bar{x}\} \subsetneq C_j^*\right\}\right|}{\binom{n}{2}}$$