Nowe podejścia do analizy skupień

S.T. Wierzchoń

Instytut Podstaw Informatyki Polskiej Akademii Nauk
Gdańsk, Warszawa

Spotkanie polskiej grupy badawczej systemów uczących się. Warszawa, 12-13 XI 2013

Plan prezentacji

1. Wybrane modyfikacje algorytmu k-średnich
2. Spektralny algorytm grupowania
3. Uogólnienia

Przykładowa literatura przeglądowa

- P. Berkhin. A survey of clustering data mining techniques. In J. Kogan, Ch. Nicholas, and M. Teboulle, eds, *Grouping Multidimensional Data*, pp. 25–72. Springer 2006
- M. Filippone, *et al.* A survey on spectral and kernel methods for clustering. *Pattern Recognition*, 41(1):176–190, 2008.
- R. Xu and D. Wunsch II. Survey of clustering algorithms. *IEEE Trans. on Neural Networks*, 16(3):645–678, 2005

- D. Graves and W. Pedrycz. Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study. *Fuzzy Sets and Systems*, 161(4):522–543, 2010.

Część I

Algorytm k-średnich

Dlaczego algorytm k-średnich?

- X. Wu *et al.* Top 10 algorithms in data mining *Knowl. Inf. Syst.* 14, 2008, 1–37:

  1. C4.5 and beyond
  2. **The k-means algorithm**
  3. Support vector machines
  4. The Apriori algorithm
  5. The EM algorithm
  6. PageRank
  7. AdaBoost
  8. kNN: k-nearest neighbor classification
  9. Naive Bayes
  10. CART

- Hans-Hermann Bock: Origins and extensions of the k-means algorithm in cluster analysis. *Electronic Journ@l for History of Probability and Statistics*, 4(2), 2008

- Anil K. Jain: Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31, 2010 651–666

Algorytm k-średnich: pomysłodawcy

- H. Steinhaus (1956): Sur la division des corp matériels en parties. *de l'Académie Polonaise des Sciences IV* (C1.III), 801–804
- S.P. Lloyd (1957): Least squares quantization in PCM. Bell Telephone Labs Memorandum, Murray Hill, NJ. Reprinted in: *IEEE Trans. Information Theory*, IT-28 (1982), vol. 2, 129-137
- E.W. Forgy (1965): Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometric Society Meeting, Riverside, California, 1965. Abstract in *Biometrics* 21 (1965) 768
- J. MacQueen (1967): Some methods for classification and analysis of multivariate observations. In: L.M. LeCam, J. Neyman (eds.): *Proc. 5th Berkely Symp. Math. Statist. Probab.* 1965/66. Univ. of California Press, Berkely, vol. I, 281-297
- G. Ball, D., Hall (1965): ISODATA, a novel method of data analysis and pattern classification. Technical report NTIS AD 699616. Stanford Research Institute, Stanford, CA
- J.C. Bezdek (1981): *Pattern Recognition with Fuzzy Objective Function Algorithms*. Plenum Press, New York

Algorytm k-średnich: sformułowanie

Dane WE: zbiór $\mathfrak{X} = \{\mathfrak{x}_1, \ldots, \mathfrak{x}_m\}$. Wektor $\mathbf{x}_i \in \mathbb{R}^n$ to reprezentacja (obraz) fizycznego obiektu $\mathfrak{x}_i$, $i = 1, \ldots, m$. Macierz $X = (\mathbf{x}_1, \ldots, \mathbf{x}_m)^\mathsf{T}$ reprezentuje zbiór $\mathfrak{X}$.

Problem: Niech $\mathfrak{J} = \{1, \ldots, m\}$ i niech $\mathcal{C} = \{C_1, \ldots, C_k\}$, $k \geqslant 2$, będzie podziałem zbioru $\mathfrak{J}$, tzn.: (a) $C_{i_1} \cap C_{i_2} = \emptyset$ jeżeli $i_1 \neq i_2$, oraz (b) $\cup_{i=1}^{k} C_i = \mathfrak{J}$. Niech

$$\boldsymbol{\mu}_i = \frac{1}{|C_i|} \sum_{j \in C_i} \mathbf{x}_j$$

oznacza centroid (prototyp, reprezentanta) grupy C_i. Należy znaleźć taki podział $\mathcal{C}$, dla którego wskaźnik

$$J_m(\mathcal{C}) = \sum_{i=1}^{k} \sum_{j \in C_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|_2^2$$

osiąga minimum. Równoważnie: znaleźć taki podział $\mathcal{C}$ i taki zbiór prototypów $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_k)^\mathsf{T}$, że

$$J'_m(\mathcal{C}, Z) = \sum_{i=1}^{k} \sum_{j \in C_i} \|\mathbf{x}_j - \mathbf{z}_i\|_2^2$$

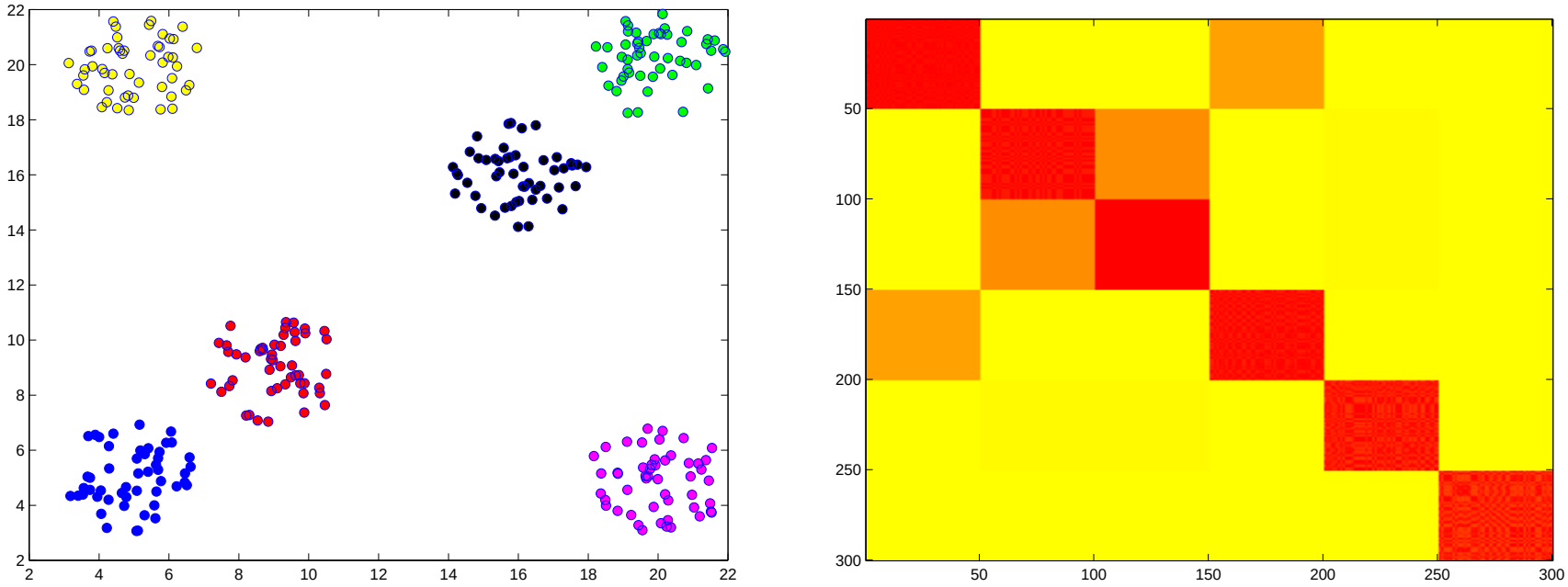
osiąga minimum.

Minimalizacja $J_m(\mathcal{C})$ jest zadaniem NP-trudnym dla $k \geqslant 2$. Jeżeli $C_i \sim N(\boldsymbol{\mu}_i, \sigma_i \mathbb{I}_n)$ oraz $\|\boldsymbol{\mu}_{i_1} - \boldsymbol{\mu}_{i_2}\| \geqslant c\sqrt{n}\max(\sigma_{i_1}, \sigma_{i_2})$ to na ogół algorytm zbiega w niewielkiej liczbie kroków do optimum globalnego (Dasgupta & Schulman, 2007).

Algorytm 1 Algorytm k-średnich (heurystyka Lloyda)

1: WE: Zbiór danych $X = (\mathbf{x}_1, \ldots, \mathbf{x}_m)^{\mathtt{T}}$, $\mathbf{x}_j \in \mathbb{R}^n$, liczba grup $k \geqslant 2$
2: Wybierz k środków grup $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$.
3: **while** (not done) **do**
4: $c(j) = \arg\min_{1 \leqslant i \leqslant k} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|_2, \ j = 1, \ldots, m$
5: $C_i = \{j : c(j) = i\}$
6: $\boldsymbol{\mu_i} = \frac{1}{|C_i|} \sum_{j \in C_i} \mathbf{x}_j$
7: **end while**
8: Zwróć wektor $\mathbf{c}$ oraz macierz $M = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)^{\mathtt{T}}$

Ograniczenia: 1. Wrażliwość na inicjalizację



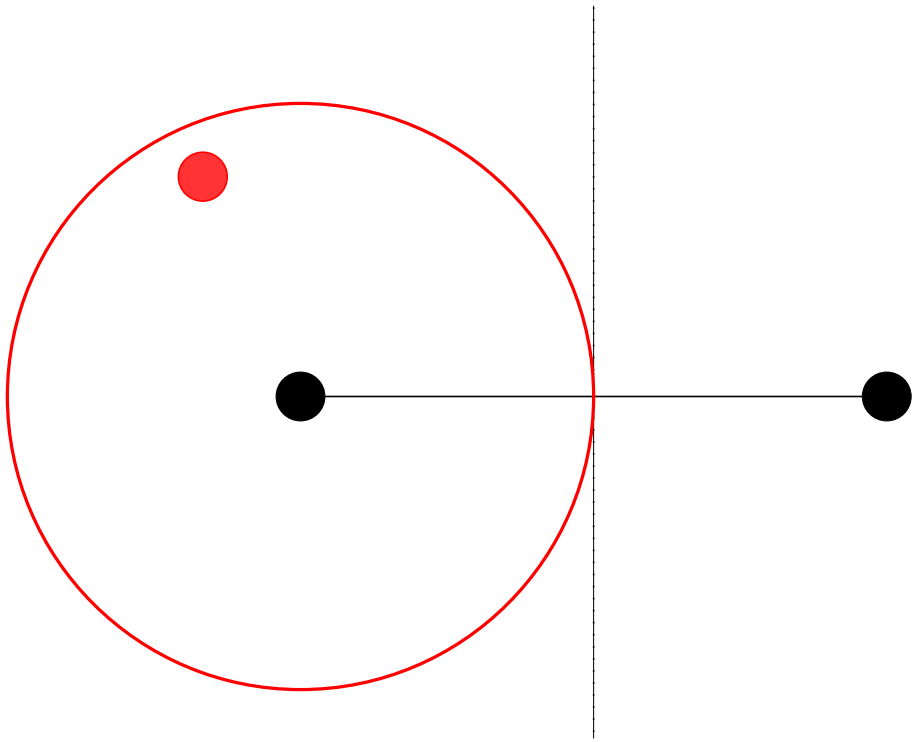
Rysunek 1: Zbiór danych i jego uśredniony (po 100 uruchomieniach) podział na 6 grup

Algorytm k-**means++** (Arthur & Vassilvitskii, 2007): kandydat na j-ty środek ciężkości wybierany jest z prawdopodobieństwem $p(\mathbf{x}) = \frac{u^2(\mathbf{x})}{\sum_{\mathbf{y}\in X} u^2(\mathbf{y})}$, gdzie $u(\mathbf{x}) = d(\mathbf{x}, \{\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_{j-1}\})$.

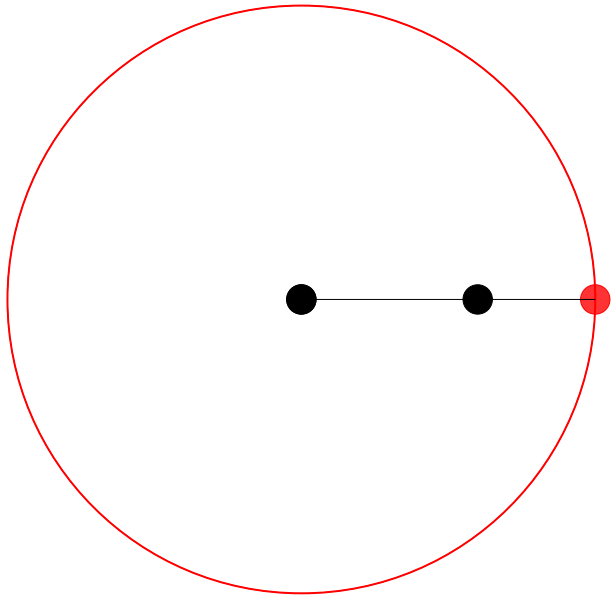
Ograniczenia: 2. Złożoność obliczeniowa

(a) Implementacje równoległe, np. "K-Means Clustering in Map Reduce", http://horicky.blogspot.com/2011/04/k-means-clustering-in-map-reduce
(b) Zastosowanie kart graficznych, np. F. Cao, *et al.* Scalable clustering using graphics processors. In *Advances in Web-Age Information Management* (pp. 372-384). Springer, 2006,
(c) Redukcja obliczeń poprzez wykorzystanie zaawansowanych struktur danych, np. k-d drzew, (Kanungo *et al.* An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Trans. on PAMI* 24(7):881–892, 2002)
(d) Redukcja obliczeń poprzez wykorzystanie nierówności trójkąta (Elkan, 2003)

Obserwacja Elkana (2003)



$d(\boldsymbol{\mu}_1, \boldsymbol{\mu_2}) \geqslant 2d(\mathbf{x}, \boldsymbol{\mu}_1) \Rightarrow d(\mathbf{x}, \boldsymbol{\mu}_2) > d(\mathbf{x}, \boldsymbol{\mu}_1)$



$d(\mathbf{x}, \boldsymbol{\mu}_2) \geqslant \max\left[0, d(\mathbf{x}, \boldsymbol{\mu}_1) - d(\boldsymbol{\mu}_1, \boldsymbol{\mu}_2)\right]$

Ograniczenia: 3. Wrażliwość na liczbę wymiarów

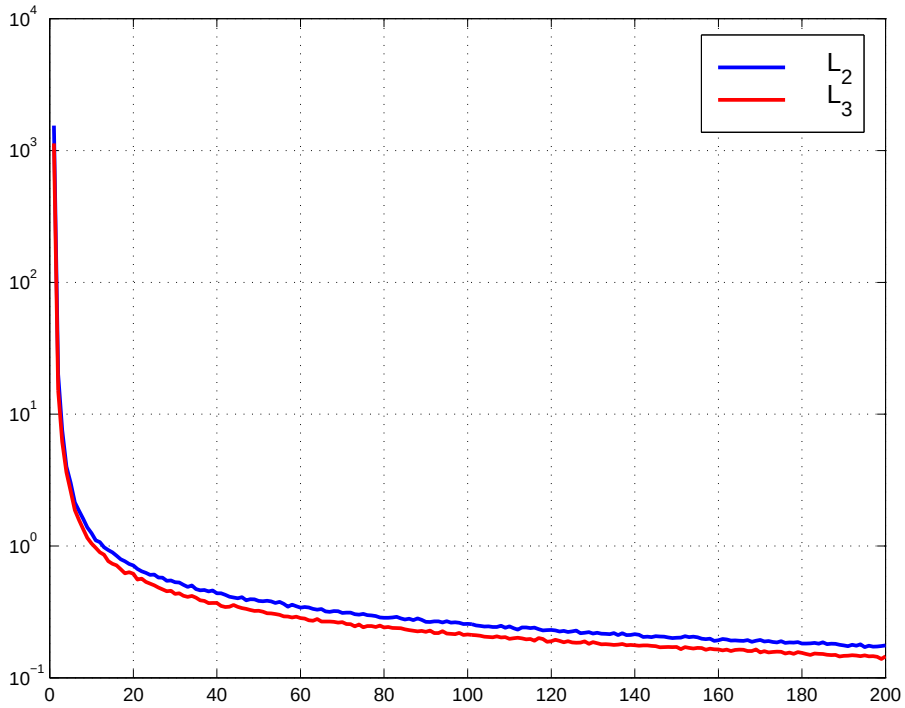
$d_{p,n}^{min}$, $d_{p,n}^{max}$ to minimalna i maksymalna (w zbiorze m losowo wygenerowanych punktów z $[0,1]^n$) odległość L_p od punktu referencyjnego, np. $\mathbf{0} = (0,\ldots,0)^{\mathsf{T}}$. Wówczas

$$C_p \leqslant \lim_{n\to\infty} \mathbb{E}\Big[\frac{d_{p,n}^{max} - d_{p,n}^{min}}{n^{1/p-1/2}}\Big] \leqslant (m-1)C_p$$

gdzie C_p jest stałą zależną od p.

W szczególności

$$d_{p,n}^{max} - d_{p,n}^{min} \to \begin{cases} C_1\sqrt{n} & \text{jeżeli } p = 1 \\ C_2 & \text{jeżeli } p = 1 \\ 0 & \text{jeżeli } p \geqslant 3 \end{cases}$$



Rysunek 2: Przeciętne wartości różnicy $d_{p,n}^{max} - d_{p,n}^{min}$ (lewy rys.) oraz ilorazu $\frac{d_{p,n}^{max} - d_{p,n}^{min}}{d_{p,n}^{min}}$ (prawy rys.) w zbiorze 100 punktów w zależności od liczby wymiarów, n.

Pierwsze modyfikacje

(a) Aggarwal, Hinnenburg i Keim (1973) proponują stosowanie ułamkowych odległości, $0 < p < 1$.

(b) Odległość kosinusowa, $d_{cos}(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x}^{\mathrm{T}}\mathbf{y}}{\|\mathbf{x}\| \cdot \|\mathbf{y}\|}$. Prowadzi to do **sferycznego** algorytmu k-średnich.

(c) Stosując odległość L_2 zakłada się brak korelacji między cechami. Jeżeli ona występuje, wprowadza się odległość Mahalanobisa $d_{\Sigma}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^{\mathrm{T}} \Sigma^{-1} (\mathbf{x}, \mathbf{y})}$, gdzie Σ to macierz kowariancji. Cechy można „dekorelować" przyjmując $\mathbf{x} \leftarrow \Sigma^{-1/2}\mathbf{x}$.

Dywergencja Bregmana (Banerjee *et al.*, 2005)

Niech $\phi\colon \mathcal{S} \to \mathbb{R}$, $\mathcal{S} = \mathrm{dom}(\phi) \subset \mathbb{R}^n$ będzie funkcją ściśle wypukłą i różniczkowalną w $rint(\mathcal{S})$. Dywergencja Bregmana to funkcja

$$d_\phi(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - (\mathbf{x} - \mathbf{y})^\mathsf{T} \nabla\phi(\mathbf{y})$$

Nie musi spełniać warunku trójkąta, ani być symetryczna.

Przykłady:

(1) Jeżeli $\phi(\mathbf{x}) = \|\mathbf{x}\|^2 = \mathbf{x}^\mathsf{T}\mathbf{x}$, to $d_\phi(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_2^2$
(2) Jeżeli $\phi(\mathbf{p}) = \Sigma_{j=1}^n p_j \log_2 p_j$, gdzie $\Sigma_j p_j = 1$, to $d_\phi(\mathbf{p}, \mathbf{q}) = KL(\mathbf{p}\|\mathbf{q})$

Algorytm BHC (*Bregman Hard Clustering*)

Niech $\mathcal{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset \mathcal{S} \subseteq \mathbb{R}^n$ i niech $\mathbf{v}$ będzie rozkładem p-stwa na $\mathcal{X}$. Poszukuje się, dla danej dywergencji d_ϕ, podziału $\mathcal{C}$ minimalizującego

$$B_m(\mathcal{C}, M) = \sum_{i=1}^{k} \sum_{j \in C_i} v_i d_\phi(\mathbf{x}_j, \boldsymbol{\mu}_i)$$

Realizujący to zadanie algorytm BHC posiada następujące własności:

- Osiąga w skończonej liczbie kroków optimum lokalne.
- Prototypy mają postać

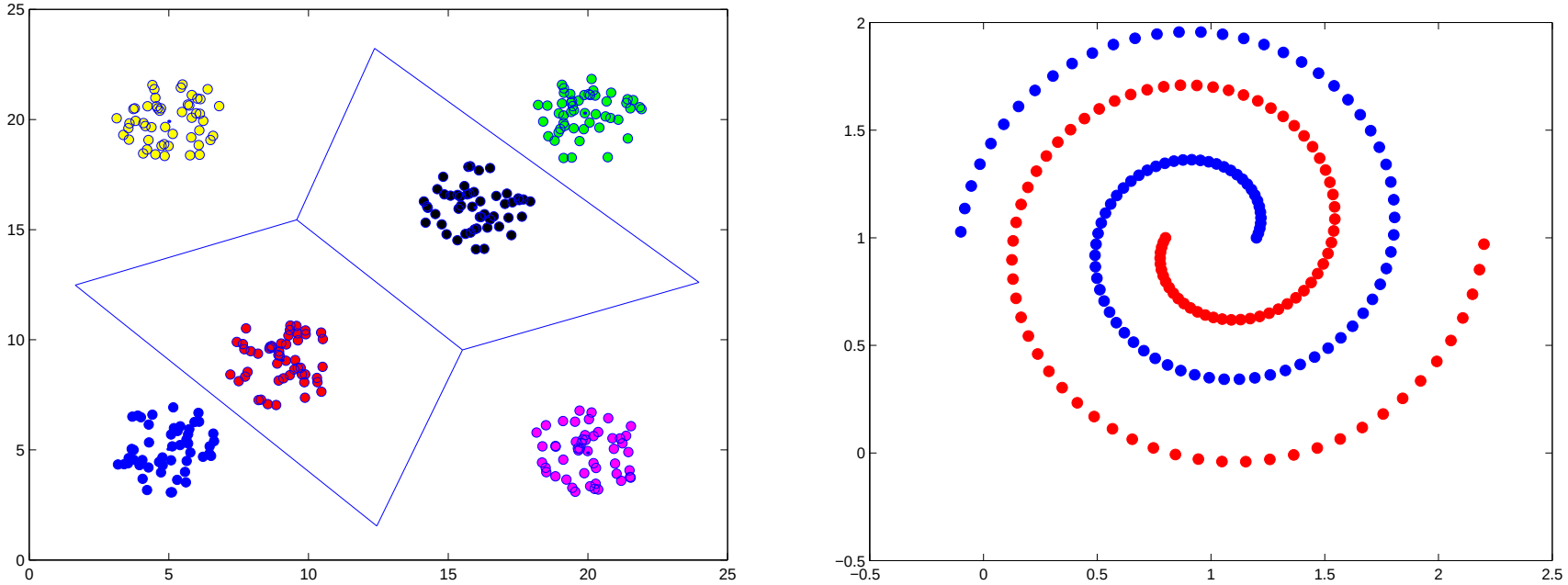
$$\boldsymbol{\mu}_i = \frac{1}{|C_i|} \sum_{j \in C_i} v_j \mathbf{x}_j$$

  Stwierdzenie prawdziwe dla wszystkich dywergencji Bregmana i tylko dla nich. W innych przypadkach nie są to wartości średnie.

- k-średnich, algorytm Linde-Buzo-Gray'a i teorio-informacyjny algorytm (Dhillon, Mallela & Kumar, 2003) to przykłady BHC.
- Granice między grupami wyznaczone są przez hiperpłaszczyzny postaci $\{\mathbf{x}\colon d_\phi(\mathbf{x}, \boldsymbol{\mu}_{i_1}) = d_\phi(\mathbf{x}, \boldsymbol{\mu}_{i_2})\}$.

Dalsze uogólnienia: α-, β- oraz γ-dywergencje (Cichocki & Amari, 2010).

Co potrafi algorytm k-średnich?



Rysunek 3: Lewy rys: zbiór liniowo separowalny, Prawy: nieliniowo separowalny

Rodziny grupowań (Cluster ensembles, consensus clustering)

Cechy charakterystyczne (Firestone, 2012):

- Agregacja grupowań uwzględniających różnorodne kryteria, różnorodne punkty widzenia i potrzeby.
- Użycie zróżnicowanych metod grupowania.
- Możliwość wykorzystania wcześniejszych wyników (np. przekroje historyczne).
- Możliwość uzyskania wyników nieosiągalnych za pomocą pojedynczej metody grupowania.
- Możliwość uzyskania rozwiązań odpornych na obecność *outlier*'ów i innych niejednorodności.
- (Hore, Hall, & Goldgof, 2009): "The advantage of these approaches is that they provide a final partition of data that is comparable to the best existing approaches, yet scale to extremely large datasets. They can be 100,000 times faster while using much less memory".

Rodziny grupowań: formalizacja

$\mathfrak{C} = \{\mathcal{C}^1, \ldots, \mathcal{C}^K\}$ - rodzina podziałów. Szukamy podziału $\mathcal{C}^*$, który jest: (a) zgodny z podziałami $\mathcal{C}^1, \ldots, \mathcal{C}^K$, (b) odporny na drobne zaburzenia w $\mathfrak{C}$.

(1) Jak otrzymać $\mathfrak{C}$?
   (a) Uruchom wielokrotnie algorytm z losowo inicjowanymi centrami.
   (b) Uruchom wielokrotnie algorytm z $k \in \{2, \ldots, k_0 + 10\}$, gdzie k_0 poprawna l-ba skupień (Fred & Jain, 2002), (Greene, *at al.*, 2004).
   (c) Zastosuj różne algorytmy grupowania.
   (d) Inne metody omawia Kuncheva (2004)

(2) Jak reprezentować podziały?
   (a) Macierz współwystępowania $A = [a_{ij}]$ o elementach $a_{ij} = \frac{1}{K} \Sigma_{r=1}^{K} \delta(C_i^r, C_j^r)$
   (b) Zastosować miarę zgodności $q(C^s, C^t)$, np. index Rand'a.

(3) Wybrać podział optymalizujący pewną funkcję zgodności, $q(\mathfrak{C}, C^*) = \min_{1 \leqslant i \leqslant K} q(C^i, C^*)$.

Relacyjne algorytmy grupowania

Dane: macierz (relacja) podobieństwa $S : \mathfrak{J} \times \mathfrak{J} \rightarrow \mathbb{R}$.
Znaleźć taki podział C i taki zbiór reprezentantów $\mathfrak{K} = \{i_1, \ldots, i_k\} \subset \mathfrak{J}$, że

$$\sum_{j \in \mathfrak{J} \setminus \mathfrak{K}} s_{j,c_j} = max!$$

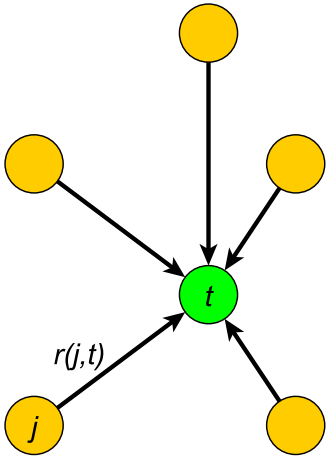
gdzie $c_j \in \mathfrak{K}$ oznacza indeks *reprezentanta* obiektu $i \in \mathfrak{J} \setminus \mathfrak{K}$ oraz $c_i = i$ gdy $i \in \mathfrak{K}$.

Rozwiązanie: algorytm k-medoids.

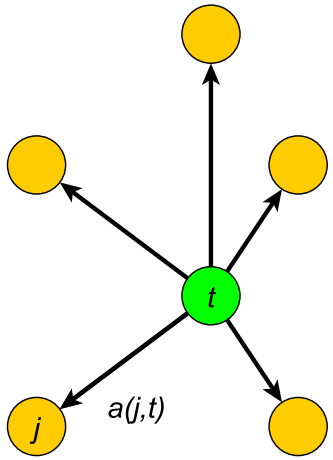
Sformułowanie alternatywne (D. Dueck (2009). "Affinity propagation: Clustering data by passing messages", Ph. D. Thesis):

$$\mathcal{S}(\mathbf{c}) = \sum_{j=1}^{m} s_{j,c_j} + \sum_{l=1}^{m} \delta_l(\mathbf{c}), \;\; \delta_l(\mathbf{c}) = \begin{cases} -\infty & \text{gdy } c_l \neq l \text{ oraz } \exists i \colon c_i = l \\ 0 & \text{w p.p.} \end{cases}$$

Affinity propagation (Frey & Dueck)



odpowiedzialność (*responsibility*), r_{jt}: jak dobrze, zdaniem węzła j, węzeł t jest dopasowany do roli reprezentanta węzła j.



dostępność (*availability*), a_{jt}: jak dobrze, zdaniem węzła t, nadaje się on na reprezentanta węzła j.

Algorytm 2 Propagacja powinowactwa, (Frey & Dueck, 2007)

1: WE: Macierz podobieństw $S = [s_ij]_{m \times m}$.
2: WY: Wskazania c_j reprezentanta obiektu j.
3: $a_{ij} = 0$ dla $i, j = 1, \ldots, m$
4: **while** (**not** warunek zakończenia) **do**
5: aktualizuj odpowiedzialności

$$r_{ij} = s_{ij} - \max_{v \neq j} (a_{iv} + s_{iv}),\ i, j = 1, \ldots, m \tag{1}$$

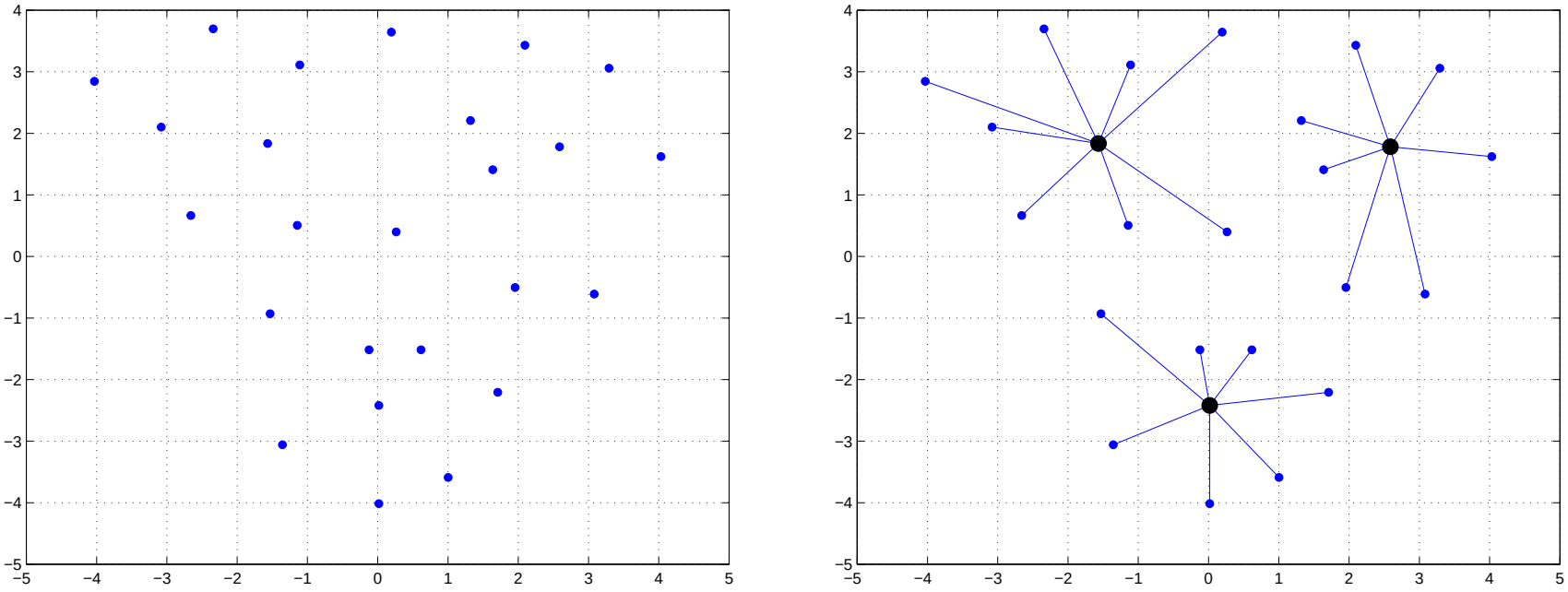
6: aktualizuj dostępności

$$a_{ij} = \begin{cases} \sum\limits_{u \neq j} \max(0, r_{u,j}) & \text{gdy } i = j \\ \min\left[0, r_{jj} + \sum\limits_{u \notin \{i,j\}} \max(0, r_{uj})\right] & \text{w p.p.} \end{cases},\ i, j = 1, \ldots, m \tag{2}$$

7: **end while**
8: wyznacz indeksy

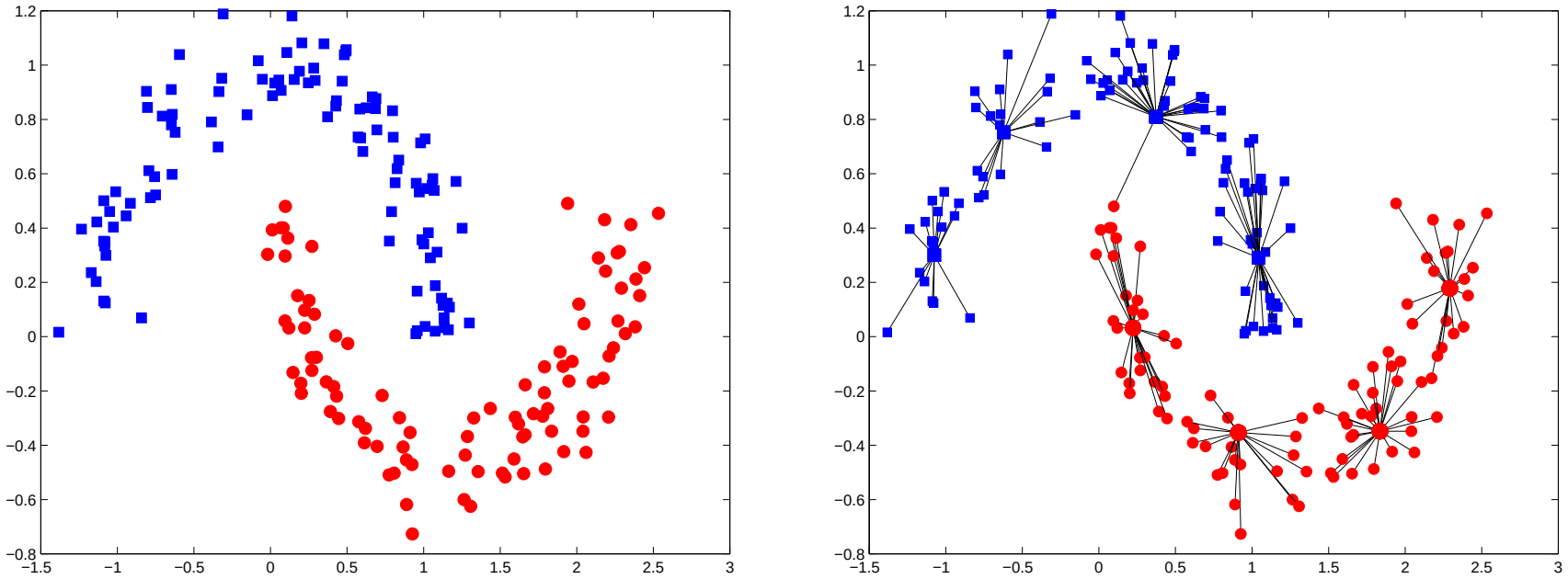
$$c_i = \arg\max_{1 \leqslant j \leqslant m} (r_{ij} + a_{ij}),\ i = 1, \ldots, m \tag{3}$$

Propagacja swoistości w działaniu (1)



Rysunek 4: Lewy rysunek: zbiór wejściowy, Prawy: reprezentanci

Propagacja swoistości w działaniu (2)



Rysunek 5: Lewy rysunek: zbiór wejściowy, Prawy: reprezentanci

Liczba skupień zależy m.in. od „preferencji” s_{jj}. Jeżeli np. $s_{jj} = \min_{i \neq j} s_{ij}$ otrzymuje się niewielką liczbę skupień, a gdy $s_{jj} = mediana(s_{ij})$ – większą liczbę.

Propagacja swoistości w działaniu (3)

Streszczenie artykułu (Frey, Dueck, 2007):

- Affinity propagation identifies exemplars by recursively sending real-valued messages between pairs of data points.
- The number of detected exemplars (number of clusters) is influenced by the values of the input preferences, but also emerges from the message-passing procedure.
- The availability $a(i, k)$ is set to the self responsibility $r(k, k)$ plus the sum of the positive responsibilities candidate exemplar k receives from other points.
- For different numbers of clusters, the reconstruction errors achieved by affinity propagation and k-centers clustering are compared.

Część II

Metody spektralne

Symetryczną macierz S **traktuje się** jak uogólnioną macierz sąsiedztw **grafu podobieństwa** $G = (V, E)$, gdzie $V = \{1, \ldots, m\}$, natomiast $\{v_i, v_j\} \in E$ jeżeli (von Luxburg, 2007):

(a) $s_{ij} \geqslant \tau$, gdzie τ – wartość progowa (w szczególności $\tau = 0$),
(b) Niech $N_k(v_i)$ – zbiór k najbliższych sąsiadów węzła v_i
- $v_j \in N_k(v_i)$ **lub** $v_i \in N_k(v_j)$ (*the k-nearest neighbor graph*), lub
- $v_j \in N_k(v_i)$ **oraz** $v_i \in N_k(v_j)$ (*the mutual k-nearest neighbor graph*).

Podstawowe definicje

$G = (V, E)$ nieskierowany graf prosty o m węzłach i $\mathfrak{e}$ krawędziach. Powiązania między węzłami opisuje symetryczna macierz $S = [s_{ij}]_{n \times n}$ o elementach $s_{ij} \in [0, 1]$, przy czym $s_{ii} = 0$ dla $i = 1, \ldots, n$.

(a) $d_i = \sum_{j=1}^{n} s_{ij}$ to (uogólniony) stopień węzła v_i. $D = \mathrm{diag}(d_1, \ldots, d_n)$ to macierz stopni.

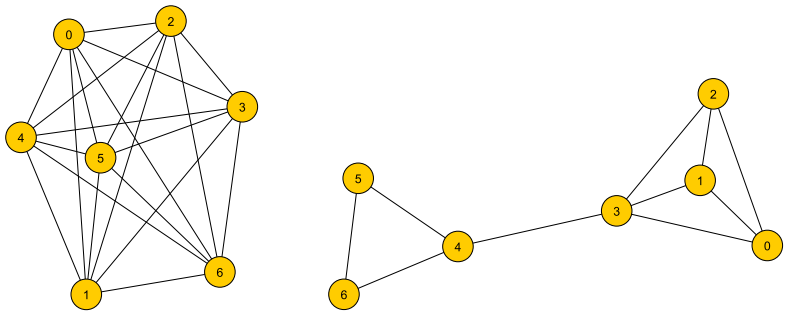
(b) $L = D - S$ to kombinatoryczny (dyskretny) laplasjan grafu G.

(c) Parę $(\lambda, \mathbf{x})$ nazywamy parą własną laplasjanu, jeżeli spełnia ona równanie $L\mathbf{x} = \lambda\mathbf{x}$. Skoro $\sum_{j=1}^{n} l_{ij} = 0$, to $(0, \alpha\mathbf{e})$ (gdzie $\mathbf{e} = (1, \ldots, 1)^{\mathtt{T}}$, α – stała normująca) jest parą własną.

(d) Sortujemy rosnąco wartości własne laplasjanu

$$0 = \lambda_1 \leqslant \lambda_2 \leqslant \ldots, \leqslant \lambda_n$$

Wartość i wektor Fiedlera: $(\lambda_2, \mathbf{x})$



$\lambda_2(G) = 7, \qquad \lambda_2(G') = 0.4$

$\mathbf{x}(G') = (- - - - + + +)$

(a) $\lambda_2 > 0$ gdy G jest grafem spójnym.

(b) Jeżeli $G = (V, E)$ oraz $G' = (V, E')$ przy czym $E \supset E'$ to $\lambda_2(G) \geqslant \lambda_2(G')$.

(c) Unormowane wektory własne macierzy L są ortogonalne, tzn. $\mathbf{e}^{\mathsf{T}}\mathbf{x} = 0$, czyli $\sum_{i=1}^{n} x_i = 0$. Niech $\boldsymbol{\chi}$ będzie binaryzacją wektora $\mathbf{x}$

$$\chi_i = \begin{cases} +1 & \text{gdy } x_i \geqslant 0 \\ -1 & \text{w p.p.} \end{cases} \qquad i = 1, \ldots, n$$

Niech $C = \{v_i \in V : \chi_i \geqslant 0\}$, $\overline{C} = V \backslash C$. Graf $(\overline{C}, E_{\overline{C}})$ jest spójny, a (C, E_C) jest grafem spójnym jeżeli $\chi_i > 0$ dla wszystkich $v_i \in C$.

Uzasadnienie (c)

Koszt rozcięcia:

$$cut(C,\overline{C}) = \sum_{\substack{v_i \in C \\ v_j \in \overline{C}}} s_{ij}, \quad cut(C,\overline{C}) = \frac{1}{4}\boldsymbol{\chi}^{\mathrm{T}} L \boldsymbol{\chi} \Rightarrow \frac{4}{m} cut(C,\overline{C}) = \frac{\boldsymbol{\chi}^{\mathrm{T}} L \boldsymbol{\chi}}{\boldsymbol{\chi}^{\mathrm{T}} \boldsymbol{\chi}}$$

Minimalizacji kosztu odpowiada problem

$$\min \frac{\boldsymbol{\chi}^{\mathrm{T}} L \boldsymbol{\chi}}{\boldsymbol{\chi}^{\mathrm{T}} \boldsymbol{\chi}} \quad p.o.: \ \chi_i \in \{-1,+1\}, i = 1,\ldots,m, \ \boldsymbol{\chi}^{\mathrm{T}} \boldsymbol{\chi} = m$$

Jego relaksacja:

$$\min \frac{\mathbf{y}^{\mathrm{T}} L \mathbf{y}}{\mathbf{y}^{\mathrm{T}} \mathbf{y}} \ p.o.: \ y_i \in [-1,+1], i = 1,\ldots,m, \ \|\mathbf{y}\| = 1$$

Gdy $\mathbf{y} = \mathbf{x}_2$ (wektor Fiedlera) to $\min \frac{\mathbf{y}^{\mathrm{T}} L \mathbf{y}}{\mathbf{y}^{\mathrm{T}} \mathbf{y}} = \lambda_2$.

Problem: Jak binaryzować $\mathbf{y}$? G.A. Tolliver, G.L. Miller (2006) Graph partitioning by spectral rounding: Applications in image segmentation and clustering. *IEEE Computer Soc. Conf. on Computer Vision and Pattern Recognition*, Vol. 1.

Kryteria rozcinania grafu

Koszt rozcięcia na k podgrafów

(a) $Mcut(C_1, \ldots, C_k)) = \sum_{i=1}^{k} cut(C_i, C \backslash C_i)$

(b) $Ncut(C_1, \ldots, C_k)) = \sum_{i=1}^{k} \frac{cut(C_i, C \backslash C_i)}{\mathrm{vol}\, C_i}, \ \mathrm{vol}\,(C_i) = \sum_{j \in C_i} d_j$

(c) $Rcut(C_1, \ldots, C_k)) = \sum_{i=1}^{k} \frac{cut(C_i, C \backslash C_i)}{|C_i|}$

(d) $MinMaxcut(C_1, \ldots, C_k) = \sum_{i=1}^{k} \frac{cut(C_i, C \backslash C_i)}{assoc(C_i)}; \quad assoc(C_i) = \sum_{j_1, j_2 \in C_i} s_{j_1, j2}$

(Ding *et al.*, 2001): Jeżeli dane posiadają wyraźną i dobrze separowalną strukturę, to wszystkie kryteria są jednakowo dobre. Jeżeli grupy są trudno separowalne, lepsze wyniki uzyskuje się stosując kryteria `Ncut` lub `MinMaxcut`, przy czym rola ostatniego kryterium rośnie proporcjonalnie do stopnia „zazębiania się" skupień.

Kryterium `Ncut` (Shi & Malik, 2000)

Minimalizacji kosztu `Ncut` odpowiada wyznaczenie wektorów własnych w uogólnionym problemie własnym

$$L\mathbf{x} = \lambda D\mathbf{x}$$

równoważnym problemowi

$$D^{-1}L\mathbf{x} = \lambda\mathbf{x}$$

Mnożąc stronami przez $D^{1/2}$

$$D^{-1/2}L\mathbf{x} = \lambda D^{1/2}\mathbf{x} \Rightarrow (D^{-1/2}LD^{-1/2})(D^{1/2}\mathbf{x}) = \lambda(D^{1/2}\mathbf{x})$$

tzn.

$$\mathcal{L}\mathbf{y} = \lambda\mathbf{y}$$

gdzie

$$\mathcal{L} = D^{-1/2}LD^{-1/2} = \mathbb{I} - D^{-1/2}SD^{-1/2},\ \mathbf{y} = D^{1/2}\mathbf{x}$$

Normalizowany laplasjan $\mathcal{L}$ jest macierzą symetryczną!

Spektralny algorytm grupowania (1)

Kryterium `Mcut`. Kod w MATLABie

```
S = load('S.adj');                  %wczytanie macierzy sąsiedztw
L = diag(sum(S,2)) - S;             %obliczenie laplasjanu
[v lamb] =eigs(L, k+1, 'SM');       %wyznaczenie k+1 wektorów i wartości wł
X = [v(:, [2:k+1])];                %odwzorowanie spektralne
[idx, C] = kmeans(X,k);             %podział zbioru \textbf{X} na k grup
```

Spektralny algorytm grupowania (2)

Algorytm 3 Spektralny algorytm grupowania Ng, Jordana i Weissa (2002)

1: Utwórz macierz podobieństwa S taką, że $s_{ii} = 0$.
2: Oblicz dopełnienie normowanego laplasjanu $\mathcal{L}^d = \mathbb{I} - \mathcal{L} = D^{-1/2} S D^{-1/2}$.
3: Wyznacz k *największych* wartości własnych macierzy $\mathcal{L}^d$. Odpowiadające im wektory własne stanowią kolumny macierzy $W \in \mathbb{R}^{m \times k}$.
4: Utwórz macierz W' normalizując wiersze macierzy W tak aby miały one jednostkową długość.
5: Traktując wiersze macierzy W' jako współrzędne k-wymiarowych obiektów, podziel ich zbiór na k klas stosując np. algorytm k-średnich.
6: Przydziel obiekt $\mathbf{x}_i$ do skupienia C_j jeżeli i-ty wiersz macierzy W' został zaklasyfikowany do j-tej grupy

Założenie: dominujące wektory własne macierzy L, lub $\mathbb{I} - \mathcal{L}$ dostarczają informację niezbędną do grupowania elementów zbioru danych. Zdaniem Shi, Belkina i Yu (2009) taka hipoteza nie zawsze jest prawdziwa (gdy np. skupienia różnią się liczebnością, gęstością, kształtem). Algorytm `DaSpec` (***data spectroscopic clustering***): pozwala wybrać właściwą liczbę nieredundantnych wektorów własnych, a w konsekwencji wskazuje prawdopodobną liczbę tych klas.

Ograniczenia metod spektralnych

- Brak jasnych wskazówek dotyczących wyznaczania podobieństwa między porównywanymi obiektami. Prosta miara $s_{ij} = \exp(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|)$ wymaga starannego doboru parametru γ.
- Stosowane miary podobieństwa zazwyczaj ignorują całkowicie rozkład przestrzenny obiektów.
- Nie wiadomo jak efektywnie analizować przypadki, w których skupienia definiowane są geometrycznie.
- (Nadler & Galun, 2007):
  (a) Algorytmy grupowania spektralnego przekształcają lokalną informację (liczby s_{ij}) w informację globalną (wektory własne) będącą podstawą grupowania. Korzystanie wyłącznie z lokalnej informacji nie pozwala traktować kosztu rozcięcia jako wiarygodnej miary jakości podziału.
  (b) Nawet jeżeli zastosuje się właściwą miarę podobieństwa, to kilka pierwszych wektorów własnych stosownej macierzy nie wystarcza do poprawnego wydzielenia grup w sytuacjach, gdy grupy te różnią się rozmiarem, gęstością i objętością.

Błądzenie losowe po grafie

Niesymetryczny normalizowany laplasjan

$$\mathfrak{L} = D^{-1}L = \mathbb{I} - D^{-1}S = \mathbb{I} - P$$

P jest wierszową macierzą stochastyczną; p_{ij} to p-stwo przejścia $i \longrightarrow j$. P opisuje odwracalny łańcuch Markowa.

Jeżeli $A, B \subset V$ to

$$P_{A\to B} = \frac{cut(A,B)}{\mathrm{vol}\,(A)}$$

W szczególności, gdy $B = \overline{A}$, to

(a) $Ncut(A,\overline{A}) = P_{A\overline{A}} + P_{\overline{A}A}$
(b) Jeżeli $\mathrm{vol}\,(A) < \frac{1}{2}\mathrm{vol}\,(V)$, to $P_{A\overline{A}} = \Phi(A)$, $\Phi(A)$ - konduktancja

Algorytm MCL, *Markov Chain Clustering*, (van Dongen, 2000)

Algorytm 4 Algorytm MCL

1: WE: A – macierz sąsiedztw analizowanego grafu, $r > 0$ – wykładnik inflacji, s – liczba kroków
2: $A \leftarrow A + I$ // dodaj pętle własne
3: $M \leftarrow AD^{-1}$ // kolumnowa macierz stochastyczna
4: **while** (A nie jest macierzą idempotentną) **do**
5: $M \leftarrow M^s$ // ekspansja
6: $m_{ij} \leftarrow \frac{m_{ij}^r}{\Sigma_l m_{lj}^r}$ // inflacja
7: $M \leftarrow \texttt{prune}(M)$
8: **end while**
9: **return** macierz M

Inne spojrzenia na skupienia

(a) Rezystancyjny obwód elektryczny: spójny nieskierowany graf $G = (V, E)$, w którym każdej krawędzi $\{u, v\}$ odpowiada rezystancja $R_{uv} > 0$, lub równoważnie – konduktancja $C_{uv} = 1/R_{uv}$. Takiemu obwodowi odpowiada łańcuch Markowa o p-stwach $p_{uv} = \frac{C_{uv}}{\Sigma_{w \in N(u)} C_{uw}}$ (Doyle & Snell, 2000). Wu i Hubermann (2004): skoro połączenia między węzłami należącymi do wspólnego skupienia (społeczności) są gęstsze niż połączenia między węzłami należącymi do różnych skupień, to potencjały węzłów z danego skupienia powinny być podobne. Czyli: skupienie to zbiór węzłów o zbliżonych wartościach potencjału.

(b) (Fouss, *et al*, 2007). Przeciętny czas komutacji c_{ij} to oczekiwana liczbę kroków, jaką musi wykonać wędrowiec, aby ze stanu i dojść do stanu j i wrócić do stanu i. Jeżeli G potraktujemy jako rezystancyjny obwód elektryczny, to c_{ij} odpowiada tzw. odległości rezystancyjnej r_{ij}, tzn. $c_{ij} = 2|R|r_{ij}$. Zarówno c_{ij} jak i $\sqrt{c_{ij}}$ są odległościami. Znając macierz $C = [c_{ij}]$ można skonstruować odpowiednik algorytmu k-średnich pozwalającego grupować węzły grafu.

(c) (Orponen, *et al.*, 2008) Niech u będzie stanem pochłaniającym. Lokalnym skupieniem zawierającym u są stany z których można szybko dotrzeć do u.

(d) Modularność (Newman).

Klasyfikacja dokumentów (Shu, *et al.*, 2011)

$T = (\mathbf{t}_1, \ldots, \mathbf{t}_m)^{\mathsf{T}}$ macierz, w której t_{ij} to waga j-tego termu w i-tym dokumencie. Niech $\|\mathbf{t}_i^{\mathsf{T}}\| = 1$. Wówczas $s_{ij} = \mathbf{t}_i^{\mathsf{T}}\mathbf{t}_j$ (podobieństwo dokumentów i oraz j), oraz $S = TT^{\mathsf{T}}$. Uogólniona macierz stopni: $D = \mathrm{diag}(T(T^{\mathsf{T}}\mathbf{e}))$.

Normalizowany laplasjan odpowiadający macierzy S można przedstawić w postaci

$$\mathcal{L} = \mathbb{I} - D^{-1/2}SD^{-1/2} = \mathbb{I} - D^{-1/2}TT^{\mathsf{T}}D^{-1/2} = \mathbb{I} - (D^{-1/2}T)(D^{-1/2}T)^{\mathsf{T}} = \mathbb{I} - CC^{\mathsf{T}}$$

Przedstawmy $C = U\Sigma V^{\mathsf{T}}$. Laplasjan $\mathcal{L}$ przyjmuje postać

$$\mathcal{L} = U(\mathbb{I} - \Sigma\Sigma^{\mathsf{T}})U^{-1}$$

Skoro $\Lambda = \mathbb{I} - \Sigma\Sigma^{T}$ jest macierzą diagonalną, to $U\Lambda U^{-1}$ jest spektralną dekompozycją laplasjanu $\mathcal{L}$ i kolumny macierzy U, które są wektorami osobliwymi macierzy C, są identyczne z wektorami własnymi macierzy $\mathcal{L}$.

Klasyfikacja dokumentów (Lin & Cohen, 2010)

Niech $P = D^{-1}S$. Wyznaczamy dominujący wektor własny stosując metodę potęgową, $\mathbf{x}^{(t)} = P\mathbf{x}^{(t-1)} = P^t\mathbf{x}^{(0)}$

$$\mathbf{x}^{(t)} = P^t\mathbf{x}^{(0)} = \sum_{i=1}^{m} \lambda_i^t \mathbf{u}_i(\mathbf{v}_i^{\mathsf{T}}\mathbf{x}^{(0)}) = \sum_{i=0}^{m} c_i\lambda_i^t\mathbf{u}_i$$

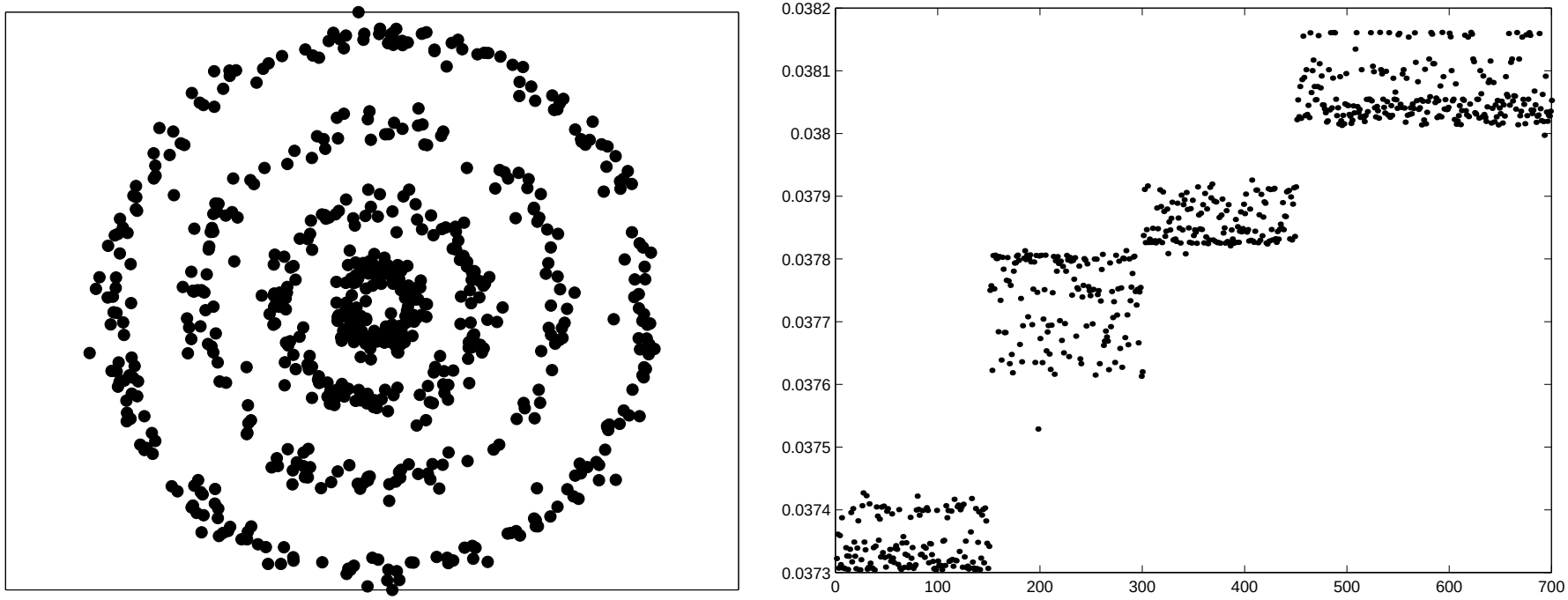
gdzie $\mathbf{u}_i$, $\mathbf{v}_i$ to, odpowiednio, prawy i lewy wektor własny macierzy P odpowiadający wartości własnej λ_i. Porządkujemy malejąco wartości własne, $1 = \lambda_1 \geqslant \lambda_2, \ldots, \lambda_m$

$$\frac{\mathbf{x}^{(t)}}{c_1\lambda_1} = \frac{1}{c_1}\mathbf{x}^{(t)} = \mathbf{u}_1 + \frac{c_2}{c_1}\lambda_2^t\mathbf{u}_2 + \cdots + \frac{c_m}{c_1}\lambda_m^t\mathbf{u}_m$$

Zatem tempo zbieżności wektora $\mathbf{x}^{(t)}$ do dominującego wektora $\mathbf{u}_1$ zależy od potęg λ_i^t. Jeżeli w zbiorze $\mathfrak{X}$ istnieje wyraźna struktura złożona z k grup, to $\lambda_i \approx 1$ dla $i = 2, \ldots, k$ (Meilă & Shi, 2001). Tak więc w początkowych iteracjach $\mathbf{x}^{(t)}$ zbiega do liniowej kombinacji k dominujących wektorów własnych, a pozostałe składniki sumy zanikają nie wolniej niż λ_{k+1}^t. Z chwilą gdy owe „szczątkowe” składniki stają się dostatecznie małe, wektor $\mathbf{x}^{(t)}$ zbiega do $\mathbf{u}_1$ w niemal stałym tempie.

Algorytm 5 Algorytm PIC: *Power Iteration Clustering*, (Lin & Cohen, 2010)

1: WE: Macierz podobieństwa S.
2: Wyznacz macierz $P = D^{-1}S$.
3: Wybierz wektor $\mathbf{x}^{(0)}$; $t = 0$
4: **repeat**
5: $\quad \mathbf{x}^{(t+1)} = \frac{P\mathbf{x}^{(t)}}{\|P\mathbf{x}^{(t)}\|_1}$.
6: $\quad \delta^{(t+1)} = \left\|\mathbf{x}^{(t+1)} - \mathbf{x}^{(t)}\right\|_1$
7: $\quad t = t + 1$
8: **until** $|\delta^{(t)} - \delta^{(t-1)} \approx 0$
9: Zastosuj algorytm k-średnich do wektora $\mathbf{x}^{(t)}$.



(a) (b)

Rysunek 6: Algorytm PIC w akcji. Zbiór danych przedstawiono na rysunku (a). Wartości dominującego wektora własnego po 1500 iteracjach (b).

Część III

Generalizacje

Programowanie dodatnio-półokreślnone (Peng & Wei, 2007

Niech $X = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$, $M = (\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k)$ i niech U oznacza przydział obiektów do grup, tzn. $u_{ij} = 1$ gdy $i \in C_j$ i $u_{ij} = 0$ w p.p. Wówczas

$$\begin{aligned} \min_{u_{ij}} \quad & \sum_{i=1}^{m} \sum_{j=1}^{k} u_{ij} \left\| \mathbf{x}_i - \frac{\Sigma_{l=1}^{m} u_{lj} \mathbf{x}_l}{\Sigma_{l=1}^{m} u_{lj}} \right\|^2 \\ p.o. \quad & \sum_{j=1}^{k} u_{ij} = 1, \quad i = 1, \ldots, m \\ & \sum_{i=1}^{m} u_{ij} > 0, \quad j = 1, \ldots, k \\ & u_{ij} \in \{0, 1\}, \quad i = 1, \ldots, m, j = 1, \ldots, k \end{aligned} \tag{4}$$

Niech $K = X^{\mathrm{T}} X$ i $Z = U(U^{\mathrm{T}} U)^{-1} U^{\mathrm{T}}$ będzie macierzą o elementach

$$z_{ij} = \begin{cases} \dfrac{1}{|C_l|} & \text{gdy } \mathbf{x}_i \in C_l \wedge \mathbf{x}_j \in C_l \\ 0 & \text{w p.p.} \end{cases}$$

Wykonując proste przekształcenia funkcji celu, zadanie (1) można przekształcić do tzw. zero-jedynkowego zadania SDP

$$\begin{array}{ll} \min & tr(K(\mathbb{I} - Z)) \\ p.o. & tr(Z) = k, \quad Z\mathbf{e} = 1 \\ & Z \succcurlyeq 0, Z = Z^{\mathrm{T}}, Z^2 = Z \end{array} \tag{5}$$

Uogólnienia:

(a) Za K można przyjąć dowolne jądro, $k_{ij} = \phi(\mathbf{x}_i, \mathbf{x}_j)$
(b) Zadanie minimalizacji `Ncut` można także sprowadzić do zadania SDP postaci (2), (Xing & Jordan, 2003).

Nieujemna faktoryzacja macierzy (Lee & Seung, 1999)

Aproksymacja ***nieujemnej*** macierzy $A \in \mathbb{R}^{n \times m}$ macierzą $\widetilde{A} \in \mathbb{R}^{n \times m}$ będącą iloczynem dwóch nieujemnych macierzy $W \in \mathbb{R}^{n \times k}$ oraz $H \in \mathbb{R}^{m \times k}$ niskiego rzędu, tzn. $\widetilde{A} = WH^{\mathrm{T}}$, $k << \min(m, n)$.

Kolumny macierzy A reprezentują na ogół obiekty, tzn. $A = (\mathbf{a}_1, \ldots, \mathbf{a}_m)$. Kolumny macierzy $W = (\mathbf{w}_1, \ldots, \mathbf{w}_k)$ nazywane są wektorami bazowymi, a wiersze macierzy H – współczynnikami. Obiekt $\mathbf{a}_i$ uzyskuje reprezentację

$$\widetilde{\mathbf{a}}_i = \sum_{j=1}^{k} \mathbf{w}_j h_{ij}$$

- Y.-X. Wang, Y.-J. Zhang. Nonnegative matrix factorization: A comprehensive review. *IEEE Trans. on Knowledge & Data Eng.*, 25(6), 2013, 1336-1353,
- M.W. Berry *et al.* Algorithms and applications for approximate nonnegative matrix factorization. *Comput. Statistics & Data Analysis*, 52, 2007, 155–173
- J.P. Brunet *et al.* Metagenes and molecular pattern discovery using matrix factorization. *PNAS*, 101(12), 2004, 4164–4169

NMF a grupowanie

(Ding, Li & Jordan, 2008):

$$J_m(\mathcal{C}) = \sum_{i=1}^{m} \sum_{j=1}^{k} u_{ij} \|\mathbf{x}_i - \boldsymbol{\mu}_j\|^2 = \|X - MU^{\mathtt{T}}\|_F^2 \tag{6}$$

Wektorom bazowym macierzy W w $\widetilde{A} = WH^{\mathtt{T}}$ odpowiadają centroidy. Ale X nie jest nieujemna! Relaksacje:

(a) częściowo nieujemna faktoryzacja (*semi NMF*): elementy macierzy F są dowolnego znaku, a elementy macierzy U są nieujemne.
(b) wypukła faktoryzacja: $M = XW$, tzn. $\boldsymbol{\mu}_j = w_{1j}\mathbf{x}_1 + \cdots + w_{mj}\mathbf{x}_m$.

Funkcja $J_m^{\alpha}(\mathcal{C}) = \|X - M(U^{\mathtt{T}})^{\alpha}\|_F^2$, $\alpha > 1$, odpowiada algorytmowi FCM.

Twierdzenie (Li & Ding): Minimum $J_m(\mathcal{C})$ jak i wariantu jądrowego

$$J_{\phi}(\mathcal{C}) = \sum_{i=1}^{k} \sum_{j \in C_i} \|\phi(\mathbf{x}_i - \overline{\phi}_j\|^2$$

można znaleźć rozwiązując problem

$$\max_{U^{\mathtt{T}}U = \mathbb{I}, U \geqslant 0} \operatorname{tr}(U^{\mathtt{T}} K U)$$

gdzie $k_{ij} = \phi^{\mathtt{T}}(\mathbf{x}_i)\phi(\mathbf{x}_j)$. Jeżeli $K = D^{-1/2} S D^{-1/2}$ to problem ten jest równoważny minimalizacji `Ncut`.