Tomáš Horváth

INTRODUCTION TO DATA SCIENCE

Lecture 2

Clustering

Data Science and Engineering Department
Faculty of Informatics
ELTE University

Basic Concepts

$$\frac{1}{F(p;\xi)} = 1+\frac{\alpha}{2\pi^2 p^2}\int \frac{d^3k}{q^2}\frac{F(k;\xi)}{k^2+\mathcal{M}^2(k;\xi)}\Bigg[\mathcal{G}(q)\Bigg\{\frac{a(k,p)}{2q^2}(-q^4+(k^2-p^2)^2)-$$

$$\left[\frac{1}{F(k;\xi)}-\frac{1}{F(p;\xi)}\right]\frac{\Omega(k,p)}{2}(k^2+p^2-q^2)-\left[\frac{b(k,p)(k^2+p^2)-c(k,p)\mathcal{M}(k;\xi)}{2q^2}\right]$$

$$(-q^4+2q^2(k^2+p^2)-(k^2-p^2)^2)\Bigg\}+\xi\Bigg\{\frac{a(k,p)}{2q^2}(q^2(k^2+p^2)-(k^2-p^2)^2)-b(k,p)$$

$$\frac{(k^2-p^2)^2}{2q^2}(k^2+p^2-q^2)+\frac{c(k,p)}{2q^2}\mathcal{M}(k;\xi)\left((k^2-p^2)^2-q^2(k^2-p^2)\right)\Bigg\}\Bigg],\ \frac{\mathcal{M}(p;\xi)}{F(p;\xi)}$$

$$=\frac{\alpha}{2\pi^2}\int\frac{d^3k}{q^2}\frac{F(k;\xi)}{k^2+\mathcal{M}^2(k;\xi)}\Bigg[\mathcal{G}(q)\Bigg\{2a(k,p)\mathcal{M}(k;\xi)-\mathcal{M}(k;\xi)\left[\frac{1}{F(k;\xi)}-\frac{1}{F(p;\xi)}\right]\Omega(k,p)$$

$$+\left[\frac{2b(k,p)\mathcal{M}(k;\xi)+c(k,p)}{2q^2}\right](-q^4+2q^2(k^2+p^2)-(k^2-p^2)^2)\Bigg\}$$

$$+\xi\Bigg\{a(k,p)\mathcal{M}(k;\xi)+b(k,p)\mathcal{M}(k;\xi)\frac{(k^2-p^2)^2}{q^2}+\frac{c(k,p)}{2q^2}(k^2-p^2)(k^2-p^2-q^2)\Bigg\}\Bigg],$$

$$\frac{1}{\mathcal{G}(q)} = 1-\frac{N_f\alpha}{2\pi^2}\int d^3k\frac{F(k;\xi)}{k^2+\mathcal{M}^2(k;\xi)}\frac{F(q;\xi)}{q^2+\mathcal{M}^2(q;\xi)}\Bigg[a(k,q)[W_1(k,p)$$

$$+W_2(k,p)\mathcal{M}(k;\xi)\mathcal{M}(q;\xi)]+b(k,q)[W_3(k,p)+W_4(k,p)\mathcal{M}(k;\xi)\mathcal{M}(q;\xi)]$$

$$-c(k,q)[W_5(k,p)\mathcal{M}(q;\xi)+W_6(k,p)\mathcal{M}(k;\xi)]\Bigg\},$$

Data Types & Attributes

Data

- raw measurements

  symbols, signals, ...
- corresponding to some **attributes**

  height, grade, heartbeat, ...

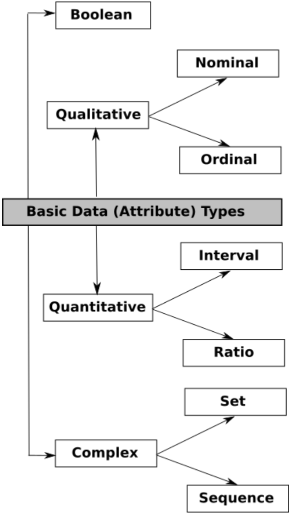
Attribute domain

- expresses the **type** of an attribute

  number, string, sequence, ...
- by the set D of **admissible values**
  - called the **domain** of the attribute

    height up to 3 m, grade from A to F, ...
- and certain **operations** allowed on D

  $1 < 3$, "A" $\geq$ "C", "Jon" $\neq$ "John", ...



What is clustering?

Given the data, **the aim is to group objects** (instances) into so-called clusters, such that objects in the same cluster are (or, at least, should be) more **similar** to each other than to the objects belonging to other clusters

- Similarity plays an important role in clustering!

Similarity of Attribute Values

- $s(x, y) \in [0, 1]$ for $x, y \in D$
  - the opposite to **dissimilarity** computed as the **difference** $d(x, y)$
    - $s(x, y) = 1 - d(x, y)$

Nominal attributes

- w.l.o.g. $D = \{1, 2, \ldots, n\}$
  - $x, y \in D$ are **symbols**
- $s(x, y) = \begin{cases} 1, & \text{if } x = y \\ 0, & \text{if } x \neq y \end{cases}$

Quantitative attributes

- w.l.o.g. $D = \mathbb{R}$
- $d(x, y) = |x - y|$
  - Be aware of the range!
    - normalization

Ordinal attributes

- w.l.o.g. $D = \{1, 2, \ldots, n\}$
  - $x, y \in D$ are **ranks**
- $d(x, y) = \frac{|x-y|}{n-1}$

$D = \{\text{worst,bad,neutral,good,best}\}$

$d(\text{bad,good}) = \frac{|2-4|}{4} = 0.5$

Boolean attributes

- $D = \{0, 1\}$
- as nominal or ordinal

Similarity of Attribute Values

Set attributes

- w.l.o.g. $D = \mathcal{P}(\{1, 2, \ldots, n\}) \setminus \emptyset$
- $s(x, y) = \frac{|x \cap y|}{|x \cup y|}$, $s(x, y) = \frac{|x \cap y|}{min\{|x|,|y|\}}$
  - Jaccard index, Overlap

		T	i	m	i
	0	1	2	3	4
T	1	**0**	1	2	3
o	2	1	**1**	2	3
m	3	2	2	**1**	**2**

Sequence attributes (strings)

- w.l.o.g. $D = \{1, 2, \ldots, n\}^{<\mathbb{N}}$
- $d(x, y) = d_{x,y}(|x|, |y|)$
  - $d_{x,y}(i, j) = \begin{cases} max\{i, j\} & \text{, if } min\{i, j\} = 0 \\ min \begin{cases} d_{x,y}(i-1, j) + 1 \\ d_{x,y}(i, j-1) + 1 \\ d_{x,y}(i-1, j-1) + 1_{x_i \neq y_j} \end{cases} & \text{, otherwise} \end{cases}$
  - Levenshtein distance
    - Be aware of the range!

Similarity of Attribute Values

- *For longer strings, other similarity measures could be beneficial*
  - *longest common substring or subsequence, ...*
- *How would you compute the similarity of two texts?*

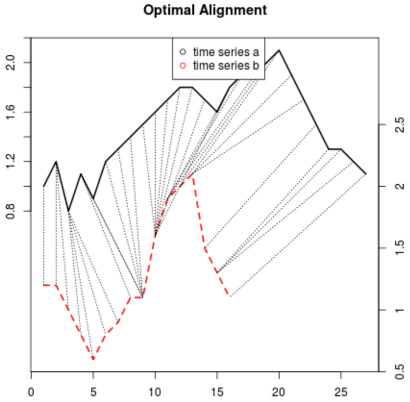
  Will talk about it later in this course...

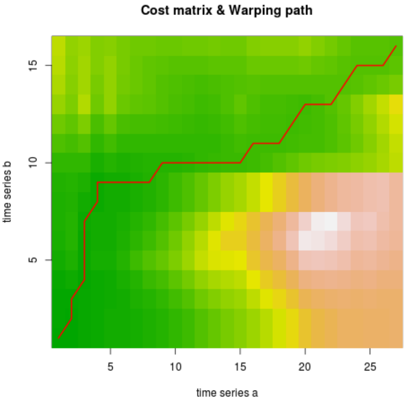
Sequence attributes (time series)

- w.l.o.g. $D = \mathbb{R}^{<\mathbb{N}}$
- $d(x, y) = d_{x,y}(|x|, |y|)$
  - $d_{x,y}(i,j) = \begin{cases} 0 & \text{, if } i + j = 0 \\ |x_i - y_j| + min \begin{cases} d_{x,y}(i-1, j) \\ d_{x,y}(i, j-1) \\ d_{x,y}(i-1, j+1) \end{cases} & \text{, if } i.j > 0 \\ \infty & \text{, otherwise} \end{cases}$
  - Dynamic Time Warping distance
    - Be aware of the range!

Similarity of Attribute Values

Illustration of Dynamic Time Warping





Similarity of Attribute Values

The Basic DTW Algorithm

procedure DTW($x = (x_1, x_2 \dots, x_p)$, $y = (y_1, y_2 \dots, y_q)$)

  $d_{x,y} \leftarrow \mathbb{R}^{(p+1)\times(q+1)}$ ▷ cost matrix $d_{x,y}$

  for all $i \in \{1, 2, \dots, p\}$ **do**

    $d_{(}x, y)(i, 0) \leftarrow \infty$

  for all $j \in \{1, 2, \dots, q\}$ **do**

    $d_{x,y}(0, j) \leftarrow \infty$

  $d_{x,y}(0, 0) \leftarrow 0$

  for $i = 1 \rightarrow p$ **do**

    for $j = 1 \rightarrow q$ **do**

      $d \leftarrow |x_i - y_j|$ ▷ distance of x_i and y_j

$d_{x,y}(i, j) \leftarrow d + min\{d_{x,y}(i-1, j),\ d_{x,y}(i, j-1),\ d_{x,y}(i-1, j-1)\}$

  return $d_{x,y}(p, q)$

Objects, Records, Observations

Object

- A collection of **record**ed measurements (attributes) representing an **entity of observation** (context, meaning)

  e.g a student represented by ID (nominal), age (quantitative), sex (boolean), English proficiency (ordinal), list of absolved courses (set), yearly scores from IQ tests (time-series), ...

- $\mathbf{x} = (x_1, x_2, \ldots, x_m) \in D_1 \times D_2 \times \cdots \times D_m$

- Objects with **mixed types of attributes can be transformed** to objects having boolean or/and quantitative attribute types
  - Be aware of the possible loss of information!
  - *Can you propose some approaches to such transformation?*

Similarity of Binary Instances

Contingency table

- $\mathbf{x} = (x_1, x_2, \ldots, x_m)$
- $\mathbf{y} = (y_1, y_2, \ldots, y_m)$

		$\mathbf{x}$ 1	$\mathbf{x}$ 0	Sum
$\mathbf{y}$	1	a	b	$a+b$
	0	c	d	$c+d$
Sum		$a+c$	$b+d$	m

- $a = \sum_{i=1}^{m} 1_{x_i=1=y_i}$
- $b = \sum_{i=1}^{m} 1_{0=x_i\neq y_i=1}$
- $c = \sum_{i=1}^{m} 1_{1=x_i\neq y_i=0}$
- $d = \sum_{i=1}^{m} 1_{x_i=0=y_i}$

Treating a and d equally

- $s(\mathbf{x}, \mathbf{y}) = \frac{a+d}{m}$
  - Simple matching
- $d(\mathbf{x}, \mathbf{y}) = \sqrt{b+c}$
  - Euclidean distance
  - Be aware of the range!

Treating a and d unequally

- $s(\mathbf{x}, \mathbf{y}) = \frac{a+d/2}{m}$
  - Faith's similarity

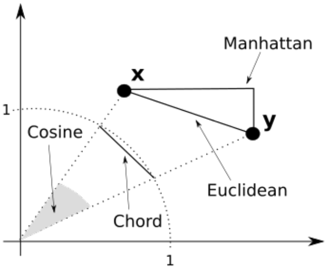
Ignoring d

- $s(\mathbf{x}, \mathbf{y}) = \frac{a}{a+b+c}$
  - Jaccard index

$\mathbf{x} = (0,1,0,1,0,1), \mathbf{y} = (0,1,1,1,1,0),\ a=2, b=2, c=1, d=1$

Similarity of Numerical Instances

Objects are points in an m-dimensional Euclidean space



Cosine similarity

- $s(\mathbf{x}, \mathbf{y}) = \dfrac{\sum_{i=1}^{m} x_i y_i}{\left(\sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} y_i^2\right)^{\frac{1}{2}}}$
  - Be aware of the range!

Minkowski distance

- $d(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{m} |x_i - y_i|^r\right)^{\frac{1}{r}}$
  - Manhattan distance ($r = 1$)
  - Euclidean distance ($r = 2$)
  - Be aware of the range!

Chord distance

- $d(\mathbf{x}, \mathbf{y}) = \left(2\left(1 - \dfrac{\sum_{i=1}^{m} x_i y_i}{\left(\sum_{i=1}^{m} x_i^2 \sum_{i=1}^{m} y_i^2\right)^{\frac{1}{2}}}\right)\right)^{\frac{1}{2}}$
  - Be aware of the range!

$\|\mathbf{x} - \mathbf{y}\|^2 = (\mathbf{x} - \mathbf{y})^\mathsf{T}(\mathbf{x} - \mathbf{y}) =$

$\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2 - 2\mathbf{x}^\mathsf{T}\mathbf{y} = 2(1 - cos(\mathbf{x}, \mathbf{y}))$

if $\|\mathbf{x}\|^2 = \|\mathbf{y}\|^2 = 1$

Similarity of Nominal, Ordinal and Mixed Instances

Ordinal Instances

- $$s(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} o_{ij}^{\mathbf{x}} o_{ij}^{\mathbf{y}}}{\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} |o_{ij}^{\mathbf{x}}|\,|o_{ij}^{\mathbf{y}}|}$$
  - $$o_{ij}^{\mathbf{x}} = \begin{cases} 1 & \text{, if } x_i > x_j \\ -1, & \text{if } x_i < x_j \\ 0 & \text{, if } x_i = x_j \end{cases}$$
  - $o_{ij}^{\mathbf{y}}$ defined as $o_{ij}^{\mathbf{x}}$
  - Goodman & Kruskal
  - Be aware of the range!

$s(\mathbf{x} = (1,2,3), \mathbf{y} = (1,2,3)) =$

$\frac{(-1).(-1)+(-1).(-1)+(-1).(-1)}{3} = \frac{3}{3} = 1$

$s(\mathbf{x} = (1,2,3), \mathbf{y} = (3,2,1)) =$

$\frac{(-1).1+(-1).1+(-1).1}{3} = \frac{-3}{3} = -1$

Nominal Instances

- $$s(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{m} 1_{x_i = y_i}}{m}$$

Mixed Instances

- $$s(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{m} w_i s(x_i, y_i)}{\sum_{i=1}^{m} w_i}$$
  - Gower's index
  - $$w_i = \begin{cases} 1, & \text{if } x_i \neq \text{NA} \neq y_i \\ 0, & \text{otherwise} \end{cases}$$
  - $s(x_i, y_i)$ is a suitable attribute similarity measure

Use-cases

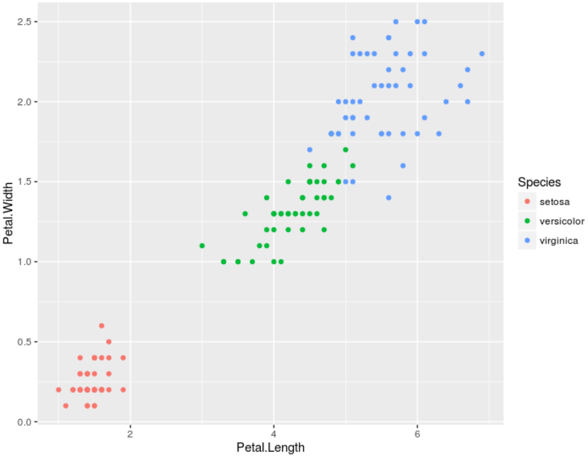
related to clustering, grouping

- Ethnographers would like to create a hierarchy of villages in a broader region such that strongly related regions according to similarity of their folk heritage are at lower levels.
- Marketers would like to divide a broad target market into smaller subsets of customers with similar characteristics in order to estimate their needs and interests.
- Biologists would like to know densely populated clusters of a certain plant in the forest based on satellite images.

An old classic. . .

The **Iris dataset**

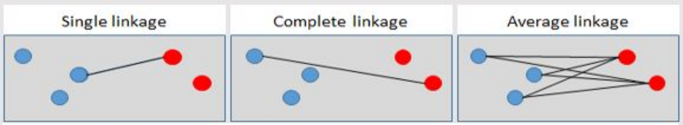
- Iris plants of the class Setosa, Versicolour, Virginica
  - 150 instances, 4 attributes
    - sepal length and width in cm, petal length and width in cm



Hierarchical Agglomerative Clustering

Given

- $D \subseteq D_1 \times D_2 \times \cdots \times D_m$
- a distance measure d (or similarity measure s)
- **linkage** criterion
  - the distance measure between $A, B \subset D$
  - **single** linkage
    - $l(A,B) = min\{d(\mathbf{a},\mathbf{b}) \mid \mathbf{a} \in A, \mathbf{b} \in B\}$
  - **complete** linkage
    - $l(A,B) = max\{d(\mathbf{a},\mathbf{b}) \mid \mathbf{a} \in A, \mathbf{b} \in B\}$
  - **average** linkage
    - $l(A,B) = \frac{1}{|A||B|} \sum_{\mathbf{a} \in A} \sum_{\mathbf{b} \in B} d(\mathbf{a},\mathbf{b})$



Hierarchical Agglomerative Clustering

the goal is to find

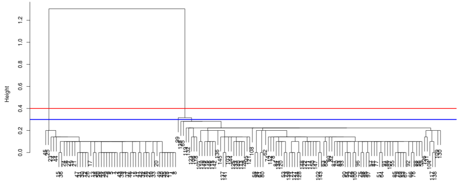
- clusterings $C_1, C_2, \ldots, C_{|D|} \subset \mathcal{P}(D) \setminus \emptyset$ of objects in D such that
  - $C_1 = \{\{\mathbf{x}_1\}, \{\mathbf{x}_2\}, \ldots, \{\mathbf{x}_{|D|}\}\}$
    - initially, each object is in a separate cluster
- and for each $i \in \{2, \ldots, k\}$
  - $C_i = (C_{i-1} \setminus \{A^\star, B^\star\}) \cup (A^\star \cup B^\star)$
    - $A^\star, B^\star \in C_{i-1}$ and $l(A^\star, B^\star) = min\{l(A, B) \mid A, B \in C_{i-1}\}$

Thus, in each step $i \in \{2, \ldots, k\}$

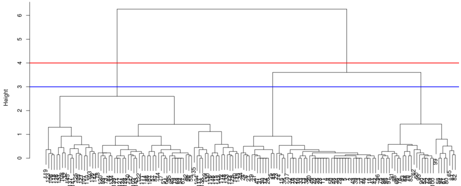
- $|C_i| - |C_{i-1}| = -1$
  - two closest clusters are removed, merged and added as new cluster
- each item is assigned exactly to one cluster

Dendrograms

Single linkage



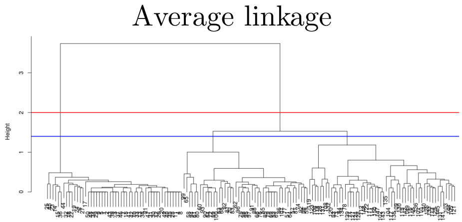
Complete linkage



Cluster	Set.	Vers.	Virg.
Cut at 2 clusters			
1	50	0	0
2	0	50	50
Cut at 3 clusters			
1	50	0	0
2	0	49	50
3	0	1	0

Cluster	Set.	Vers.	Virg.
Cut at 2 clusters			
1	50	29	0
2	0	21	50
Cut at 3 clusters			
1	50	0	0
2	0	21	50
3	0	29	0

Dendrograms



Cluster	Set.	Vers.	Virg.
Cut at 2 clusters			
1	50	0	0
2	0	50	50
Cut at 3 clusters			
1	50	0	0
2	0	45	1
3	0	5	49

Pros of Aggl. Clustering

- easily interpretable
- setting of the parameters is not hard

Cons of Aggl. Clustering

- computationally complex
- subjective interpretation of dendrograms
- obtain quite often local optima

k-Means Clustering

Given

- $D \subseteq D_1 \times D_2 \times \cdots \times D_m$
- a distance measure d (or similarity measure s)
- the number k of clusters
  - $k \ll n$

the goal is to find

- cluster centers $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k$
- and a mapping $p : D \to \{1, 2, \ldots, k\}$ such that

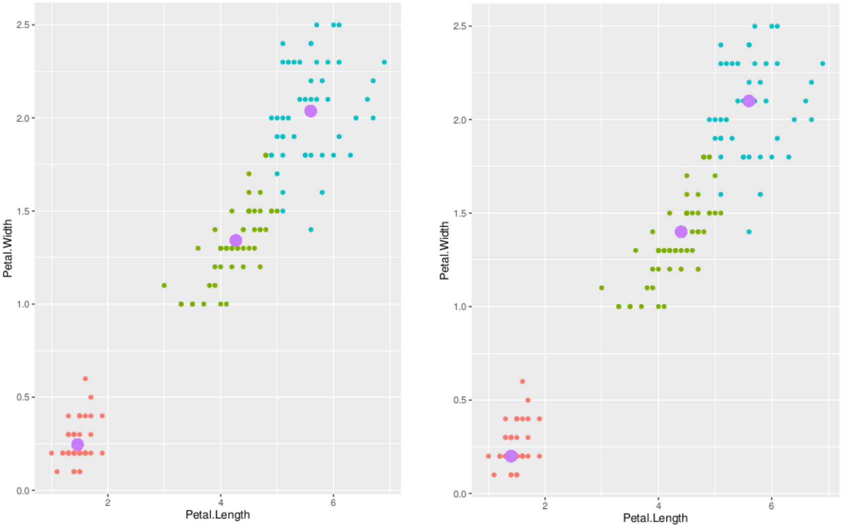
$$\sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{c}_{p(\mathbf{x}_i)}) \text{ is minimal}$$

k-Means Clustering

The algorithm

1. Initialize $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k$ such that for all $i = \{1, 2, \ldots, k\}$
   - $\mathbf{c}_i \in D$ (random initialization), or
   - $\mathbf{c}_i = \frac{\sum_{\mathbf{x}, p(\mathbf{x})=i} \mathbf{x}}{\sum_{\mathbf{x}, p(\mathbf{x})=i} 1}$ for a random mapping p (random partition)
2. compute p such that
   - $\sum_{i=1}^{n} d(\mathbf{x}_i, \mathbf{c}_{p(\mathbf{x}_i)})$ is minimal
3. update $\mathbf{c}_i$ for all $i = \{1, 2, \ldots, k\}$ such that
   - $\mathbf{c}_i = \frac{\sum\limits_{\mathbf{x}, p(\mathbf{x})=i} \mathbf{x}}{\sum\limits_{\mathbf{x}, p(\mathbf{x})=i} 1}$
4. if p or $\mathbf{c}_i$ for some $i = \{1, 2, \ldots, k\}$ were changed then goto step 2
5. return p and $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_k$

Means vs. Medoids



k-Means "Good to know"

Pros

- Computationally efficient
- Obtains, quite often, good results, i.e. global optima

Cons

- The necessity of defining k
- Multiple runs with random initialization recommended
- Can only find partitions with convex shape
- Influence of outliers to cluster centers

External Evaluation of Clusters

- Class labels of instances are known
  - e.g. Setosa, Versicolor, Virginica
  - based on **contingency table**

Object pairs in the same		*Class* Yes	No
Cluster	Yes	a	b
	No	c	d

Rand index

- $RI = \frac{a+d}{a+b+c+d}$

Jaccard index

- $J = \frac{a}{a+b+c}$

Could we use some measure from Information Theory?

$$\text{e.g.} \; -\sum_{i=1}^{k} \sum_{\substack{\mathbf{x}\in D,\\ p(\mathbf{x})=i}} \heartsuit_i \, log \, \heartsuit_i \quad \ldots ?$$

Precision

- $P = \frac{a}{a+b}$

Recall

- $R = \frac{a}{a+c}$

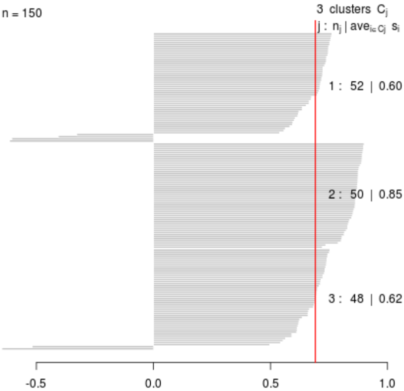
F-measure

- $F_\beta = \frac{(\beta^2+1)P.R}{\beta^2.P+R}$

Internal Evaluation of Clusters

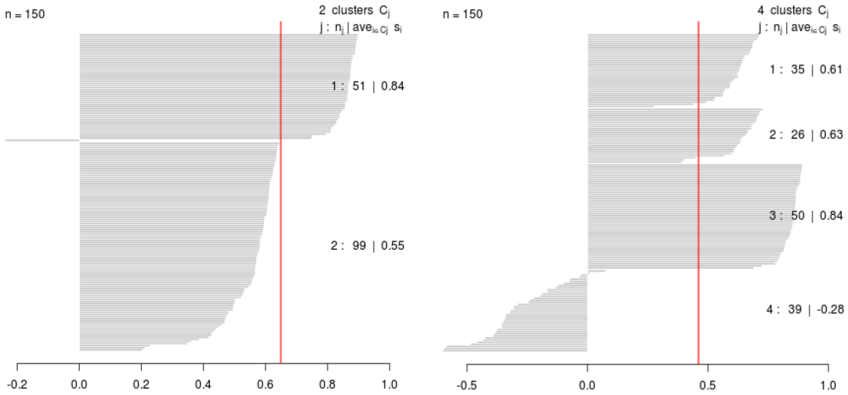
Silhouette

- $S = \frac{1}{|D|} \sum_{\mathbf{x} \in D} sil(\mathbf{x})$
  - $sil(\mathbf{x}) = \frac{b(\mathbf{x}) - a(\mathbf{x})}{max\{a(\mathbf{x}), b(\mathbf{x})\}}$
  - $a(\mathbf{x}) = \frac{\sum_{\mathbf{y} \in D, p(\mathbf{y}) = p(\mathbf{x})} d(\mathbf{x}, \mathbf{y})}{\sum_{\mathbf{y} \in D, p(\mathbf{y}) = p(\mathbf{x})} 1}$
  - $b(\mathbf{x}) =$
    $$\min_{\substack{i \in \{1,2,\ldots,k\}, \\ i \neq p(\mathbf{x})}} \left\{ \frac{\sum_{\mathbf{y} \in D, p(\mathbf{y}) = i} d(\mathbf{x}, \mathbf{y})}{\sum_{\mathbf{y} \in D, p(\mathbf{y}) = i} 1} \right\}$$



- $sil(\mathbf{x}) \in [-1, 1]$
  - $sil(\mathbf{x}) = 1 \Rightarrow \mathbf{x}$ is far away from the neighboring clusters
  - $sil(\mathbf{x}) = 0 \Rightarrow \mathbf{x}$ is on the boundary between two neighboring clusters
  - $sil(\mathbf{x}) = -1 \Rightarrow \mathbf{x}$ is probably assigned to the wrong cluster

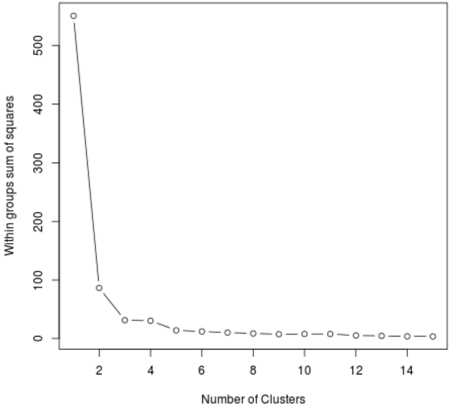
Internal Evaluation of Clusters



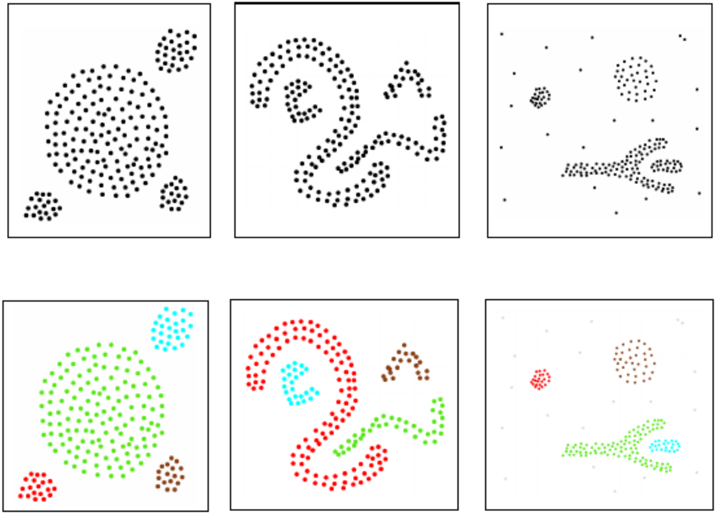
Internal Evaluation of Clusters

Within sum group of squares

- $W = \sum_{i=1}^{k} \sum_{\substack{\mathbf{x} \in D, \\ p(\mathbf{x})=i}} \|\mathbf{x} - \mathbf{c}_i\|^2$



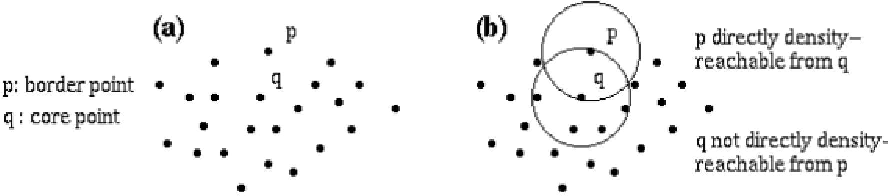
Non-convex Clusters

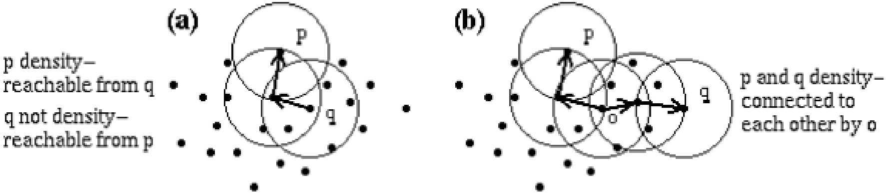


Neighborhood and Reachability

- ϵ-neighborhood of $\mathbf{p} \in D$ defined as $N_\epsilon(\mathbf{p}) = \{\mathbf{x} \in D \mid d(\mathbf{p}, \mathbf{x}) \le \epsilon\}$
- $\mathbf{p}$ is directly density-reachable from $\mathbf{q} \in D$ w.r.t. some ϵ and δ if
  - $\mathbf{p} \in N_\epsilon(\mathbf{q})$
  - $|N_\epsilon(\mathbf{q})| \ge \delta$, i.e. is a core point
- $\mathbf{p}$ is density-reachable from $\mathbf{q}$ w.r.t. some ϵ and δ if
  - $\exists \mathbf{p}_1, \dots, \mathbf{p}_n \in D$ such that $\mathbf{p}_1 = \mathbf{q}$, $\mathbf{p}_n = \mathbf{p}$, and
  - $\mathbf{p}_{i+1}$ is directly density-reachable from $\mathbf{p}_i$ for $2 \le i \le n$
- $\mathbf{p}$ is density-connected to $\mathbf{q}$ w.r.t. some ϵ and δ if
  - $\exists \mathbf{o} \in D$ such that both $\mathbf{p}$ and $\mathbf{q}$ are density-reachable from $\mathbf{o}$
- $C \subseteq D$ $(C \ne \emptyset)$ is a cluster w.r.t. some ϵ and δ if
  - $\forall \mathbf{p}, \mathbf{q} \in D$: if $\mathbf{p} \in C$ and $\mathbf{q}$ is density-reachable from $\mathbf{p}$ then $\mathbf{q} \in C$
  - $\forall \mathbf{p}, \mathbf{q} \in C$: $\mathbf{p}$ is density-connected to $\mathbf{q}$
- $noise = \{\mathbf{p} \in D : \ | : \ \mathbf{p} \notin C_1 \cup \dots \cup C_k\}$ where
  - $C_1, \dots, C_k \subseteq D$ are clusters

Neighborhood and Reachability





DBSCAN

```
procedure DBSCAN(D, ϵ, δ)
    for all x ∈ D do
        p(x) ← −1                              ▷ mark points as unclastered
    i ← 1                                      ▷ the noise cluster have id 0
    for all p ∈ D do
        if p(p) = −1 then
            if ExpandCluster(D, p, i, ϵ, δ) then
                i ← i + 1
```

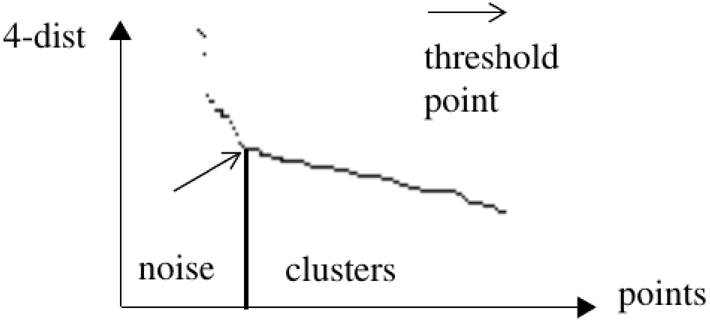
DBSCAN

```
function ExpandCluster(D, p, i, \epsilon, \delta)
    if |N_\epsilon(p)| < \delta then
        p(p) \leftarrow 0                              ▷ mark p as noise
        return false
    else
        for all x \in N_\epsilon(p) do
            p(x) \leftarrow i                          ▷ assign all x to cluster i
        S \leftarrow N_\epsilon(p) \setminus \{p\}
        while S \neq \emptyset do
            s \leftarrow S_1                           ▷ Get the first point from S
            if |N_\epsilon(s)| \geq \delta then
                for all x \in N_\epsilon(s) do
                    if p(x) \leq 0 then
                        if p(x) = -1 then
                            S \leftarrow S \cup \{x\}
                        p(x) \leftarrow i
            S \leftarrow S \setminus \{s\}
        return true
```

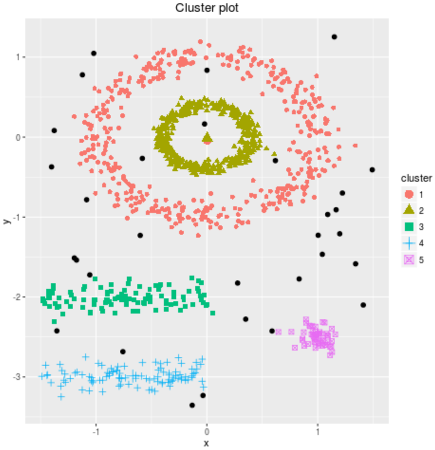
How to guess ϵ and δ?

k-distance

- k-$dist$: $D \to \mathbb{R}$
- k-$dist(\mathbf{x})$ is the distance of $\mathbf{x}$ to its k-th nearest neighbor



DBSCAN – "good to know"



Pros

- Clusters of an arbitrary shape
- Robust to outliers

Cons

- Computationally complex
- Hard to set the parameters

Final remarks

- **domain knowledge might help** in choosing the right similarity measure
- **be aware of the range** of values of the attributes
  - e.g. similarities between $\mathbf{x} = (3.2, 178)$ and $\mathbf{y} = (3.1, 170)$ affected more by the second co-ordinate
- there are **various other approaches** to similarity computation
  - Janos Podani (2000). *Introduction to the Exploration of Multivariate Biological Data. Chapter 3: Distance, similarity, correlation...* Backhuys Publishers, Leiden, The Netherlands, ISBN 90-5782-067-6.


That's all Folks!
Thanks for your attention

References

- Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis (2001). *On Clustering Validation Techniques*. Journal on Intelligent Information Systems 17, 2-3.

- Pang-Ning Tan, Michael Steinbach, and Vipin Kumar(2005). *Introduction to Data Mining*, (First Edition). Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

- Chris Ding and Xiaofeng He (2004). *K-means clustering via principal component analysis*. In Proceedings of the twenty-first international conference on Machine learning (ICML '04). ACM, New York, NY, USA.

- Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu (1996). *A density-based algorithm for discovering clusters in large spatial databases with noise*. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, AAAI Press.

Homework

- Download a clustering dataset from the UCI Machine Learning Repository
- Cluster the dataset using
  - Agglomerative clustering
  - k-means method
  - DBSCAN method
- Justify the choice of the values for the hyper-parameters
  - similarity, linkage, k, δ, ϵ, ...

Questions?

tomas.horvath@inf.elte.hu