

Tomáš Horváth

INTRODUCTION TO DATA SCIENCE

Lecture 2

Clustering

Data Science and Engineering Department
Faculty of Informatics
ELTE University



Basic Concepts

$$\begin{aligned}
\frac{1}{F(p;\xi)} &= 1 + \frac{\alpha}{2\pi^2 p^2} \int \frac{d^3 k}{q^2} \frac{F(k;\xi)}{k^2 + \mathcal{M}^2(k;\xi)} \left[\mathcal{G}(q) \left\{ \frac{a(k,p)}{2q^2} (-q^4 + (k^2 - p^2)^2) - \right. \right. \\
&\left. \left[\frac{1}{F(k;\xi)} - \frac{1}{F(p;\xi)} \right] \frac{\Omega(k,p)}{2} (k^2 + p^2 - q^2) - \left[\frac{b(k,p)(k^2 + p^2) - c(k,p)\mathcal{M}(k;\xi)}{2q^2} \right] \right. \\
&\left. \left. (-q^4 + 2q^2(k^2 + p^2) - (k^2 - p^2)^2) \right\} + \xi \left\{ \frac{a(k,p)}{2q^2} (q^2(k^2 + p^2) - (k^2 - p^2)^2) - b(k,p) \right. \right. \\
&\left. \left. \frac{(k^2 - p^2)^2}{2q^2} (k^2 + p^2 - q^2) + \frac{c(k,p)}{2q^2} \mathcal{M}(k;\xi) ((k^2 - p^2)^2 - q^2(k^2 - p^2)) \right\} \right], \frac{\mathcal{M}(p;\xi)}{F(p;\xi)} \\
&= \frac{\alpha}{2\pi^2} \int \frac{d^3 k}{q^2} \frac{F(k;\xi)}{k^2 + \mathcal{M}^2(k;\xi)} \left[\mathcal{G}(q) \left\{ 2a(k,p)\mathcal{M}(k;\xi) - \mathcal{M}(k;\xi) \left[\frac{1}{F(k;\xi)} - \frac{1}{F(p;\xi)} \right] \Omega(k,p) \right. \right. \\
&\quad \left. \left. + \left[\frac{2b(k,p)\mathcal{M}(k;\xi) + c(k,p)}{2q^2} \right] (-q^4 + 2q^2(k^2 + p^2) - (k^2 - p^2)^2) \right\} \right. \\
&\quad \left. + \xi \left\{ a(k,p)\mathcal{M}(k;\xi) + b(k,p)\mathcal{M}(k;\xi) \frac{(k^2 - p^2)^2}{q^2} + \frac{c(k,p)}{2q^2} (k^2 - p^2)(k^2 - p^2 - q^2) \right\} \right], \\
\frac{1}{\mathcal{G}(q)} &= 1 - \frac{N_f \alpha}{2\pi^2} \int d^3 k \frac{F(k;\xi)}{k^2 + \mathcal{M}^2(k;\xi)} \frac{F(q;\xi)}{q^2 + \mathcal{M}^2(q;\xi)} \left[a(k,q)[W_1(k,p) \right. \\
&\quad \left. + W_2(k,p)\mathcal{M}(k;\xi)\mathcal{M}(q;\xi)] + b(k,q)[W_3(k,p) + W_4(k,p)\mathcal{M}(k;\xi)\mathcal{M}(q;\xi)] \right. \\
&\quad \left. - c(k,q)[W_5(k,p)\mathcal{M}(q;\xi) + W_6(k,p)\mathcal{M}(k;\xi)] \right],
\end{aligned}$$

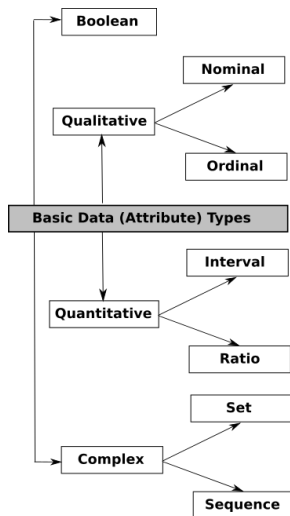
Data Types & Attributes

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”, ...



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. $D = \{1, 2, \dots, n\}$
 - $x, y \in D$ 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. $D = \{1, 2, \dots, n\}$
 - $x, y \in D$ are **ranks**
- $d(x, y) = \frac{|x-y|}{n-1}$
 $D = \{\text{worst}, \text{bad}, \text{neutral}, \text{good}, \text{best}\}$
 $d(\text{bad}, \text{good}) = \frac{|2-4|}{4} = 0.5$

Boolean attributes

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

Similarity of Attribute Values

Set attributes

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

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

Sequence attributes (strings)

- w.l.o.g. $D = \{1, 2, \dots, n\}^{<\mathbb{N}}$
- $d(x, y) = d_{x,y}(|x|, |y|)$

$$d_{x,y}(i, j) = \begin{cases} \max\{i, j\} & , \text{ if } \min\{i, j\} = 0 \\ \min \begin{cases} d_{x,y}(i-1, j) + 1 \\ d_{x,y}(i, j-1) + 1 \\ d_{x,y}(i-1, j-1) + 1_{x_i \neq y_j} \end{cases} & , \text{ otherwise} \end{cases}$$

- 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...

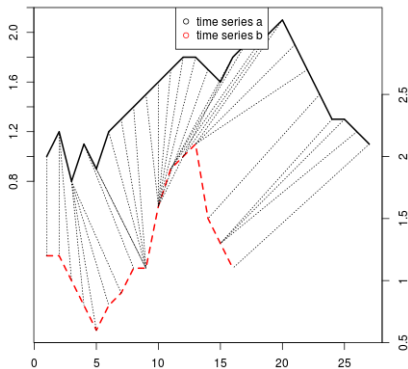
Sequence attributes (time series)

- 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!

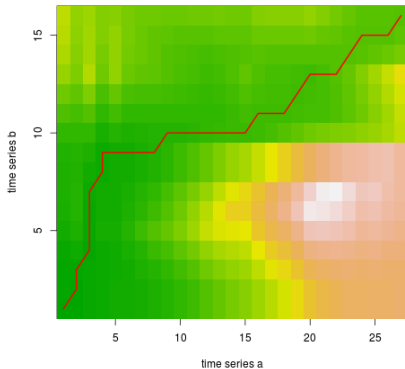
Similarity of Attribute Values

Illustration of Dynamic Time Warping

Optimal Alignment



Cost matrix & Warping path



Similarity of Attribute Values

The Basic DTW Algorithm

```
1: procedure DTW( $x = (x_1, x_2 \dots, x_p)$ ,  $y = (y_1, y_2 \dots, y_q)$ )
2:    $d_{x,y} \leftarrow \mathbb{R}^{(p+1) \times (q+1)}$  ▷ cost matrix  $d_{x,y}$ 
3:   for all  $i \in \{1, 2, \dots, p\}$  do
4:      $d_{x,y}(i, 0) \leftarrow \infty$ 
5:   for all  $j \in \{1, 2, \dots, q\}$  do
6:      $d_{x,y}(0, j) \leftarrow \infty$ 
7:    $d_{x,y}(0, 0) \leftarrow 0$ 
8:   for  $i = 1 \rightarrow p$  do
9:     for  $j = 1 \rightarrow q$  do
10:       $d \leftarrow |x_i - y_j|$  ▷ distance of  $x_i$  and  $y_j$ 
11:       $d_{x,y}(i, j) \leftarrow d + \min\{d_{x,y}(i-1, j), d_{x,y}(i, j-1), d_{x,y}(i-1, j-1)\}$ 
12:   return  $d_{x,y}(p, q)$ 
```


Object

- A collection of **recorded** 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), ...
- $\mathbf{x} = (x_1, x_2, \dots, x_m) \in D_1 \times D_2 \times \dots \times D_m$
- 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

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

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

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

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

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

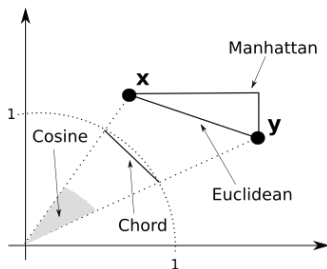
- $s(\mathbf{x}, \mathbf{y}) = \frac{a+d/2}{m}$
 - Faith's similarity

Ignoring d

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

Similarity of Numerical Instances

Objects are points in an m-dimensional Euclidean space



Cosine similarity

- $$s(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^m x_i y_i}{\left(\sum_{i=1}^m x_i^2 \sum_{i=1}^m y_i^2 \right)^{\frac{1}{2}}}$$
- 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

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

Similarity of Nominal, Ordinal and Mixed Instances

Ordinal Instances

$$\bullet s(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{m-1} \sum_{j=i+1}^m o_{ij}^{\mathbf{x}} o_{ij}^{\mathbf{y}}}{\sum_{i=1}^{m-1} \sum_{j=i+1}^m |o_{ij}^{\mathbf{x}}| |o_{ij}^{\mathbf{y}}|}$$

$$\bullet o_{ij}^{\mathbf{x}} = \begin{cases} 1 & , \text{ if } x_i > x_j \\ -1 & , \text{ if } x_i < x_j \\ 0 & , \text{ if } x_i = x_j \end{cases}$$

$$\bullet o_{ij}^{\mathbf{y}} \text{ defined as } o_{ij}^{\mathbf{x}}$$

• Goodman & Kruskal

• Be aware of the range!

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

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

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

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

Nominal Instances

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

Mixed Instances

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

• Gower's index

$$\bullet w_i = \begin{cases} 1, & \text{if } x_i \neq \text{NA} \neq y_i \\ 0, & \text{otherwise} \end{cases}$$

• $s(x_i, y_i)$ is a suitable attribute similarity measure

Clustering Methods



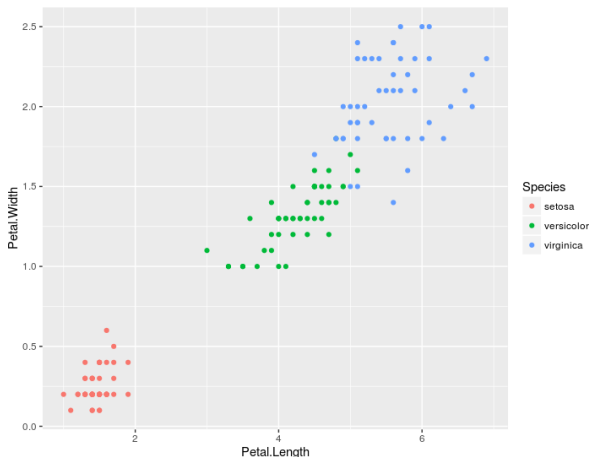
related to clustering, grouping

- 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.
- 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.
- 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

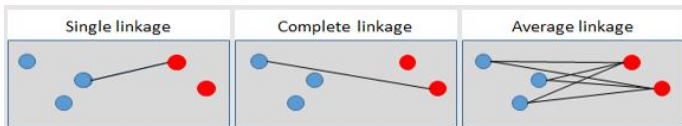
- Iris plants of the class Setosa, Versicolour, Virginica
 - 150 instances, 4 attributes
 - sepal length and width in cm, petal length and width in cm



Hierarchical Agglomerative Clustering

Given

- $D \subseteq D_1 \times D_2 \times \dots \times D_m$
- 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

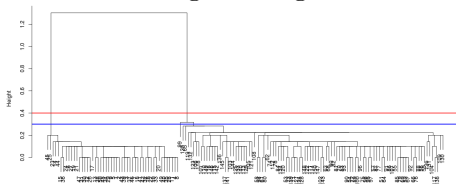
- clusterings $C_1, C_2, \dots, C_{|D|} \subset \mathcal{P}(D) \setminus \emptyset$ of objects in D such that
 - $C_1 = \{\{\mathbf{x}_1\}, \{\mathbf{x}_2\}, \dots, \{\mathbf{x}_{|D|}\}\}$
 - initially, each object is in a separate cluster
- and for each $i \in \{2, \dots, k\}$
 - $C_i = (C_{i-1} \setminus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

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

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

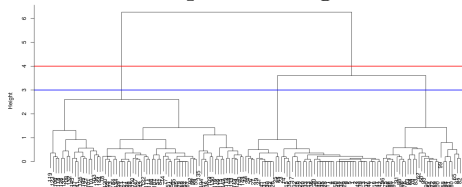
Dendrograms

Single linkage



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

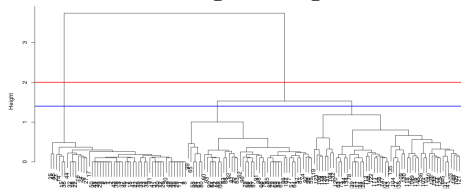
Complete linkage



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

Average linkage



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

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

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 $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_k$
- and a mapping $p : D \rightarrow \{1, 2, \dots, k\}$ such that

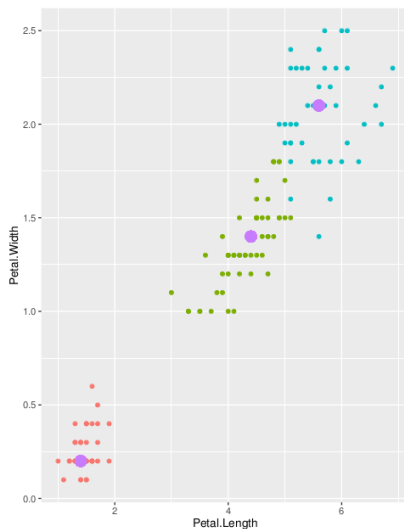
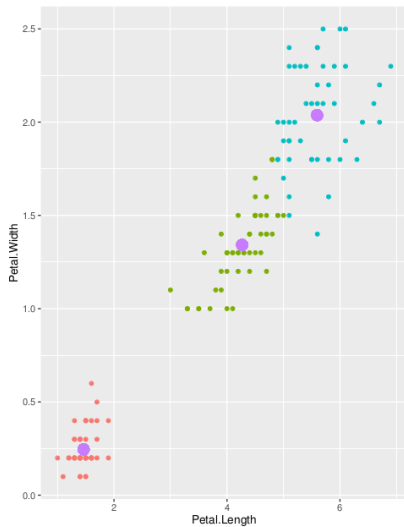
$$\sum_{i=1}^n d(\mathbf{x}_i, \mathbf{c}_{p(\mathbf{x}_i)}) \text{ is minimal}$$

k-Means Clustering

The algorithm

- ① Initialize $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_k$ such that for all $i = \{1, 2, \dots, k\}$
 - $\mathbf{c}_i \in D$ (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 $i = \{1, 2, \dots, k\}$ such that
 - $\mathbf{c}_i = \frac{\sum_{\mathbf{x}, p(\mathbf{x})=i} \mathbf{x}}{\sum_{\mathbf{x}, p(\mathbf{x})=i} 1}$
- ④ if p or \mathbf{c}_i for some $i = \{1, 2, \dots, k\}$ were changed then goto step 2
- ⑤ return p and $\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_k$

Means vs. Medoids



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**

<i>Object pairs in the same</i>		<i>Class</i>	
		Yes	No
<i>Cluster</i>	Yes	<i>a</i>	<i>b</i>
	No	<i>c</i>	<i>d</i>

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?*

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

Precision

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

Recall

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

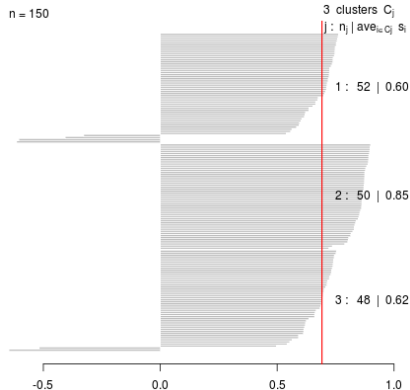
F-measure

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

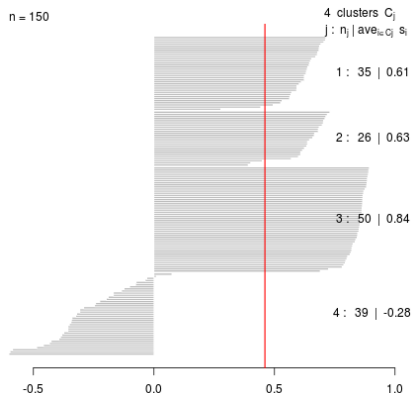
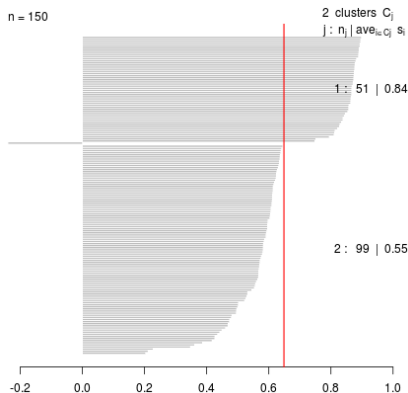
Internal Evaluation of Clusters

Silhouette

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



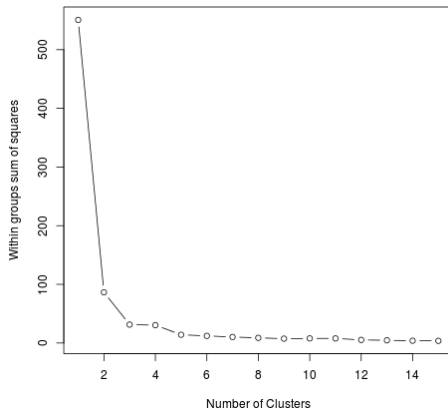
Internal Evaluation of Clusters



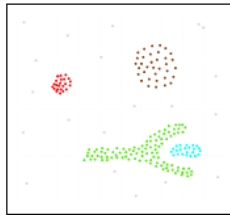
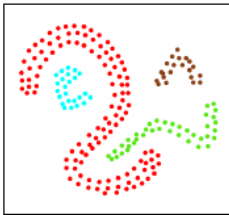
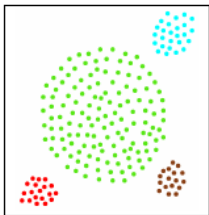
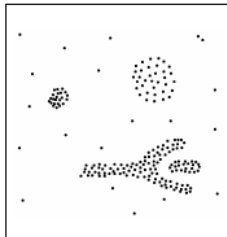
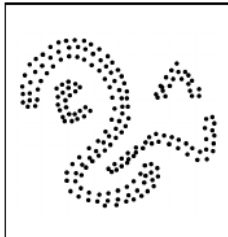
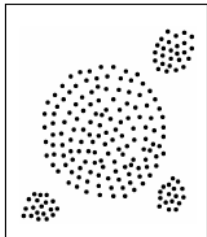
Internal Evaluation of Clusters

Within sum group of squares

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



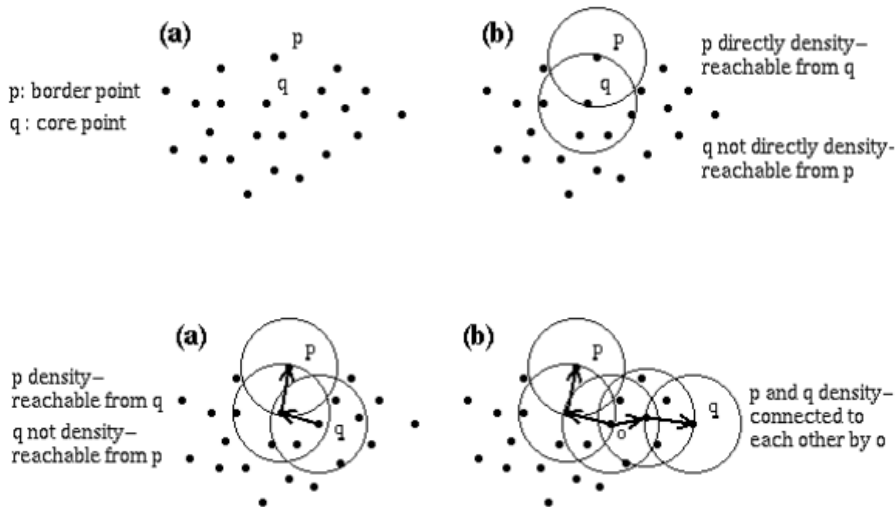
Non-convex Clusters



Neighborhood and Reachability

- ϵ -neighborhood of $\mathbf{p} \in D$ defined as $N_\epsilon(\mathbf{p}) = \{\mathbf{x} \in D \mid d(\mathbf{p}, \mathbf{x}) \leq \epsilon\}$
- \mathbf{p} is directly density-reachable from $\mathbf{q} \in D$ w.r.t. some ϵ and δ if
 - $\mathbf{p} \in N_\epsilon(\mathbf{q})$
 - $|N_\epsilon(\mathbf{q})| \geq \delta$, 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 $2 \leq i \leq n$
- \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}
- $C \subseteq D$ ($C \neq \emptyset$) 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}
- $\text{noise} = \{\mathbf{p} \in D : \mid : \mathbf{p} \notin C_1 \cup \dots \cup C_k\}$ where
 - $C_1, \dots, C_k \subseteq D$ are clusters

Neighborhood and Reachability



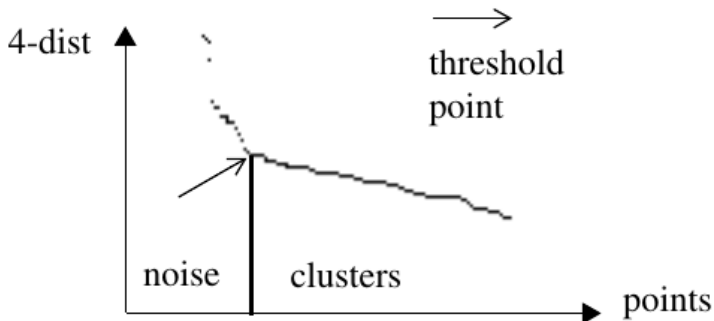
```
1: procedure DBSCAN( $D, \epsilon, \delta$ )
2:   for all  $\mathbf{x} \in D$  do
3:      $p(\mathbf{x}) \leftarrow -1$                                  $\triangleright$  mark points as unclustered
4:    $i \leftarrow 1$                                           $\triangleright$  the noise cluster have id 0
5:   for all  $\mathbf{p} \in D$  do
6:     if  $p(\mathbf{x}) = -1$  then
7:       if ExpandCluster( $D, \mathbf{p}, i, \epsilon, \delta$ ) then
8:          $i \leftarrow i + 1$ 
```

```
1: function EXPANDCLUSTER( $D, \mathbf{p}, i, \epsilon, \delta$ )
2:   if  $|N_\epsilon(\mathbf{p})| < \delta$  then
3:      $p(\mathbf{x}) \leftarrow 0$  ▷ mark  $\mathbf{p}$  as noise
4:     return false
5:   else
6:     for all  $\mathbf{x} \in N_\epsilon(\mathbf{p})$  do
7:        $p(\mathbf{x}) \leftarrow i$  ▷ assign all  $\mathbf{x}$  to cluster  $i$ 
8:      $S \leftarrow N_\epsilon(\mathbf{p}) \setminus \{\mathbf{p}\}$ 
9:     while  $S \neq \emptyset$  do
10:       $\mathbf{s} \leftarrow S_1$  ▷ Get the first point from  $S$ 
11:      if  $|N_\epsilon(\mathbf{s})| \geq \delta$  then
12:        for all  $\mathbf{x} \in N_\epsilon(\mathbf{s})$  do
13:          if  $p(\mathbf{x}) \leq 0$  then
14:            if  $p(\mathbf{x}) = -1$  then
15:               $S \leftarrow S \cup \{\mathbf{x}\}$ 
16:             $p(\mathbf{x}) \leftarrow i$ 
17:           $S \leftarrow S \setminus \{\mathbf{s}\}$ 
18:      return true
```

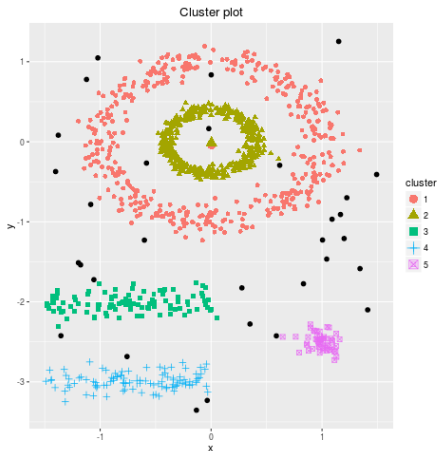

How to guess ϵ and δ ?

k -distance

- $k\text{-dist}: D \rightarrow \mathbb{R}$
- $k\text{-dist}(\mathbf{x})$ 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

- **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

- 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.

- 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`