WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Data Clustering

Marzena Kryszkiewicz

HUMAN CAPITAL HUMAN – BEST INVESTMENT! EUROPEAN UNION EUROPEAN SOCIAL FUND

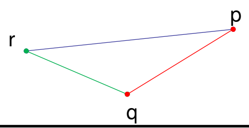
Project is co-financed by European Union within European Social Fund

Basic Notions

Distance Metric

A *distance metric* is defined as a measure that satisfies the following conditions:

- ∀p, distance(p, p) = 0;
- ∀p,q, distance(p, q) = distance(q, p);
- ∀p,q,r, distance(p, r) ≤ distance(p, q) + distance(q, r) /* triangle inequality property */.



Example Distance Metrics

- Euclidean(p, q) = $\sqrt{\sum_{i=1..n}(p_i - q_i)^2}$
- Manhattan(p, q) = $\sum_{i=1..n}|p_i - q_i|$
- Minkowski(p, q) = $\sqrt[m]{\sum_{i=1..n}|p_i - q_i|^m}$

DBSCAN: Density-Based Clustering with Noise

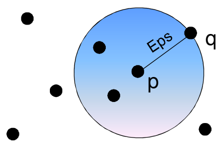
Eps-Neighborhood

- *Eps-neighborhood of a point p* (denoted by $N_{Eps}(p)$) is defined as the set of all points q in dataset D that are distant from p by no more than Eps; that is,

$$N_{Eps}(p) = \{q \in D \mid distance(p,q) \leq Eps\}.$$

Example: Eps-Neighborhood

$|N_{Eps}(p)| = 4$.



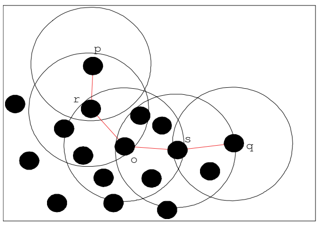
Core Points

- A point p is defined as a *core point* if its *Eps-neighborhood* contains at least MinPts points; that is, if $|N_{Eps}(p)| \geq$ MinPts.

Example: Core Points

For MinPts = 6:

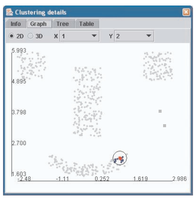
- r is a core point;
- p is not a core.

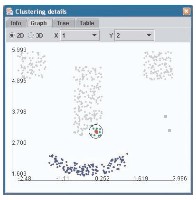


Clusters in DBSCAN…

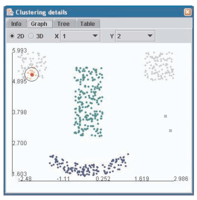
- A core point is treated as a seed from which a cluster is built.
- Whenever any core point is included in the cluster, all points in its Eps-neighborhood are also included in the cluster.

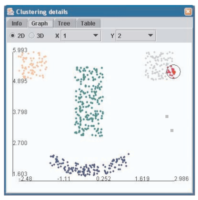
Clusters in DBSCAN…



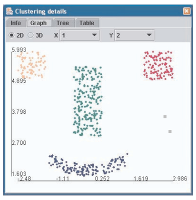


Clusters in DBSCAN…





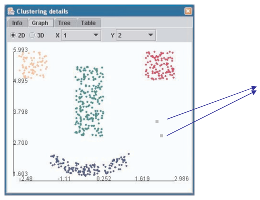
Clusters in DBSCAN



13

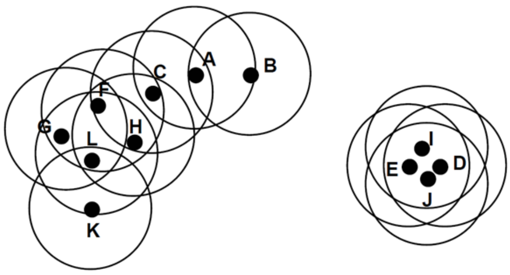
Clusters and Noise in DBSCAN

- The points that are not included in any cluster constitute *noise*.



14

DBSCAN – Definition's Illustration



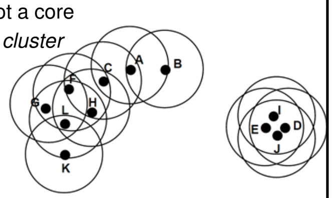
minPts = 4

15

DBSCAN – Example Execution

- Dealing with a point q from an analysed neighborhood:

if q.ClusterId = Unclassified **then** // q is not classified
  assign q *to currently created cluster*
  add q *to seeds*
else if q.ClusterId = N // q is not a core
  assign q *to currently created cluster*
else
  do nothing



minPts = 4

16

Major Challenges in DBSCAN

- Efficient calculation of Eps-neighborhood for each point.
- To this end, DBSCAN uses the R*-tree index.
- The use of such indices helps in the case of low dimensional data only.

17

TI-DBSCAN: DBSCAN with Efficient Calculation of Eps-Neighboorhoods

- Use the triangle inequality property (TI) to reduce the number of candidates for being a member of Eps-neighboorhood of a given point.

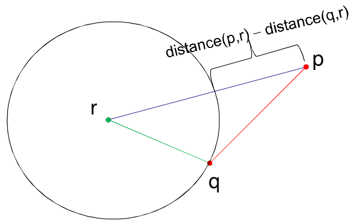
18

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

TI for pessimistic estimation of distance

For any three points p, q, r:

- distance(p,q) + distance(q,r) ≥ distance(p,r).
- **distance(p,q)** ≥ distance(p,r) – distance(q,r).



19

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

TI & Eps-Neighborhood…

Lemma. Let D be a set of points. For any two points p, q in D and any point r:

distance(p,r) – distance(q,r) > Eps ⇒

distance(p,q) ≥ distance(p,r) – distance(q,r) > Eps ⇒

by TI

$q \notin N_{Eps}(p) \wedge p \notin N_{Eps}(q)$.

20

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

TI & Eps-Neighborhood…

Theorem. Let:

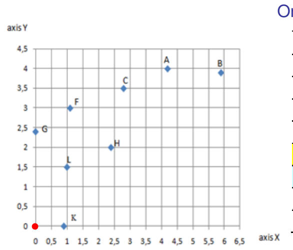
- r be any point,
- D be a set of points ordered in a non-decreasing way wrt. their distances to r;
- p be any point in D;
- q be a point following point p in D such that distance(q,r) – distance(p,r) > Eps.

Then q and all points following q in D do not belong to $N_{Eps}(p)$.

21

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: TI & Eps-Neighborhood…



Ordered set of points D; Eps = 0.2

q∈D	X	Y	distance(q,r)
K	0,9	0,0	0,9
L	1,0	1,5	1,8
G	0,0	2,4	2,4
H	2,4	2,0	3,1
F	1,1	3,0	3,2
C	2,8	3,5	4,5
A	4,2	4,0	5,8
B	5,9	3,9	7,1

$\notin N_{Eps}(F)$

22

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

TI & Eps-Neighborhood…

Theorem. Let:

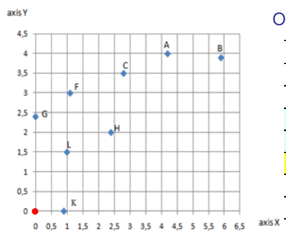
- r be any point,
- D be a set of points ordered in a non-decreasing way wrt. their distances to r;
- p be any point in D;
- q be a point preceding point p in D such that distance(p,r) – distance(q,r) > Eps.

Then q and all points preceding q in D do not belong to $N_{Eps}(p)$.

23

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: TI & Eps-Neighborhood…



Ordered set of points D; Eps = 0.2

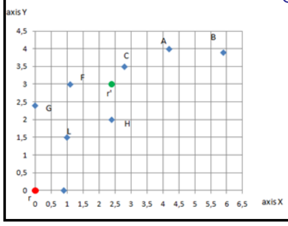
q∈D	X	Y	distance(q,r)
K	0,9	0,0	0,9
L	1,0	1,5	1,8
G	0,0	2,4	2,4
H	2,4	2,0	3,1
F	1,1	3,0	3,2
C	2,8	3,5	4,5
A	4,2	4,0	5,8
B	5,9	3,9	7,1

$\notin N_{Eps}(F)$

24

Using Many Reference Points

Example. Let r(0, 0), r'(2.4, 3.0), Eps = 0.2. Then:
distance(F,r) – distance(H,r) = 3.2 – 3.1 = 0.1 ≤ Eps.
distance(F,r') – distance(H,r') = 1.3 – 1.0 = 0.3 > Eps.
Hence, H $\notin$ N_{Eps}(p).



Ordered set of points D; Eps = 0.2

q∈D	X	Y	distance(q,r)
K	0,9	0,0	0,9
L	1,0	1,5	1,8
G	0,0	2,4	2,4
H	2,4	2,0	3,1
F	**1,1**	**3,0**	**3,2**
C	2,8	3,5	4,5
A	4,2	4,0	5,8
B	5,9	3,9	7,1

$\notin N_{Eps}(F)$

25

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

NBC: Clustering Based on k+-neighbours and Reversed k+- neighbours

- *k+-neighbourhood of a point p* (k+NN(p)) is the set of all points in D that are distant from p by no more than any k-neighbour of p.
- A reversed *k+-neighbourhood of a point p* (Rk+NN(p)) is the set of all points in D having p as their k+-neighbour,

Rk+NN(p) = {q ∈ D| p ∈ k+NN(q)}.

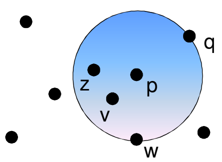
26

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: k+-neighbourhood and reversed k+-neighbourhood

Let k = 3. Then |k+-NN(p)| = 4 and:

- Point q is a k+-neighbour of point p (i.e., q∈k+NN(p)).
- Point p is a reversed k+- neighbour of q (i.e., p∈Rk+NN(q)).



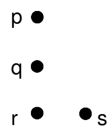
27

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: k+-neighbourhood and reversed k+- neighbourhood

Let k = 2. Then:

- k+-NN(p) = {q, r}; k+-NN(q) = {p, r}; k+-NN(r) = {q, s}; k+-NN(s) = {r, q}
- Rk+-NN(p) = {q}
- Rk+-NN(q) = {p,r,s}



28

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Clusters generated by NBC

- Density of a subspace is expressed by means of density factor NDF understood as the ration of the cardinality of k+-neighbourhood to the cardinality of reversed k+-neighbourhood :

$$\mathrm{NDF}(p) = \frac{|Rk^{+}NN(p)|}{|k^{+}NN(p)|}.$$

- Point p playes a role of a core point if NDF(p) ≥ 1.
- A core point is understood as a seed that together with its k+-neighbourhood represents a dense space, which can be regarded as a cluster or its part.

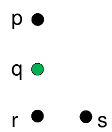
29

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: k+-NN, Rk+-NN and density factor NDF

Let k = 2. Then:

- k+-NN(p) = {q, r}; k+-NN(q) = {p, r}; k+-NN(r) = {q, s}; k+-NN(s) = {r, q}
- Rk+-NN(p) = {q}; NDF(p) = 1/2
- Rk+-NN(q) = {p,r,s}; NDF(q) = 3/2



30

TI-NBC: NBC Clustering with Efficient Calculation of k+-neighboorhood

By applying:

- multiple estimation of a decreasing radius within which k+-neighbourhood is guaranteed to be found
- triangle inequality property (TI)

in order to reduce the number of distance calculations.

31

Hierarchical Clustering

Hierarchical Clustering

- Agglomerative:
  Initially, each point is treated as a different cluster.
- Divisive:
  Initially, the whole dataset is treated as one cluster.
- Note: Both approaches use measures of *dis(similarities) between clusters.*

33

Measuring Dissimilarity of Clusters

- Single link:
  $d(C1,C2) = \min\{d(x,y) \mid x \in C1, y \in C2\}$.
- Average link:
  $d(C1,C2) = \text{avg}\{d(x,y) \mid x \in C1, y \in C2\}$.
- Complete link:
  $d(C1,C2) = \max\{d(x,y) \mid x \in C1, y \in C2\}$.
- Based on representatives of clusters:
  $d(C1,C2) = d(R1,R2)$, where
  - R1 is *a representative* (*a set of representatives*) of C1,
  - R2 is *a representative* (*a set of representatives*) of C2.

34

Iterative-Optimization Clustering

Partitioning Algorithm: k-Means

1) Randomly choose k data points in D to play a role of centroids of k distinct clusters.
2) Each point p in D assign to the cluster represented by p’s closest centroid.
3) Recompute the centroids of obtained clusters.
4) If stopping criterion is not met, go to 2.

36

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Alternative Stopping Criteria

- No point is reassigned to a different cluster.
- There is no (or minimal) change of centroids.
- In iteration j:

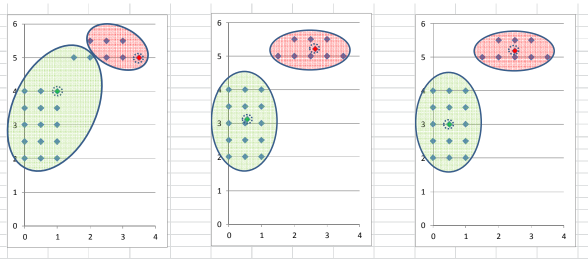
$$\frac{E_{j-1}-E_j}{E_j} < \varepsilon,$$

  - $E_j = \sum_{i=1}^{k} \sum_{X \in C_i} d(X, M_i)$, where
    C_i is the i-th cluster, M_i is the centroid of C_i;
  - ε is a user-defined threshold value.

37

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Example: Algorithm k-means



WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Scaling of Attribute Values

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Scaling of Continuous Attributes: range

Let:

- v be a value of a continuous attribute A,
- v_{min} be the least value of A,
- v_{max} be the greatest value of A.

Then:

$$range(v) = \frac{v - v_{min}}{v_{max} - v_{min}}.$$

40

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Scaling of Continuous Attributes: Z-score

Let D consist of *n* data points that have values $v_1, \ldots, v_n$, for a continuous attribute A. Then:

$$Z - score(v) = \frac{v - \mu}{s}, \text{ where}$$

- the mean for A:

$$\mu = \frac{1}{n}(v_1 + \cdots + v_n),$$

- the mean absolute deviation for A:

$$s = \frac{1}{n}(|v_1 - \mu| + \cdots + |v_n - \mu|).$$

41

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Quality of Clustering

Evaluation of Clustering

- Evaluation based on external information: calculated clusters can be compared with real clusters (e.g. determined by a knowledgeable user).
- Evaluation based on internal information.

External Evaluation of Clustering with Purity

$Purity = \frac{1}{n}\sum_{g \in G} max_{c \in C} |g \cap c|$, where

C – real clusters,

G – discovered clusters,

n – the number of points.

Example: External Evaluation of Clustering with Purity

Real clusters	Discovered clusters	Correct assignment of points to clusters
L	1	
L	3	
L	2	Yes (L)
L	2	Yes (L)
L	2	Yes (L)
H	2	
H	1	Yes (H)
H	1	Yes (H)
H	3	Yes (H)
H	3	Yes (H)

Purity = 7/10

External Evaluation of Clustering with Rand

$Rand = \frac{|TP| + |TN|}{\binom{n}{2}}$, where

- TP – the set of pairs of objects each of which is contained in some real cluster and is contained in some discovered cluster,
- TN – the set of pairs of objects each of which is neither contained in any real cluster nor is contained in any discovered cluster,
- *n* – the number of objects.

Example: External Evaluation of Clustering with Rand

Id	Real clusters	Discovered clusters
1	L	1
2	L	3
3	L	2
4	L	2
5	L	2
6	H	2
7	H	1
8	H	1
9	H	3
10	H	3

- Pairs of objects each of which is contained in some real cluster and is contained in some discovered cluster:

  TP = {(3,4), (3,5), (4,5), (7,8), (9,10).}

- Pairs of objects each of which is neither contained in any real cluster nor is contained in any discovered cluster:

  TN = {(1,6), (1,9), (1,10), (2,6), (2,7), (2,8), (3,7), (3,8), (3,9), (3,10), (4,7), (4,8), (4,9), (4,10), (5,7), (5,8), (5,9), (5,10)}.

$Rand = \frac{|TP| + |TN|}{\binom{10}{2}} = \frac{5+18}{45} \approx 0.51$

Internal Evaluation of Clustering with Davies-Bouldin

$Davies\text{−}Bouldin = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq i}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right)$, where

- n – the number of discovered clusters,
- c_k – the centroid of cluster k,
- σ_k – the average distance of points in cluster k to its centroid c_k,
- $d(c_i, c_j)$ – the distance between centroids c_i, c_j.

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

Internal Evaluation of Clustering with Silhouette Coefficient

Quality of assigning a point *i* to its cluster:

$$s(i) = \frac{b(i)-a(i)}{\max\{b(i),a(i)\}}, \text{ where}$$

- $a(i)$ – the average distance of point *i* to all other points in its cluster,
- $b(i)$ – the least average distance of point *i* to all points of a cluster that does not contain point *i*.

Quality of a cluster *C* – the average $s(i)$ over all points *i* in *C*.

Quality of clustering – the average $s(i)$ over all points *i* in the whole data set.

49

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

References…

– Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. KDD 1996: 226-231

– Jiawei Han, Micheline Kamber, Jian Pei: Data Mining: Concept and Techniques, The Morgan Kaufmann Series in Data Management Systems, 2011

50

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

References…

– Marzena Kryszkiewicz, Bartlomiej Janczak: Basic Triangle Inequality Approach Versus Metric VP-Tree and Projection in Determining Euclidean and Cosine Neighbors. Intelligent Tools for Building a Scientific Information Platform 2014: 27-49

– Marzena Kryszkiewicz, Piotr Lasek: TI-DBSCAN: Clustering with DBSCAN by Means of the Triangle Inequality. RSCTC 2010: 60-69

51

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

References…

– Marzena Kryszkiewicz, Piotr Lasek: A Neighborhood-Based Clustering by Means of the Triangle Inequality. IDEAL 2010: 284-291

– Shuigeng Zhou, Yue Zhao, Jihong Guan, Joshua Zhexue Huang: A Neighborhood-Based Clustering Algorithm. PAKDD 2005: 361-371

52

WARSAW UNIVERSITY OF TECHNOLOGY DEVELOPMENT PROGRAMME

References

–ttps://en.wikipedia.org/wiki/Cluster_analysis

–https://en.wikipedia.org/wiki/Silhouette_(clustering)

53