PLATFORMA PRZEMYSŁU 4.0

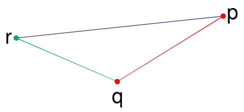
Grupowanie danych

Pojęcia wstępne

Metryka

Metryka jest funkcja odległości (*distance*) spełniająca następujące warunki:

- $\forall$p,q distance(p, q) $\geq$ 0;
- $\forall$p, distance(p, p) = 0;
- $\forall$p,q, distance(p, q) = distance(q, p);
- $\forall$p,q,r, distance(p, r) $\leq$ distance(p, q) + distance(q, r) /* nierówność trójkąta, ang. triangle inequality */.



3

Typowe miary odległości

- 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) = $\sqrt[m]{\sum_{i=1..n}|p_i - q_i|^m}$

Uwaga: Jeżeli m $\geq$ 1, to miara Minkowskiego jest metryką.

4

Grupowanie gęstościowe

Otoczenie epsilonowe

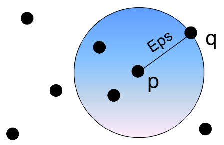
- *Otoczeniem epsilonowym* (ang. *Eps-neighborhood)* punktu p (oznaczanym jako N_{Eps}(p)) jest zbiór wszystkich punktów q w danym zbiorze D, które są odległe od punktu p o nie więcej niż Eps:

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

6

Przykład: Otoczenie epsilonowe

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



7

Punkty rdzeniowe

- Punkt p jest *punktem rdzeniowym, jeżeli* jego otoczenie epsilonowe zawiera co najmniej MinPts punktów, czyli gdy

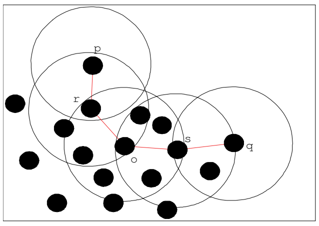
$$|N_{Eps}(p)| \geq MinPts.$$

8

Przykład: Punkty rdzeniowe

Dla MinPts = 6:

- r jest punktem rdzeniowym;
- p nie jest punktem rdzeniowym.



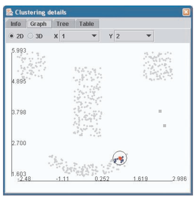
9

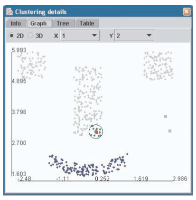
Grupy wg DBSCAN…

- Punkt rdzeniowy jest interpretowany jako ziarno, które wraz ze swoim otoczeniem epsilonowym reprezentuje gęstą przestrzeń, którą można uznać za grupę lub część grupy.
- Kiedykolwiek punkt rdzeniowy jest dołączany do grupy, wszystkie punkty w jego otoczeniu epsilonowym także są dołączane do grupy.

10

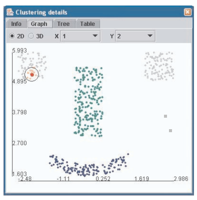
Grupy wg DBSCAN…

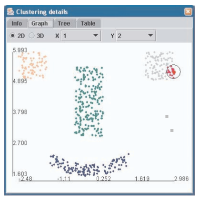




11

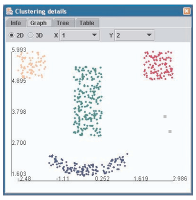
Grupy wg DBSCAN…





12

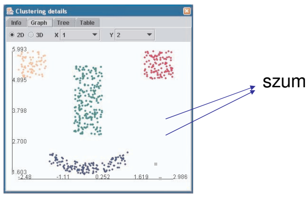
Grupy wg DBSCAN



13

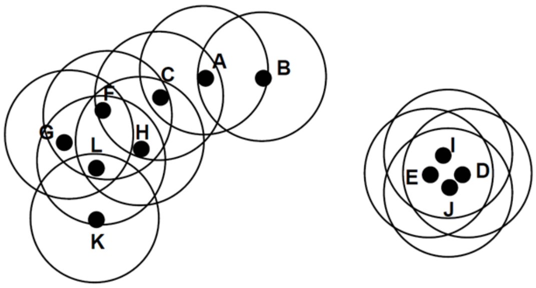
Grupy i szum w DBSCAN

- Punkty, które nie są dołączone do żadnej grupy, są traktowane jako *szum (punkty osobliwe, anomalie)*.



14

DBSCAN – ilustracja definicji

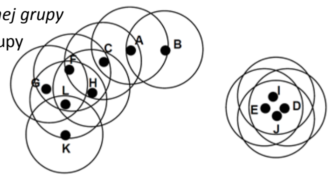


minPts = 4

DBSCAN – przykładowe wykonanie

- Postępowanie w przypadku punktu q z analizowanego otoczenia:

if q.ClusterId = Unclassified // q nie jest sklasyfikowany
przypisz q do aktualnie tworzonej grupy
dodaj q do ziaren
else if q.ClusterId = N **then** // q nie jest rdzeniem
przypisz q do aktualnie tworzonej grupy
else // q jest przypisany do grupy
nic nie rób



16

minPts = 4

Główne wyzwanie wydajnościowe w DBSCAN

- Efektywne wyznaczanie otoczeń epsilonowych dla wszystkich punktów.
- W tym celu DBSCAN używa indeksu przestrzennego R*-drzewa.
- Użycie tego typu indeksów jest pomocne wyłącznie w przypadku danych o małej liczbie wymiarów.

17

TI-DBSCAN: Grupowanie DBSCAN z efektywnym wyznaczaniem sąsiedztwa epsilonowego punktu p

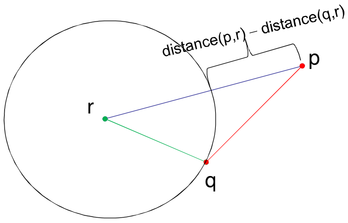
- poprzez zastosowanie nierówności trójkąta (ang. the triangle inequality property, TI) w celu wyeliminowania konieczności wyznaczania odległości od punktu p do wszystkich punktów w zbiorze D.

18

Nierówność trójkąta jako pesymistyczne oszacowanie odległości

Dla dowolnych punktów p, q, r:

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



19

TI i sąsiedztwo epsilonowe…

Lemat. Niech D będzie zbiorem punktów. Dla dowolnych punktów p, q w D i dowolnego punktu r:

distance(p,r) – distance(q,r) > Eps ⇒

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

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

by TI

20

TI i sąsiedztwo epsilonowe…

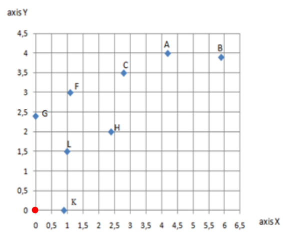
Twierdzenie. Niech:

- r będzie dowolnym punktem,
- D będzie zbiorem punktów uporządkowanych niemalejąco względem odległości do punktu r;
- p będzie dowolnym punktem w D;
- q będzie takim punktem występującym po punkcie p w D, że distance(q,r) – distance(p,r) > Eps.

Wtedy q i wszystkie punkty występujące po punkcie q w D nie należą do $N_{Eps}(p)$.

21

Przykład: TI i sąsiedztwo epsilonowe…



Uporządkowany zbiór punktów 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

TI i sąsiedztwo epsilonowe…

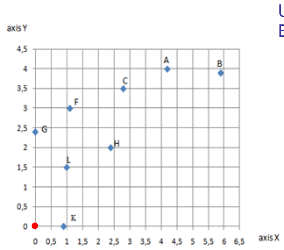
Twierdzenie. Niech:

- r będzie dowolnym punktem,
- D będzie zbiorem punktów uporządkowanych niemalejąco względem odległości do punktu r;
- p będzie dowolnym punktem w D;
- q będzie takim punktem występującym przed punktem p w D, że distance(q,r) – distance(p,r) > Eps.

Wtedy q i wszystkie punkty występujące przed punktem q w D nie należą do $N_{Eps}(p)$.

23

Przykład: TI i sąsiedztwo epsilonowe



Uporządkowany zbiór punktów 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

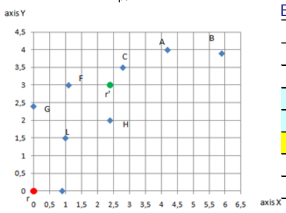
Użycie wielu punktów referencyjnych

Przykład. Niech r(0, 0), r'(2.4, 3.0), Eps = 0.2. Wtedy:

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.

Stąd, H $\notin N_{Eps}(p)$.



Uporządkowany zbiór punktów 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

$\}\notin N_{Eps}(F)$ (K, L, G)

25

NBC: Grupowanie ze względu na k^+-sąsiadów i odwrotnych k^+-sąsiadów…

- *k^+-sąsiedztwo punktu p* (k^+NN(p)) jest zbiorem wszystkich punktów w zbiorze D różnych od p, których odległość do punktu p nie przekracza odległości jej dowolnego najdalszego k-tego sąsiada do punktu p.
- *Odwrotne k^+-sąsiedztwo punktu p* (Rk^+NN(p)) jest zbiorem punktów w zbiorze D, dla których p jest k^+-sąsiadem,

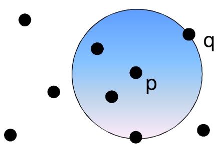
$$\mathrm{Rk^+NN(p)} = \{q \in D | p \in \mathrm{k^+NN(q)}\}.$$

26

Przykład: k^+-sąsiedztwo

Niech k = 3. Wtedy:

- |k^+-NN(p)| = 4.
- Punkt q jest k^+-sąsiadem punktu p (czyli, q∈k^+NN(p)).

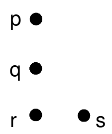


27

Przykład: k^+-sąsiedztwo i odwrotne k^+-sąsiedztwo

Niech k = 2. Wtedy:

- 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

Grupy wg algorytmu NBC

- Gęstość podprzestrzeni jest wyrażana za pomocą *współczynnika gęstości* NDF rozumianego jaki stosunek liczności k^+-sąsiedztwa do liczności odwrotnego k+-sąsiedztwa:

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

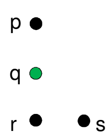
- Punkt p pełni rolę punktu rdzeniowego, jeżeli NDF(p) ≥ 1.
- Punkt rdzeniowy jest interpretowany jako ziarno, które wraz ze swoim k^+-sąsiedztwiem reprezentuje gęstą przestrzeń, którą można uznać za grupę lub część grupy.
- Kiedykolwiek punkt rdzeniowy jest dołączany do grupy, wszystkie punkty w jego k^+-sąsiedztwie także są dołączane do grupy.

29

Przykład: k^+-NN, Rk^+-NN i współczynnik gęstości NDF

Niech k = 2. Wtedy:

- 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) = ½; p nie jest punktem rdzeniowym
- Rk^+-NN(q) = {p,r,s}; NDF(q) = 3/2; q jest punktem rdzeniowym



30

TI-NBC: Grupowanie NBC z efektywnym wyznaczaniem sąsiedztwa epsilonowego punktu p

Poprzez zastosowanie:

- wielokrotnego szacowania minimalnego promienia, gwarantującego znalezienie k+-sąsiedztwa punktu w jego otoczeniu o tym promieniu

oraz

- nierówności trójkąta (ang. the triangle inequality property, TI)

w celu wyeliminowania konieczności wyznaczania odległości od punktu p do wszystkich punktów w zbiorze D.

31

Grupowanie hierarchiczne

Grupowanie hierarchiczne

- Aglomeracyjne (ang. agglomerative) zwane także wstępującym:
  Początkowo każdy punkt danych jest traktowany jako osobna grupa. Następnie najbliższe sobie grupy są iteracyjne łączone w większe grupy.
- Podziałowe (and.divisive) zwane także zstępującym:
  Początkowo cały zbiór danych punkt jest traktowany jako jedna grupa. Następnie grupy są dzielone na mniejsze podgrupy.
- Uwaga: Obydwa podejścia stosują miary *nie*(*podobieństwa*) *pomiędzy grupami.*

33

Miary niepodobieństwa grup

- Pojedyncze połączenie (ang. single link):
  $d(C1,C2) = \min\{d(x,y) \mid x \in C1, y \in C2\}$.
- Średnie połączenie (ang. average link):
  $d(C1,C2) = \text{avg}\{d(x,y) \mid x \in C1, y \in C2\}$.
- Całkowite połączenie (ang. complete link):
  $d(C1,C2) = \max\{d(x,y) \mid x \in C1, y \in C2\}$.
- Oparte na reprezentantach grup:
  $d(C1,C2) = d(R1,R2)$, gdzie
  - R1 jest *reprezentantem* (*zbiorem reprezentantów*) grupy C1,
  - R2 jest *reprezentantem* (*zbiorem reprezentantów*) grupy C2.

34

Grupowanie iteracyjno-optymalizacyjne

Podstawowy algorytm k-średnich (ang. k-means)

1) Losowo wybierz k punktów danych w D, które będą pełnić rolę reprezentantów k różnych grup.
2) Każdy punkt p w D przypisz do grupy wskazywanej przez reprezentanta najbliższego punktowi p.
3) Wyznacz reprezentantów uzyskanych grup jako ich środki geometryczne.
4) Jeśli kryterium stopu nie jest spełnione, ich do punktu 2.

36

Alternatywne warunki stopu

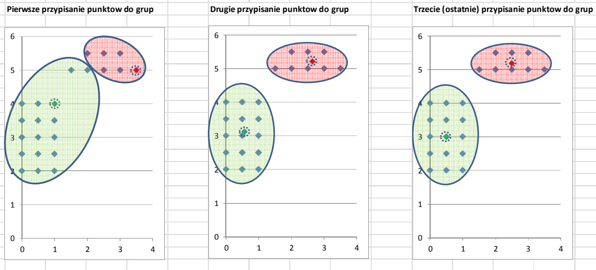
- Żaden punkt nie został przypisany do innej grupy.
- Reprezentanci grup nie uległi zmianie lub uległi nieznacznej zmianie.
- W iteracji *j* spełniony jest warunek:

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

  - $E_j = \sum_{i=1}^{k} \sum_{p \in C_i} d(p, M_i)$, gdzie:
    C_i jest *i*-tą grupą, a M_i jest jej reprezentantem;
  - ε jest wartością progową zdefiniowaną przez użytkownika.

37

Przykład: Algorytm k-średnich



Skalowanie atrybutów

Skalowanie atrybutów ciągłych: range

Niech:

- v będzie wartością atrybutu ciągłego A,
- v_{min} będzie najmniejszą wartością atrybutu A,
- v_{max} będzie najmniejszą wartością atrybutu A.

Wtedy:

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

40

Skalowanie atrybutów ciągłych: Z-score

Niech D składa się z *n* punktów danych o wartościach $v_1, \ldots, v_n$ dla ciągłego atrybutu A. Wtedy:

$$Z - score(v) = \frac{v-\mu}{S}, \text{ gdzie}$$

- średnia dla A:

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

- odchylenie przeciętne dla A:

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

41

Ewaluacja grupowania

Ewaluacja grupowania

- Ewaluacja zewnętrzna: odkryte grupy są porównywane z wzorcowymi grupami (np. określonymi przez eksperta).
- Ewaluacja wewnętrzna: odkryte grupy nie są porównywane z wzorcowymi grupami.

43

Zewnętrzna miara ewaluacji grupowania Purity – wskaźnik czystości

Miara ewaluacji grupowania Purity:

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

- *C* – klasy (decyzyjne),
- *G* – grupy,
- *n* – liczba obiektów.

44

Przykład: Zewnętrzna miara ewaluacji grupowania – Purity

Id	Wzorcowe grupy	Odkryte grupy	Prawidłowe przypisanie punktów do grup
1	L	1	
2	L	3	
3	L	2	Tak (L)
4	L	2	Tak (L)
5	L	2	Tak (L)
6	H	2	
7	H	1	Tak (H)
8	H	1	Tak (H)
9	H	3	Tak (H)
10	H	3	Tak (H)

Purity = 7/10

45

Zewnętrzna miara ewaluacji grupowania Rand

Miara ewaluacji grupowania Rand:

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

- TP – zbiór par obiektów, z których każda jest zawarta w pewnej wzorcowej grupie i jest zawarta w pewnej odkrytej grupie,
- TN – zbiór par obiektów, z których każda nie jest zawarta w żadnej wzorcowej grupie i nie jest zawarta w żadnej odkrytej grupie,
- *n* – liczba obiektów.

46

Przykład: Zewnętrzna miara ewaluacji grupowania – Rand

Id	Wzorcowe grupy	Odkryte grupy
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

- Pary obiektów, z których każda jest zawarta w pewnej wzorcowej grupie i jest zawarta w pewnej odkrytej grupie:

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

- Pary obiektów, z których każda nie jest zawarta w żadnej wzorcowej grupie i nie jest zawarta w żadnej odkrytej grupie:

  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)}.

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

47

Wewnętrzna miara ewaluacji grupowania Davies-Bouldin

Miara ewaluacji grupowania Davies-Bouldin:

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

- *n* – liczba odkrytych grup,
- c_k – centroid *k*–tej grupy,
- σ_k – średnia odległość elementów *k*-tej grupy do jej centroidu c_k,
- $d(c_i, c_j)$ – odległość pomiędzy centroidami c_i, c_j.

48

Wewnętrzna miara ewaluacji grupowania Silhouette – wskaźnik sylwetkowy

- Ewaluacja grupowania dla punktu *i*:

$$S(i) = \frac{b(i)-a(i)}{\max\{b(i),a(i)\}}, \text{ gdzie}$$

  - *a*(*i*) – średnia odległość od punktu *i* do innych punktów w grupie punktu *i*,
  - *b*(*i*) – najmniejsza średnia odległość od punktu *i* do wszystkich punktów grupy, która nie zawiera punktu *i*.

- Ewaluacja dla grupy – średnia wartość *S*(*i*) dla punktów tej grupy.
- Ewaluacja grupowania całego zbioru danych – średnia wartość *S*(*i*) dla wszystkich punktów zbioru danych.

49

Literatura…

- 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
- Jacek Koronacki, Jan Ćwik: Statystyczne systemy uczące się, Akademicka Oficyna Wydawnicza EXIT
- 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

50

Literatura…

- Marzena Kryszkiewicz, Piotr Lasek: TI-DBSCAN: Clustering with DBSCAN by Means of the Triangle Inequality. RSCTC 2010: 60-69
- Marzena Kryszkiewicz, Piotr Lasek: A Neighborhood-Based Clustering by Means of the Triangle Inequality. IDEAL 2010: 284-291
- Tadeusz Morzy, Eksploracja danych: Metody i algorytmy, Wydawnictwo Naukowe PWN, 2013
- Shuigeng Zhou, Yue Zhao, Jihong Guan, Joshua Zhexue Huang: A Neighborhood-Based Clustering Algorithm. PAKDD 2005: 361-371

51

Literatura

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

52

Ćwiczenia…

1. Niech minPts = 3, a parameter odległości ε = 1.2. Wyznacz grupy i szum, które zostałyby znalezione przez algorytm DBSCAN w zbiorze danych z poniższego rysunku, przy założeniu, że stosowana jest Euklidesowa miara odległości.

								•	•	
			•				•	•	•	
	•		•					•		
•		•					•	•		
•	•		•							

Ćwiczenia…

2. Niech k = 2. Wyznacz grupy i szum, które zostałyby znalezione przez algorytm DBSCAN w zbiorze danych z poniższego rysunku, przy założeniu, że stosowana jest Euklidesowa miara odległości.

								•	•	
			•				•	•	•	
	•		•					•		
•		•					•	•		
•	•		•							

Ćwiczenia

3. Dotyczy użycia nierówności trójkąta przy określaniu ε-otoczenia w algorytmie TI-DBSCAN: Niech D będzie zbiorem 2-wymiarowych punktów jak w poniższej tabeli, uporządkowanych według ich odległości Euklidesowych do pewnego punktu referencyjnego r.

   Niech ε = 0.6, a A będzie punktem, którego ε-otoczenie wyznaczono za pomocą nierówności trójkąta.

punkt q	X	Y	odległość(q, r)
K	0.9	0.0	0.9
L	1.0	1.5	1.9
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	5.0
A	4.2	4.0	5.8
B	5.9	3.9	6.1

- Dla których punktów różnych od A wykonywano pesymistyczne oszacowanie ich Euklidesowych odległości do punktu A?
- Dla których punktów różnych od A wyznaczano ich rzeczywiste, Euklidesowe odległości do punktu A?