Carnegie Mellon University
Department of Computer Science
15-826 Multimedia Databases and Data Mining
C. Faloutsos, Fall 2024

Homework 3 - Solutions

Due: **pdf, on canvas, at 2:00pm, on 11/01/2024**

VERY IMPORTANT:

- Upload **e-copy** of your answers, on canvas.

Reminders:

- *Plagiarism*: Homework is to be completed *individually*.
- *Typeset* your answers. Illegible handwriting may get zero points.
- *Late homeworks*: Follow the published policy

For your information:

- Explanations are *optional*, and will only be used to for partial credit, if the main answer is off.
- Graded out of **100** points; **5** questions total
- Rough time estimate: *1-2 hours (≈ 10-20 minutes per question)*

Revision : 2024/11/02 20:20

Question	Points	Score
R-trees	10	
Fractals - warm-up	20	
Fractals - T/F	20	
Correlation integrals	25	
Power laws	25	
Total:	100	

Question 1: R-trees ....[10 points]

A parent node in an R-tree of d=4 dimensions, has an MBR with sizes x_1=0.1, x_2=0.2, x_3=0.3 and x_4=0.4.

Remember: Explanations are *optional*, and will only be used to for partial credit, if the main answer is off.

(a) **[10 points]** What is the probability it will be retrieved by a point query? (Assume uniform distribution of queries; assume the address space is the unit cube).

(a) **0.0024**

Solution: the product: $x_1 * x_2 * x_3 * x_4 = 0.0024$

Question 2: Fractals - warm-up .... [20 points]

Remember: Explanations are *optional*, and will only be used to for partial credit, if the main answer is off.

(a) **[10 points]** Consider 1M points on the Koch snowflake (see course lectures, slides #13-17, or wikipedia). Suppose that point P has n_1=10 neighbors within r_1=0.001 unit of distance. How many neighbors do you expect it to have, when we triple the radius, that is, how many neighbors within r_2=3*r_1 units of distance?

(a) **40**

Solution: n_2=40: triple the distance, quadruple the neighbors, for (most) of the points on the Koch curve.

Grading info: *full points for nearby approximations (eg., 39)*

(b) **[10 points]** Same as above, but with 1M points on the Cantor Dust (see course lectures, slides #19-20): Point Q has m_1=10 neighbors within r_1=0.001. How many neighbors do you expect it to have, within r_2=3*r_1 units of distance?

(b) **20**

Solution: Triple the distance, double the neighbors.

Grading info: *again, full points for nearby approximations.*

Question 3: Fractals - T/F .... [20 points]

Mark your response (Yes or No), for the following statements. No need for explanations. As in the class, by 'fractal dimension' we mean the 'correlation fractal dimension'.

Is it possible that a cloud of points in $E(\geq 1)$ dimensions may have:

(a) [**2 points**] fractal dimension equal to E?
■ **Yes** □ No

(b) [**2 points**] fractal dimension equal to $E - 1$?
■ **Yes** □ No

(c) [**2 points**] fractal dimension equal to $E/2$?
■ **Yes** □ No

(d) [**2 points**] fractal dimension equal to $E + 1$?
□ Yes ■ **No**

(e) [**2 points**] fractal dimension = 1, *without* the dataset being points from a straight line?
■ **Yes** □ No

(f) [**2 points**] fractal dimension < 1?
■ **Yes** □ No

(g) [**2 points**] negative fractal dimension?
□ Yes ■ **No**

(h) [**2 points**] fractal dimension equal to zero?
■ **Yes** □ No

(i) [**2 points**] fractal dimension = 0.5 (a rational number)?
■ **Yes** □ No

(j) [**2 points**] *several, different* fractal dimensions, in different ranges of scales?
■ **Yes** □ No

Question 4: Correlation integrals .... [25 points]

Consider some clouds of points, some of which are shown in Figure 1. We want to describe their correlation integrals. No need for explanations.

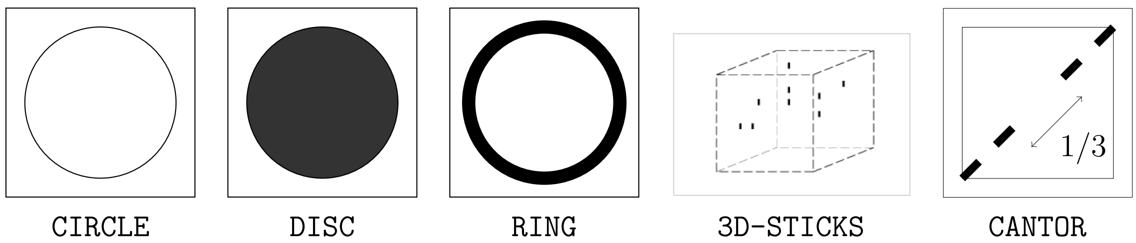

Figure 1: Clouds of points: for `3D-STICKS` we show a few 'sticks'; for `CANTOR`, we show the result after the first 2 iterations of eliminating the *middle third*

Each cloud has N points (say, N=100,000 - the exact value is not needed). The descriptions of the clouds follow. For each cloud, list the slopes (left-to-right). Figure 2 gives reminders of how correlation integrals look like.

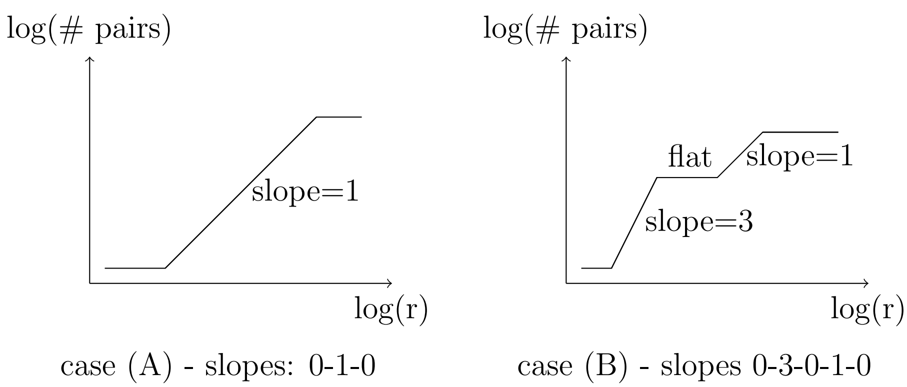

Figure 2: Illustrations of some correlation integrals and their slopes

(a) **[1 point]** `LINE`: points uniformly distributed on the major diagonal of the unit square (zero thickness)

(a) **0,1,0**

(b) **[3 points]** `CIRCLE`: points uniformly distributed along the periphery of the unit circle (zero thickness)

(b) **0,1,0**

(c) **[3 points]** `DISC`: points uniformly distributed in the unit circle

(c) **0,2,0**

(d) **[6 points]** `RING`: points uniformly distributed inside a ring of radius=1 and width w=0.01 (that is, `RING` is like '`CIRCLE`', but with non-zero thickness).

(d) **0,2,1,0**

(e) **[6 points]** `3D-STICKS`: points on 1,000 line segments; each segment is vertical, with height 10^{-4} (and obviously, zero width); the segment-centers are uniformly distributed in the unit **cube**.

Grading info: -2pt if they miss the plateau, ie if they give 0-1-3-0

(e) **0,1,0,3,0**

(f) **[6 points]** `CANTOR`: points on the (zero width) diagonal line of the unit square, after eliminating the middle-third, recursively. In figure 1, we show the first two iterations of the above procedure.

(f) **0, 0.63, 0**

Grading info: slope = log(2)/log(3) = 0.6309

Grading info: -2pt if linear extra slope, like 0-1-0-0.63 - 0

Question 5: Power laws .... [25 points]

We are told that some country has N=10^7 citizens, and that their income X follows a power law:

$$P(X \geq x) = x^{-1} \quad x \geq X_{min} = 1 \tag{1}$$

That is, the CCDF (complementary cummulative distribution function) is a power law with exponent −1, and the lowest income X_{min} is 1.

Remember: Explanations are *optional*, and will only be used to for partial credit, if the main answer is off.

(a) [**5 points**] What is median income?

(a) **2**

Solution: $x_{median}^{-1} = 1/2$

(b) [**10 points**] What is your estimate for the income X_{max} of the richest citizen?

(b) **10M**

Solution: $1/X_{max} = 1/N$, since there is only 1 person out of N that has greater or equal to X_{max}

(c) [**5 points**] What is your estimate for the income X_2 of the second richest person?

(c) **5M**

Solution: $1/X_2 = 2/N$, since there is only 2 persons out of N

(d) [**5 points**] From the lectures, we know that the pdf (probability density function) of the income $p(x) \propto x^{\beta}$, is also a power law. What is the value of the exponent β?

(d) **-2**

Solution: We know that if the CCDF has slope $-\alpha$, the pdf has one-less for slope: $-\alpha - 1$

Grading info: *-1 if they just say '2'*

Grading info: *no penalty if they just say '2', but give the correct formula for the pdf*