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

Homework 4

Due: **pdf, on canvas, at 2:00pm, on 11/22/2024**

VERY IMPORTANT:

- Upload **e-copy** of your answers, on canvas.
- Time estimate: **HEAVY: 15-30 hours** ($\approx$ 1 hour for Q1; 5-10 hours per programming question)

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.
- No need to provide your code.
- Graded out of **100** points; **4** questions total

Revision : 2024/11/07 18:17

Question	Points	Score
Density paradox	10	
Mean-median paradox	30	
Similarities and SVD	30	
Fourier	30	
Total:	100	

Question 1: Density paradox .....[10 points]

See Figure 1: Consider a cloud of N=1,000 2-d points uniformly distributed across the diagonal of the unit square. Also consider the point $\mathcal{P}$= (0.5, 0.5) in the middle. Compute the density of the cloud around $\mathcal{P}$.

Use squares of side s ($s = 1, 1/2, 1/4$), centered around $\mathcal{P}$. Define as:

- N_s: the estimated number of points inside the square
- $D(s) = N_s/s^2$: the density estimate

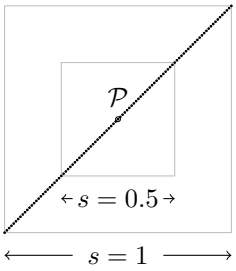


Figure 1: Compute the density around $\mathcal{P}$

.

(a) **[3 points]** Compute the density $D(1)$ for the square of side s=1 (centered at point $\mathcal{P}$)

(a) _____

(b) **[3 points]** : Repeat for $D(0.5)$ (side $s = 0.5$)

(b) _____

.....
.....
.....
.....

(c) **[3 points]** : Repeat, for side $s = 1/4$

(c) _____

(d) **[0 points]** Why is the density fluctuating?

.....
.....
.....
.....

(e) **[1 point]** What should you do if this happens in a real dataset you are studying? (Mark all the answers that you agree with)

A. Compute the (local) fractal dimension instead

B. Fix the side s to a reasonable value.

C. Give the limit of the density $D(s)$ as $s \to 0$

D. none of the above

Question 2: Mean-median paradox ..... [30 points]

(*This is from a real story:*) Suppose you are interning at a financial institution, and your mentor would like you to study the statistics of the amounts of their financial transactions, and specifically the average.

The file `amounts1M.csv.gz` in the folder `http://www.cs.cmu.edu/~christos/courses/826.F24/HOMEWORKS/hw4-data/` mimics this scenario: It has $N = 2^{20}$ lines (plus a header line), each data line corresponding to a transaction, and each with a single number that stands for the amount of the transaction.

Write code to estimate the mean $m(n)$ for the first n transactions (n=2^{10}, 2^{11}, ..., 2^{20}=N); and also to estimate the median $\mu(n)$ for the same settings.

(a) **[10 points]** Give the plots for $m(n)$ and $\mu(n)$

(b) **[10 points]** You suspect that the data may follow a Pareto distribution. Plot the CCDF (= Prob($X \geq x$)) of the full dataset, in log-log scales.

(c) **[5 points]** The CCDF plot seems to confirm that a Pareto distribution fits well - you conjecture that the formula is:

$$CCDF(x) \equiv \text{Prob}(X \geq x) = 1/x \quad (x \geq 1) \tag{1}$$

Compute the theoretical median μ, from Eq. 1.

(c) _____

(d) **[5 points]** Compute the theoretical average m, as the sample size $n \to \infty$

(d) _____

(e) **[0 points]** What would you report to your mentor?

A. Just report the average for the $N=2^{20}$ given amounts
B. Suggest that you also compute the variance
C. Suggest a deep-dive, to make sure that the amounts follow a Pareto distribution
D. Suggest that you use the median instead

Question 3: Similarities and SVD . . . . . [30 points]

(*Also based on (multiple) true stories:*) Suppose you have N=25 objects (eg., documents) and you are given all the cosine similarities in an $N \times N$ similarity matrix. Such a matrix is available as `similarities25.csv` at `http://www.cs.cmu.edu/~christos/courses/826.F24/HOMEWORKS/hw4-data/`

Clearly, the diagonal is all '1'. You want to visualize the dataset, and specifically, you want to find and scatter-plot the best 2-dimensional points on the unit circle, that will produce close similarity scores.

(a) **[10 points]** Do the SVD to the similarity matrix. What is its effective rank r? ('effective' means that you can ignore the small singular values, say ≤ 0.001)

(a) _____

(b) **[5 points]** How many dimensions d would you need to find N=25 points that produce the same similarity matrix?

(b) _____

(c) **[10 points]** Find N=25 points in 2-d, whose cosine similarities are as close to the original similarity matrix as possible; plot them as a scatterplot, like the one of Figure 2(a). Notice that there are multiple correct solutions (since rotation and mirroring preserve angles).

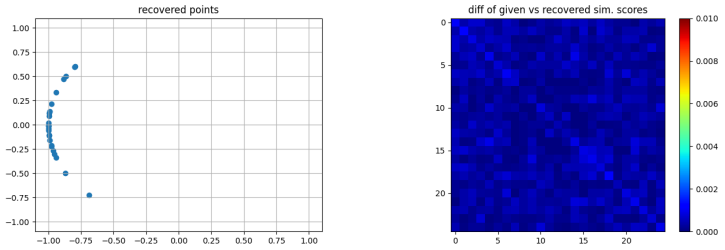
. . . . . .

. . . . . .

. . . . . .

. . . . . .

(d) **[5 points]** Plot (eg, `imshow()` of python) the $N \times N$ matrix of the differences between the original similarities, and the recovered ones - something like Figure 2(b)



(a) example of scatterplot (b) example of diff matrix

Figure 2: Visualizations: (a) of the recovered 2-d points; and (b) of the differences between original and recovered similarities

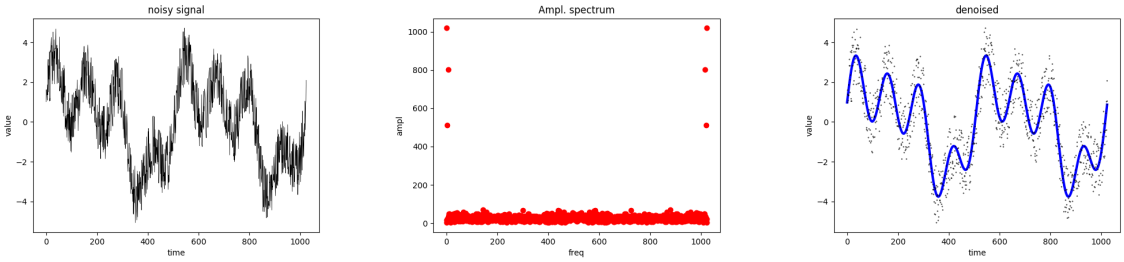
Question 4: Fourier ..... [30 points]

(*Denoising is a popular task in signal processing*). Suppose you have a noisy time sequence $s(t)$ of duration N, like in Figure 3(a). It could be an intercepted intelligence signal; or the sales of a product over time; or the sound of a honeybee (main pollinators for our agriculture), or light from a star in astrophysics.

You suspect there are periodicities, of the form:

$$\begin{aligned} s(t) \quad = \quad & A_1 \sin(2\pi f_1 t/N + \phi_1) + \\ & \dots \\ & A_k \sin(2\pi f_k t/N + \phi_k) + \\ & noise(t) \end{aligned} \tag{2}$$

You want to find how many (k) such frequencies are there, as well as their parameters: Amplitudes A_i, frequenceies f_i, and phases ϕ_i ($i = 1, \dots, k$)



(a) noisy signal (b) its Fourier spectrum (c) denoised signal

Figure 3: Example of a signal and its plots.

The signal you are to work on is at `sounds1024.csv` in the folder `http://www.cs.cmu.edu/~christos/courses/826.F24/HOMEWORKS/hw4-data/`. It consists of 1024 lines, each having a value. There is no header.

(a) **[10 points]** Compute and plot the amplitude spectrum, as in Figure 3(b).

(b) **[5 points]** Spot the highest amplitude(s); how many are high, in your plot? (in the sample plot of Figure 3(b), one would argue that there are k=3 dominant frequencies).

(b) _____

(c) **[10 points]** Give the amplitude, frequency, and phase, for each of the dominant components.

.....

.....

.....

.....

(d) **[5 points]** Drop all the other frequencies, and plot your de-noised signal (blue curve) along with the dots of the original signal as in Figure 3(c).

Hint: Use an existing library, like scipy fft. Any other package (matlab, octave, R), is fine.