OM-S20: Manifold Learning Problem Set

C. V. Jawahar

IIIT Hyderabad

Submission URL: `https://forms.gle/Xh1zHjhuifEtiw76A`

07 Apr 2020

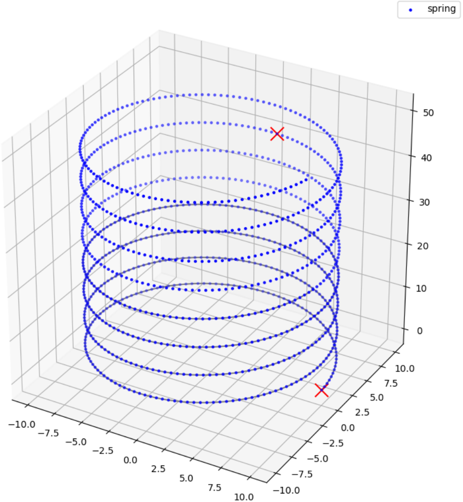
Context for Q1, Q2

Consider that we have N=1000 points on a spring. A spring can be expressed parametrically as

$$(x, y, z) = (R\cos(t), R\sin(t), Ht)$$

Let $R = 10, H = 1, t \in \{0, 0.05, 0.1, 0.15, .., 50\}$
Starter python code to generate data can be downloaded from here:
`https://drive.google.com/file/d/1KhQrqtxkbTBWVQjYOXpH78WuL-d2AzVQ/`.
This code outputs (x, y, z) coordintes of 1000 points.

Context for Q1, Q2



Q1

Find the graph distance (distance on the nearest neighbor graph) between points $X[0]$ and $X[784]$ for K=1, K=2, K=3, and report these numbers. You need only to give the three distances.

Q2

For $X[0]$ and $X[784]$, report

(A) Distance on the nearest neighbour graph when K=5

(B) True distance on the manifold (you may have to analytically compute it)

(C) Euclidean distance between these two points

You need only to give the three distances.

Q3

We had seen the eigen vector based optimization at many places. In fact, it can come in 10 different forms, in practice.

Reference: `https://arxiv.org/pdf/1903.11240.pdf` summarizes first five in Sec 3 (0 to 4) next 5 in Sec 4 (5 to 9).

Elaborate the formulation corresponding to your roll-number mod 10. Fill any missing details, elaborate steps. Find an application problem where this formulation appears.

(your write up should not be more than one page as PNG.)

Q4

We had seen PCA. A related dimensionality reduction technique is Fisher Discriminant Analysis. See Sec 5.2.2 of Reference:
`https://arxiv.org/pdf/1903.11240.pdf`
Explain the objective. Provide intuition and then formulate and derive the solution.
(your write up should not be more than one page as PNG.)

Q5

Write the pseudocode for LLE.
(your write up should not be more than one page as PNG.)

Q6

Take 100 points on the line $x - y = 1$, compute the covariance matrix, eigen vectors and explain the eigen vector directions with respect to the slope line.
(your write up should not be more than one page as PNG.)

Q7

Given a dataset (of 100 points), run K means with K=2. We initialize with first 50 points as first cluster and the next 50 as second cluster. Compute the objective at

(A) initializtaion

(B) after iteration 1

(C) after iterations 2

(D) after iterations 3

(E) at convergence

Data can be downloaded from here:
`https://drive.google.com/file/d/1PT3Wcndp2JRojP2IqqerFI2Eui79vt_V/`
You need only to give the objective function values.

Q8

Given a data set D, if we increase K, objective of K means will monotonically decrease.

(A) Agree/Disagree with 2 sentences.

(B) How large K can go and what will be the objective then?

(C) Answer part A with respect to normalized objective ie objective divided by K.

Data can be downloaded from here:
`https://drive.google.com/file/d/1Thf1KRppeHRsCdzNjXM0za8o0Q-L4smV/`
(your write up should not be more than one page as PNG.)