Computational Data Analysis

CSE/ISYE 6740

Midterm Exam I– Feb. 14, 2018

Writing Time: 50 Minutes

Total Score: 100

If you think a question is unclear or multiple answers are reasonable, please write a brief explanation of your answer, to be safe. Also, show your work if you want wrong answers to have a chance at some credit: it lets us see how much you understood.

I have neither given nor received any unauthorized aid on this exam. I understand that this exam must be taken without the aid of notes, textbooks, the use of the internet, or any other aid. The work contained herein is wholly my own.

I understand that violation of these rules, including using an authorized aid or copying from another person, may result in my receiving a 0 on this exam .

Name:

GT ID:

GT Account:

Question 1	
Question 2	
Question 3	
Question 4	
Total	

1 Clustering [25 pts]

K-means

Given $m = 5$ data points configuration in Figure 1. Assume $K = 2$ and use Euclidean distance. Assuming the initialization of centroid as shown, after one iteration of k-means algorithm, answer the following questions.

(a) Show the cluster assignment;

Euclidean distance:

$$A = \{2, 5\} \quad B = \{1, 3, 4\}$$

Manhattan distance:

$$A = \{4, 5\} \quad B = \{1, 2, 3\}$$

(b) Show the location of the new center;

Euclidean distance:

$$\mu_A = (-1.5, -0.5) \quad \mu_B = \{\frac{5}{3}, \frac{2}{3}\}$$

Manhattan distance:

$$\mu_A = (-1, -1.5) \quad \mu_B = \{2, \frac{4}{3}\}$$

(c) Will it terminate in one step?

The new assignment after one step is:

Euclidean distance:

$$A = \{2, 4, 5\} \quad B = \{1, 3\}$$

Manhattan distance:

$$A = \{2, 4, 5\} \quad B = \{1, 3\}$$

So the assignment changed, thus it will not be terminated for both cases.

Spectral clustering

Consider the data point setting in Figure 2. We will use spectral clustering to divide these points into two clusters. Our version of spectral clustering uses a neighbourhood graph obtained by connecting each point to its two nearest neighbors (breaking ties randomly), and by weighting the resulting edges between points xi and x_j by $W_{ij} = \exp(-\|x_i - x_j\|)$.

(d) Indicate on Figure 2b the clusters that we will obtain from spectral clustering. Provide a brief justification.

The weight matrix W is:

$$W = \begin{bmatrix} 0 & exp(-1) & exp(-\sqrt{2}) & 0 & 0 & 0 & 0 & 0 \\ exp(-1) & 0 & exp(-1) & 0 & 0 & 0 & 0 & 0 \\ exp(-\sqrt{2}) & exp(-1) & 0 & exp(-1) & exp(-\sqrt{2}) & exp(-1) & 0 & 0 \\ 0 & 0 & exp(-1) & 0 & exp(-1) & 0 & 0 & 0 \\ 0 & 0 & exp(-\sqrt{2}) & exp(-1) & 0 & 0 & 0 & 0 \\ 0 & 0 & exp(-1) & 0 & 0 & 0 & exp(-1) & exp(-\sqrt{2}) \\ 0 & 0 & 0 & 0 & 0 & exp(-1) & 0 & exp(-1) \\ 0 & 0 & 0 & 0 & 0 & exp(-\sqrt{2}) & exp(-1) & 0 \end{bmatrix}$$

The eigenvectors of the 2 smallest eigenvalues are

$$\begin{bmatrix} 0.30092 & 0.35355 \\ 0.28391 & 0.35355 \\ 0.16792 & 0.35355 \\ 0.28391 & 0.35355 \\ 0.30092 & 0.35355 \\ -0.2983 & 0.35355 \\ -0.5044 & 0.35355 \\ -0.5347 & 0.35355 \end{bmatrix}$$

So we may have the following clusters: $cluster1 : \{12345\}$,$cluster2 : \{678\}$

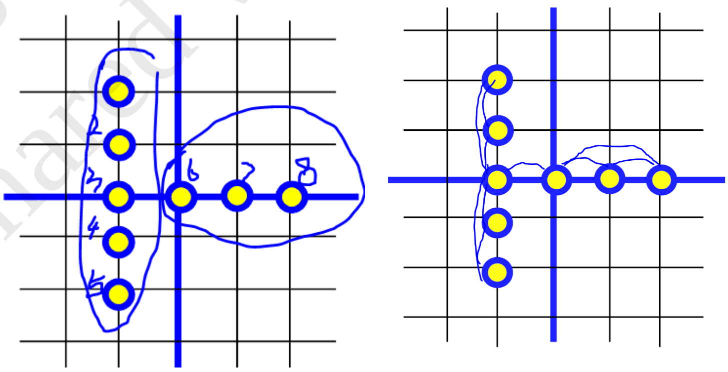


Figure 1: Question 2.

Intuitively, we can see it only have one edge between 3 and 6. Thus the two parts $\{12345\}$ and $\{678\}$ are connected respectively tightly.

Any reasonable argument are acceptable.

2 Principal Component Analysis [25 pts]

Suppose we have 4 points in 3-dimensional Euclidean space, namely $(4, -2, 4)$, $(5, -3, 5)$, $(2, 0, 2)$, and $(3, -1, 3)$.

(a) Find the first principal direction.

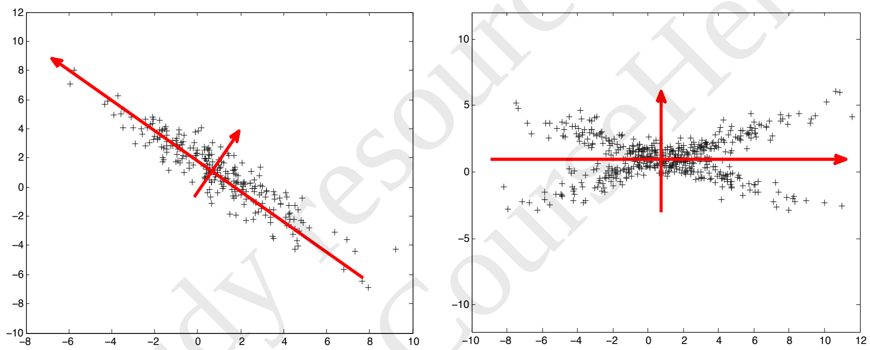
Answer:
$\frac{1}{\sqrt{3}}(1, -1, 1)$

Hints:
1. Visualize the data, and you will find all the data points lie in a straight line.
2. Check differences of any two points, and you will find the differences are proportional to $(1, -1, 1)$. For example:
$(5, -3, 5) - (4, -2, 4) = (1, -1, 1); (2, 0, 2) - (5, -3, 5) = -3(1, -1, 1); ...$

(b) When we reduce the dimensionality from 3 to 1 based on the principal direction you found in (a), what is the reconstruction error in terms of variance?

Answer:
0. Because the rank of the centered data matrix is 1.

(c) You are given the following 2-D datasets, approximately draw the first and second principal directional on each plot.



(a) (b)

3 Density Estimation [25 pts]

(a) We have a random variable X drawn from a Poisson distribution. The Poisson distribution is a discrete distribution and X can be any non-negative integer. The probability of X at a point x is $p(x) = \frac{\lambda^x e^{-\lambda}}{x!}$. Given points x_1, …, x_n, write down the maximum likelihood estimate (MLE) of λ. [15 pts]

Answer: The joint likelihood given points $x_1, \ldots, x_n$ is:

$$\mathcal{L} = \prod_{i=1}^{n} \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}.$$

The log-likelihood is:

$$\ell = \sum_{i=1}^{n} \left[x_i \log(\lambda) - \lambda - \log(x_i!) \right].$$

Take derivative with respect to λ and set the derivative 0:

$$\begin{aligned} \frac{d\ell}{d\lambda} &= \sum_{i=1}^{n} \left[\frac{x_i}{\lambda} - 1 \right] = 0 \\ &\Rightarrow \frac{\sum_{i=1}^{n} x_i}{\lambda} - n = 0. \end{aligned}$$

Therefore, we obtain the MLE of λ:

$$\widehat{\lambda}_{\text{MLE}} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}.$$

Remark: More rigorously, you also need to check the second derivative of ℓ with respect to λ and show

$$\frac{d^2\ell}{d\lambda^2} = \sum_{i=1}^{n} \left[-\frac{x_i}{\lambda^2} \right] < 0.$$

(b) Non-parametric models do not have parameters. [2 pts]

- Yes / No

(c) In kernel density estimation, a large kernel bandwidth will results in low bias. [2 pts]

- Yes / No

(d) Non-parametric models are usually more efficient than parametric models in terms of model storage. [2 pts]

- Yes / No

(e) Suppose K_1 and K_2 are valid kernels for KDE. Is $K = \alpha K_1 + \beta K_2$, $\alpha, \beta \in \mathbb{R}$ a valid kernel? [2 pts]

- Yes / No

Hint: Find a counterexample: suppose $\alpha < 0$ and $\beta < 0$.

4 Probability and Bayes' Rule [25 pts]

(a) A probability density function (pdf) is defined by

$$f(x,y) = \begin{cases} C(x+2y) & \text{if } 0 < y < 1 \text{ and } 0 < x < 2 \\ 0 & \text{otherwise} \end{cases}$$

(i) Find the value of C.

$$C = \frac{1}{4}$$

(ii) Find the marginal distribution of X.

$$f(x) = (x+1)/4$$

(iii) Find the joint cumulative density function (cdf) of X and Y.

$$F(x,y) = P(X \le x, Y \le y) = (0.5x^2y + xy^2)/4$$