

Main Examination period 2022 – May/June – Semester B

## MTH793P: Advanced Machine Learning

You should attempt ALL questions. Marks available are shown next to the questions.

In completing this assessment:

- You may use books and notes.
- You may use calculators and computers, but you must show your working for any calculations you do.
- You may use the Internet as a resource, but not to ask for the solution to an exam question or to copy any solution you find.
- You must not seek or obtain help from anyone else.

All work should be **handwritten** and should **include your student number**.

The exam is available for a period of **24 hours**. Upon accessing the exam, you will have **4 hours** in which to complete and submit this assessment.

When you have finished:

- scan your work, convert it to a **single PDF file**, and submit this file using the tool below the link to the exam;
- e-mail a copy to **maths@qmul.ac.uk** with your student number and the module code in the subject line;
- with your e-mail, include a photograph of the first page of your work together with either yourself or your student ID card.

Please try to upload your work well before the end of the submission window, in case you experience computer problems. **Only one attempt is allowed – once you have submitted your work, it is final.**

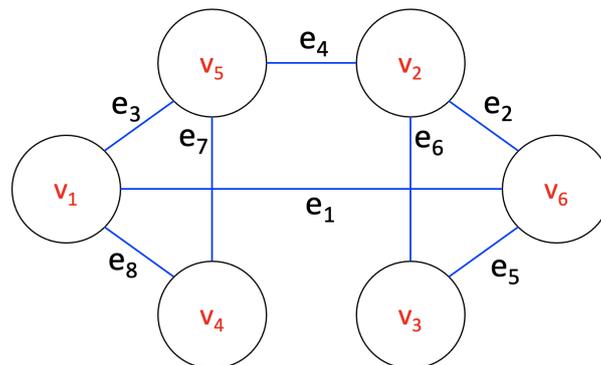
Examiners: O. Bobrowski, P. Skraba

(1) For involved mathematical computations (e.g., inverting a matrix, computing eigenvectors, etc.) you are encouraged to use a calculator or a computer, unless stated otherwise in the problem. You should make it clear where you used a computer and how.

(2) When asked to perform a certain machine-learning task (e.g., K-means, PCA), you should present all the steps of execution for the algorithm, and **not** run the algorithm on a computer.

### Question 1 [35 marks]. Graph clustering

Let  $G$  be the following graph:



We assign weights to the edges in the following way:  $w(e_i) = i$ .

- (a) Write down the graph Laplacian  $L$  for this weighted graph. Make sure the rows of the Laplacian have the same ordering as the vertices. [5]
- (b) Suppose that we are given that  $v_1$  is labelled as ‘0’, and  $v_6$  as ‘1’. Use the semi-supervised graph labelling method discussed in class to label all the other vertices. You may use a computer to solve linear systems, but should explain all the steps of how you arrive at the system, and how to interpret the output. [10]
- (c) Suppose we modify the weights of  $e_2, e_4$  to be  $w(e_2) = 20$  and  $w(e_4) = 40$ . Repeat the labelling procedure above with the new weights. What is the new graph labelling that you get? Provide an intuitive explanation of your results. [5]
- (d) In this part we assume **no label information** is given, with the original weights on the edges ( $w(e_i) = i$ ). Compute the eigenvector of  $L$  corresponding to the second smallest eigenvalue, and use it to divide the vertices into  $k = 2$  clusters. Do your result agree with part (b)? [10]
- (e) Suppose that the true clusters for this graph are given by:

$$C_1^* = \{v_1, v_4, v_5\}, \quad C_2^* = \{v_2, v_3\}, \quad C_3^* = \{v_6\}.$$

Compute the Rand Index, to rank the quality of the clustering result you received in (d). Show all the steps in the calculation. [5]

**Question 2 [25 marks]. K-means clustering**

In this problem, you are asked to run the k-means algorithm **by hand**, showing all the steps and calculations. You may use a calculator if you find it helpful. For your convenience, you can use the tables on the last page of this exam paper, where you can document the results of each step in the iteration of the algorithm. If you do, make sure to attach them to your submitted pdf. Note that the tables may be longer than the number of steps required.

- (a) Suppose that we are given the following dataset in  $\mathbb{R}^2$ :

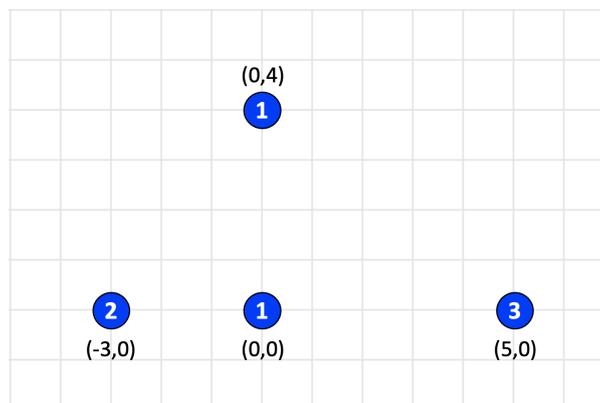
$$x_1 = (0, 0), \quad x_2 = (0, 4), \quad x_3 = (-3, 0), \quad x_4 = (5, 0).$$

Run the k-means algorithm (until it converges) with  $k = 2$  and with initial centroids  $\mu_1^{(0)} = (0, 0)$ ,  $\mu_2^{(0)} = (-3, 0)$ . Explain the steps in each iteration. What are the resulting clusters and centroids? [5]

- (b) Next, we assume we have data points at the same locations as in (a), but with some points appearing multiple times. Our new dataset is:

$$x_1 = (0, 0), \quad x_2 = (0, 4), \quad x_3 = x_4 = (-3, 0), \quad x_5 = x_6 = x_7 = (5, 0).$$

This is summarised in the following figure (the numbers in the circles are the number of points placed at each location).



Repeat part (a) for the new dataset, with the same initial centroids as in (a). [10]

- (c) Repeat part (b) but for initial centroids  $\mu_1^{(0)} = (0, 0)$  and  $\mu_2^{(0)} = (0, 4)$ . [5]

- (d) Are your clustering results from parts (b) and (c) the same?

**If yes** – is it true that running k-means always yields the same results? (regardless of initial centroids). **If no** – explain how is that possible. [5]

**Question 3 [25 marks]. SVD and PCA**

In this problem you must show all your calculations.

(a) Consider the matrix:

$$M = \begin{pmatrix} 1 & 2 & -1 & 0 & -2 \\ 2 & 1 & 0 & -2 & -1 \end{pmatrix}$$

Find the matrices  $U$  and  $\Sigma$  of the SVD:  $M = U\Sigma V^T$  (no need to find  $V$ ).

**HINT:** Use the matrix  $M \cdot M^T$ .

[10]

(b) Find the solution to the following problem

$$\tilde{M} = \arg \min_{M' \in \mathbb{R}^{2 \times 5}} \|M - M'\|_F, \quad \text{subject to } \text{rank}(M') = 1.$$

where  $\|\cdot\|_F$  stands for the Frobenius norm.

[5]

(c) You are given a set of points  $\{x_1, x_2, x_3, x_4, x_5\} \subset \mathbb{R}^2$ , where

$$x_1 = (1, 2), x_2 = (2, 1), x_3 = (-1, 0), x_4 = (0, -2), x_5 = (-2, -1).$$

Denote by  $\hat{x}_1, \dots, \hat{x}_5$  the lower dimensional approximation of the dataset using the first principal component. Find the coordinates of the points  $\hat{x}_1, \dots, \hat{x}_5$ , and in a single axis system provide a sketch for the following:

- The original points  $x_1, x_2, x_3, x_4, x_5$ .
- Their lower dimensional approximation  $\hat{x}_1, \hat{x}_2, \hat{x}_3, \hat{x}_4, \hat{x}_5$ .
- The line that represents the direction of the principal component.

[10]

**Question 4 [15 marks]. Matrix completion**

(a) Suppose you are given the following matrix with missing entries:

$$M = \begin{pmatrix} 1 & ? & ? & 2 & 1 \\ 0 & 1 & 2 & -1 & 0 \\ ? & 3 & ? & -3 & 0 \\ 1 & 4 & 2 & 1 & ? \end{pmatrix}.$$

Can the above matrix be completed to be rank 1? Explain your answer.

[5]

(b) Find the missing values in order for the resulting matrix to have minimum rank.

[10]

---

End of Paper.

Table for Question 2(a):

	assignments				centroids	
	$x_1$	$x_2$	$x_3$	$x_4$	$\mu_1$	$\mu_2$
<b>Step 1:</b>						
<b>Step 2:</b>						
<b>Step 3:</b>						
<b>Step 4:</b>						
<b>Step 5:</b>						

Table for Question 2(b):

	assignments							centroids	
	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$\mu_1$	$\mu_2$
<b>Step 1:</b>									
<b>Step 2:</b>									
<b>Step 3:</b>									
<b>Step 4:</b>									
<b>Step 5:</b>									

Table for Question 2(c):

	assignments							centroids	
	$x_1$	$x_2$	$x_3$	$x_4$	$x_5$	$x_6$	$x_7$	$\mu_1$	$\mu_2$
<b>Step 1:</b>									
<b>Step 2:</b>									
<b>Step 3:</b>									
<b>Step 4:</b>									
<b>Step 5:</b>									