Queen Mary
University of London

Late-Summer Examination period 2022

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.

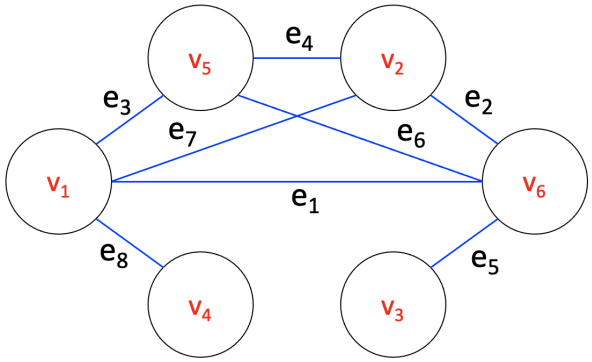
Please try to upload your work well before the end of the submission window, in case you experience computer problems. **Only one attempt is allowed – once you have submitted your work, it is final**.

Examiners: O. Bobrowski, P. Skraba

 Continue to next page

(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 [25 marks]. Graph clustering
Let G be the following graph:



We assign weights to the edges in the following way: $w(e_i) = 10 - i$.

(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 both v_2 and v_6 as '1'. Use the semi-supervised graph labelling method discussed in class to label all the other vertices. You are allowed to use a computer to solve a linear system, but should explain all the steps leading to this system, and how to interpret the output. **[10]**

(c) Suppose we modify the weights of e_5, e_8 to be $w(e_5) = 0.000005$ and $w(e_8) = 4,000,000$. Without running the algorithm again, explain what will be the effect of this change on the results. **[5]**

(d) Suppose that we repeat part (b), but now with v_1, v_2, v_6 all labelled as '1'. Without running the algorithm again, explain what will be the effect on the results. **[5]**

 Continue to next page

Question 2 [25 marks]. K-means clustering
Consider the set of data points:

$$x_1 = (1,1)^T, \quad x_2 = (2,3)^T, \quad x_3 = (-1,-2)^T, \quad x_4 = (4,4)^T, \quad x_5 = (-3,-3)^T.$$

(a) Perform the first two steps of the K-means algorithm for these points, with $k = 2$ and the initial centroids given by:

$$\mu_1 = (0,3)^T, \quad \mu_2 = (-2,-2)^T.$$

You should run the k-means algorithm **by hand**, and not use any software. You may use a calculator if you find it helpful. [**15**]

(b) Let C_1, C_2 be the two clusters produced by the K-means algorithm in part (a). Compute the Dunn-Index (DI) for these clusters, using the single-linkage inter-cluster distance, and the diameter intra-cluster distance. [**10**]

 Continue to next page

Question 3 [35 marks]. SVD and PCA

In this problem you are **not** allowed to use a computer, except for part (c).

(a) Consider the matrix:

$$M = \begin{pmatrix} 3 & 4 & 0 & 0 \\ 5 & 0 & 0 & 0 \end{pmatrix}$$

Find the matrices U,Σ, and V of the SVD: $M = U\Sigma V^T$.
HINT: Use the matrices MM^T, M^TM. [**10**]

(b) You are given a set of points $\{x_1, x_2, x_3, x_4, x_5\} \subset \mathbb{R}^3$, where

$$x_1 = (3, 2, 3), x_2 = (2, 3, -2), x_3 = (-1, 3, 2), x_4 = (1, -2, 0), x_5 = (0, -1, 2).$$

We want to find the best fit for a line approximating these points using PCA. We will break it into a few steps:

(i) Find the empirical mean of the dataset, denoted $\bar{x}$.

(ii) Centre the data points $x_1, \ldots, x_5$ using $\bar{x}$, and stack the resulting vectors as columns in a matrix called X.

(iii) Using the SVD decomposition of X, find the principal components (directions) of X.

(iv) Find the projection of X onto its first principal component.

(v) Write down the resulting approximation, denoted $\hat{x}_1, \ldots, \hat{x}_5$ (don't forget to fix the mean).

Note: In this part you are allowed to use a computer to compute eigenvectors/singular vectors, but do **not** use any implemented PCA routine. [**15**]

(c) Recall the definition of the singular value thresholding operator:

$$X = U\Sigma V^T \longrightarrow D_\tau(X) = US_\tau(\Sigma)V^T,$$

where S_τ is the soft-thresholding function.

Considering the data matrix X from part (c) – what is rank(X)? What is rank$(D_4(X))$ [**10**]

 Continue to next page

Question 4 [15 marks]. Robust PCA & Matrix completion

(a) Given a matrix

$$M = \begin{pmatrix} 2 & 1 & 4 & 3 \\ 4 & 21 & 8 & 6 \\ -2 & -1 & -4 & 13 \end{pmatrix}$$

find the decomposition $M = L + E$ where E is a sparse matrix (with at most 3 nonzero entries), and L is a low-rank matrix (lowest rank possible). [**10**]

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

$$M = \begin{pmatrix} 1 & ? & ? & ? \\ 0 & ? & 4 & 0 \\ 0 & ? & ? & 1 \end{pmatrix}$$

Can the above matrix be completed to be rank 2? Explain your answer. [**5**]

 Continue to next page

End of Paper.