Queen Mary
University of London

Late-Summer Examination period 2023

MTH793P: Advanced machine learning

Duration: 4 hours

The exam is available for a period of **4 hours**, within which you must complete the assessment and submit your work. **Only one attempt is allowed – once you have submitted your work, it is final**.

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

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.**

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.

Examiners: 1st **Dr. N. Perra,** 2nd **Dr. N. Otter**

 Continue to next page

Question 1 [25 marks].

Consider a graph $\mathbf{G}(V, E)$ defined by the following list of undirected edges $e_1 = (1,2)$, $e_2 = (2,3)$, $e_3 = (1,3)$, $e_4 = (3,4)$, $e_5 = (4,5)$, $e_6 = (5,6)$, $e_7 = (4,6)$. The edges are weighted as follows $w_{1,2} = 10$, $w_{2,3} = 10$, $w_{1,3} = 5$, $w_{3,4} = 2$, $w_{4,5} = 20$, $w_{5,6} = 15$, $w_{4,6} = 10$

(a) Draw the graph and write down the incidence matrix $\mathbf{M}$. **[5]**

(b) Using the incidence matrix write down the Laplacian matrix $\mathbf{L}$. **[5]**

(c) Write down the adjacency matrix $\mathbf{A}$ and the matrix $\mathbf{D}$ whose diagonal elements are the weighted degree (i.e., strength) of each node. Verify that $\mathbf{L} = \mathbf{D} - \mathbf{A}$. **[5]**

(d) Assume that we gather partial information about some category c of nodes 1 and 4, namely $c_1 = 1$ and $c_4 = 0$. Using the network information in a semi-supervised setting, you are tasked to predict the category of the other four nodes. We know that this problem leads to the following equation:

$$P_{I_1/I_2}^{\top} \mathbf{L} P_{I_1/I_2} \hat{\mathbf{w}} = -P_{I_1/I_2}^{\top} \mathbf{L} P_{I_2} \mathbf{v} \tag{1}$$

where $\mathbf{v}$ is the vector of known categories.

- Write down the expressions for P_{I_2} and P_{I_1/I_2}, **[5]**
- Find $\hat{\mathbf{w}}$ by solving the normal equation. **[5]**

 Continue to next page

Solution:

(a) *Variation of a problem discussed in the coursework.* The graph is drawn in Figure 1. The incidence matrix $\mathbf{M}$ reads

$$\mathbf{M} = \begin{pmatrix} -\sqrt{10} & \sqrt{10} & 0 & 0 & 0 & 0 \\ 0 & -\sqrt{10} & \sqrt{10} & 0 & 0 & 0 \\ -\sqrt{5} & 0 & \sqrt{5} & 0 & 0 & 0 \\ 0 & 0 & -\sqrt{2} & \sqrt{2} & 0 & 0 \\ 0 & 0 & 0 & -\sqrt{20} & \sqrt{20} & 0 \\ 0 & 0 & 0 & 0 & -\sqrt{15} & \sqrt{15} \\ 0 & 0 & 0 & -\sqrt{10} & 0 & \sqrt{10} \end{pmatrix} \tag{2}$$

(b) *Variation of a problem discussed in the coursework.* The Laplacian matrix is $\mathbf{L} = \mathbf{M}^\top \mathbf{M}$. Hence we have:

$$\mathbf{L} = \begin{pmatrix} 15 & -10 & -5 & 0 & 0 & 0 \\ -10 & 20 & -10 & 0 & 0 & 0 \\ -5 & -10 & 17 & -2 & 0 & 0 \\ 0 & 0 & -2 & 32 & -20 & -10 \\ 0 & 0 & 0 & -20 & 35 & -15 \\ 0 & 0 & 0 & -10 & -15 & 25 \end{pmatrix} \tag{3}$$

(c) *Variation of a problem discussed in the coursework.* The adjacency matrix is:

$$\mathbf{A} = \begin{pmatrix} 0 & 10 & 5 & 0 & 0 & 0 \\ 10 & 0 & 10 & 0 & 0 & 0 \\ 5 & 10 & 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 & 20 & 10 \\ 0 & 0 & 0 & 20 & 0 & 15 \\ 0 & 0 & 0 & 10 & 15 & 0 \end{pmatrix} \tag{4}$$

the weighted degree (i.e., strength) is

$$\mathbf{L} = \begin{pmatrix} 15 & 0 & 0 & 0 & 0 & 0 \\ 0 & 20 & 0 & 0 & 0 & 0 \\ 0 & 0 & 17 & 0 & 0 & 0 \\ 0 & 0 & 0 & 32 & 0 & 0 \\ 0 & 0 & 0 & 0 & 35 & 0 \\ 0 & 0 & 0 & 0 & 0 & 25 \end{pmatrix} \tag{5}$$

Clearly $\mathbf{D} - \mathbf{A} = \mathbf{L}$

(d) *Variation of a problem discussed in the coursework.* The operator P_{I_2} projects $\mathbf{v} \in \mathbb{R}^{2\times 1}$ into a $\mathbf{y}_v \in \mathbb{R}^{6\times 1}$ selecting the known data, hence we have

$$P_{I_2} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \tag{6}$$

 Continue to next page

The operator P_{I_1/I_2} selects instead the unknown data, hence we have

$$P_{I_1/I_2} = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{7}$$

(e) *Variation of a problem discussed in the coursework.* The normal equation reads:

$$\begin{pmatrix} 20 & -10 & 0 & 0 \\ -10 & 17 & 0 & 0 \\ 0 & 0 & 35 & -15 \\ 0 & 0 & -15 & 25 \end{pmatrix} \hat{\mathbf{w}} = - \begin{pmatrix} -10 \\ -5 \\ 0 \\ 0 \end{pmatrix} \tag{8}$$

which implies $\hat{\mathbf{w}} = \left(\frac{11}{12}, \frac{5}{6}, 0, 0\right)^\top$. Hence, we can infer that $c_2 = c_3 = c_1 = 1$ and $c_5 = c_6 = c_4 = 0$

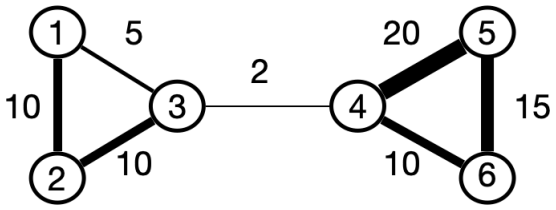

Figure 1: Representation of graph **G**

 Continue to next page

Question 2 [40 marks].

(a) Considering the following data points $p_1 = -3$, $p_2 = -1$, $p_3 = 1$, $p_4 = 2$, $p_5 = 3$, $p_6 = 5$, $p_7 = 8$, cluster them applying k-means starting with centroids $\mu_1^{(0)} = 0$ and $\mu_2^{(0)} = 9$. Write out all steps of the algorithm by hand. **[10]**

(b) Compute the Rand Index considering as $\mathcal{P}_1$ the partition outcome of the clustering in the previous point and $\mathcal{P}_2 = (C_1', C_2', C_3')$ where $C_1' = (p_1, p_2)$, $C_2' = (p_3, p_4)$ and $C_3' = (p_5, p_6, p_7)$. Write out all steps by hand. **[15]**

(c) We are now given a representation of the points in a two dimensional space $\mathbf{p}_1 = (-1, -1)^\top$, $\mathbf{p}_2 = (-3, -2)^\top$, $\mathbf{p}_3 = (1, 0)^\top$, $\mathbf{p}_4 = (2, 0)^\top$, $\mathbf{p}_5 = (3, 4)^\top$, $\mathbf{p}_6 = (5, 2)^\top$, $\mathbf{p}_7 = (8, 2)^\top$. Cluster the data points applying k-means starting with centroids $\mu_1^{(0)} = (0, -1)^\top$ and $\mu_2^{(0)} = (1, 1)^\top$. Write out all steps of the algorithm by hand. **[15]**

 Continue to next page

Solution:

(a) *Variation of a problem discussed in the coursework.* As first step we need to compute the distance between all points and the two centroids $d_{ik}^{(1)}$ where $i \in \{1,2,3,4,5,6,7\}$ and $k \in \{1,2\}$

$$d_{ik}^{(1)} = \begin{pmatrix} 3 & 12 \\ 1 & 10 \\ 1 & 8 \\ 2 & 7 \\ 3 & 6 \\ 5 & 4 \\ 8 & 1 \end{pmatrix} \tag{9}$$

which implies

$$z_{ik}^{(1)} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \tag{10}$$

in words, the first fives data points at the first iteration of the algorithm are closer to the first centroid, the last two data points instead are closer to the second one. Given these new assignment, we can update the coordinates of the two centroids obtaining

$$\begin{aligned} \mu_1^{(1)} &= \frac{\sum_j z_{j1}^{(1)} p_j}{\sum_j z_{j1}^{(1)}} = \frac{-3-1+1+2+3}{5} = \frac{2}{5} \\ \mu_2^{(1)} &= \frac{\sum_j z_{j2}^{(1)} p_j}{\sum_j z_{j2}^{(1)}} = \frac{5+8}{2} = \frac{13}{2} \end{aligned} \tag{11}$$

we can now compute the distance between the points and the updated centroids

$$d_{ik}^{(2)} = \begin{pmatrix} 17/5 & 19/2 \\ 7/5 & 15/2 \\ 3/5 & 11/2 \\ 8/5 & 9/2 \\ 13/5 & 7/2 \\ 23/5 & 3/2 \\ 38/5 & 3/2 \end{pmatrix} \tag{12}$$

 Continue to next page

which implies

$$z_{ik}^{(2)} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \tag{13}$$

as $z_{ik}^{(2)} = z_{ik}^{(1)}$ we are at convergence.

(b) *Variation of a problem discussed in the coursework.* To compute the Rand Index we need to compare $\mathcal{P}_1 = \{C_1, C_2\}$ with $C_1 = \{p_1, p_2, p_3, p_4, p_5\}$ and $C_2 = \{p_6, p_7\}$ with $\mathcal{P}_2 = \{C_1', C_2', C_3'\}$ with $C_1' = \{p_1, p_2\}$ $C_2' = \{p_3, p_4\}$ and $C_3' = \{p_5, p_6, p_7\}$. The true positives (pairs of points that are in the same clusters in both partitions) are $TP = 3$ which are (p_1, p_2), (p_3, p_4) (p_6, p_7). The true negative (pairs of points that are in different clusters in both) are $TN = 8$ which are (p_1, p_6), (p_1, p_7), (p_2, p_6), (p_2, p_7), (p_3, p_6), (p_3, p_7), (p_4, p_6), (p_4, p_7). Hence,

$$RI = \frac{TP + TN}{N(N-1)/2} = \frac{11}{21} \approx 0.524 \tag{14}$$

(c) *Variation of a problem discussed in the coursework.* As first step we need to compute the distance between all points and the two centroids $d_{ik}^{(1)}$ where $i \in \{1, 2, 3, 4, 5, 6, 7\}$ and $k \in \{1, 2\}$

$$d_{ik}^{(1)} = \begin{pmatrix} 1 & 2\sqrt{2} \\ \sqrt{10} & 5 \\ \sqrt{2} & 1 \\ \sqrt{5} & \sqrt{2} \\ \sqrt{34} & \sqrt{13} \\ \sqrt{34} & \sqrt{17} \\ \sqrt{73} & 5\sqrt{2} \end{pmatrix} \tag{15}$$

which implies

$$z_{ik}^{(1)} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \tag{16}$$

Given these new assignment, we can update the coordinates of the two

centroids obtaining

$$\begin{aligned} \mu_1^{(1)} &= \left(\frac{-1-3}{2}, \frac{-1-2}{2}\right)^\top = \left(-2, -\frac{3}{2}\right)^\top \\ \mu_2^{(1)} &= \left(\frac{1+2+3+5+8}{5}, \frac{3+2+2}{5}\right)^\top = \left(\frac{19}{5}, \frac{7}{5}\right)^\top \end{aligned} \tag{17}$$

Considering the geometry of the problem the new assignment will not change with the new position of centroids, hence the k-means converges after one iteration.

 Continue to next page

Question 3 [20 marks]. Consider the following data points $\mathbf{p}_1 = (1,2)^\top$, $\mathbf{p}_2 = (2,3)^\top$, $\mathbf{p}_3 = (4,1)^\top$.

(a) Write down the correspondent $\mathbf{X} \in \mathbb{R}^{2\times 3}$ data matrix, compute its singular values, and the left singular vectors $\mathbf{X}$. **[10]**

(b) Compute the matrix $\text{soft}_\tau(\mathbf{\Sigma})$ obtained by applying the soft thresholding operator to each element of the matrix of singular values $\mathbf{\Sigma}$. Set τ equal to the last digit of your student ID. Compute the nuclear norm of $D_\tau(\mathbf{X}) = \mathbf{U}\text{soft}_\tau(\mathbf{\Sigma})\mathbf{V}^\top$ and compare it with the nuclear norm of the original matrix ($\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors respectively). **[5]**

(c) Compute a lower rank approximation $\hat{\mathbf{L}} \in \mathbb{R}^{2\times 3}$ of the matrix $\mathbf{X}$ by hand, considering $rank(\hat{\mathbf{L}}) = 1$ and such that $\|\hat{\mathbf{L}} - \mathbf{X}\| \leqslant \|\mathbf{L} - \mathbf{X}\|$ for all $\mathbf{L} \in \mathbb{R}^{2\times 3}$ and $rank(\mathbf{L}) = 1$. **[5]**

 Continue to next page

Solution:

(a) *Variation of a problem discussed in the coursework.* The data matrix reads:

$$\mathbf{X} = \begin{pmatrix} 1 & 2 & 4 \\ 2 & 3 & 1 \end{pmatrix} \tag{18}$$

(b) *Variation of a problem discussed in the coursework.* The singular values are conveniently computed considering the eigenvalues of $\mathbf{X}\mathbf{X}^\top$. This squared matrix is

$$\mathbf{X}\mathbf{X}^\top = \begin{pmatrix} 21 & 12 \\ 12 & 14 \end{pmatrix} \tag{19}$$

The eigenvalues are the square of the singular values and are equal to $\sigma_1^2 = 30$, $\sigma_2^2 = 5$. The left singular vectors are the correspondent eigenvalues which, after normalisation, read $\mathbf{u}_1 = \left(\frac{4}{5}, \frac{3}{5}\right)^\top$ and $\mathbf{u}_2 = \left(-\frac{3}{5}, \frac{4}{5}\right)^\top$.

(c) *We have discussed the operator at length in the lectures and a related problem was discussed in the coursework.* The matrix $D_\tau(\mathbf{X})$ is obtained applying the singular thresholding operator to the matrix of singular values in the SVD. This implies soft thresholding each singular value. In practice

$$D_\tau(\mathbf{X}) = \mathbf{U}\text{soft}_\tau(\mathbf{\Sigma})\mathbf{V}^\top \tag{20}$$

where soft_τ is the soft thresholding operator, defined as:

$$\text{soft}_\tau(\sigma_i) = \begin{cases} \sigma_i - \tau, & \text{if } \sigma_i > \tau \\ 0, & \text{if } -\tau \leqslant \sigma_i \leqslant \tau \\ \sigma_i + \tau, & \text{if } \sigma_i < -\tau \end{cases} \tag{21}$$

The matrix $\mathbf{\Sigma}$ reads:

$$\mathbf{\Sigma} = \begin{pmatrix} 30 & 0 & 0 \\ 0 & 5 & 0 \end{pmatrix} \tag{22}$$

Hence, we have to cases. In case the last digit of the ID is $\tau \leqslant 5$ we have

$$\text{soft}_\tau(\mathbf{\Sigma}) = \begin{pmatrix} 30-\tau & 0 & 0 \\ 0 & 5-\tau & 0 \end{pmatrix} \tag{23}$$

In case the last digit of the ID is $\tau > 5$ instead we have we have

$$\text{soft}_\tau(\mathbf{\Sigma}) = \begin{pmatrix} 30-\tau & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{24}$$

If $\tau \leqslant 5$ the nuclear norm of $D_\tau(\mathbf{X})$ is $\|D_\tau(\mathbf{X})\|_* = \sqrt{30-\tau} + \sqrt{5-\tau}$. Instead if $\tau > 5$ is $\|D_\tau(\mathbf{X})\|_* = \sqrt{30-\tau}$. Hence, the nuclear norm is always smaller than the nuclear norm of the orginal matrix ($\|\mathbf{X}\|_* = \sqrt{30} + \sqrt{5}$), unless $\tau = 0$.

 Continue to next page

(d) *Variation of a problem discussed in the coursework.* The best lower rank approximation for rank equal to one is

$$\hat{\mathbf{L}} = \mathbf{u}_1 \mathbf{u}_1^\top \mathbf{X} \tag{25}$$

hence we can write

$$\hat{\mathbf{L}} = \begin{pmatrix} \frac{4}{5} \\ \frac{3}{5} \end{pmatrix} \begin{pmatrix} \frac{4}{5} & \frac{3}{5} \end{pmatrix} \begin{pmatrix} 1 & 2 & 4 \\ 2 & 3 & 1 \end{pmatrix} = \frac{1}{25} \begin{pmatrix} 40 & 68 & 76 \\ 30 & 51 & 57 \end{pmatrix} \tag{26}$$

 Continue to next page

Question 4 [15 marks].

(a) Consider the matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & 5 & 0 \\ 0 & 1 & 1 \end{pmatrix} \tag{27}$$

diagonalise the matrix $\mathbf{X}\mathbf{X}^\top$ and discuss its connection with the SVD of $\mathbf{X}$. **[10]**

(b) Consider a matrix $\mathbf{M} \in \mathbb{R}^{m \times n}$. Show that $Trace(\mathbf{M}\mathbf{M}^\top + \mathbf{M}^\top\mathbf{M}) = 2\sum_{i=1}^{r} \sigma_i^2$ where σ_i are its singular values. **[5]**

 Continue to next page

Solution:

(a) *We have discussed about this problem in the lectures.* The matrix $\mathbf{X}\mathbf{X}^\top$ reads:

$$\mathbf{X}\mathbf{X}^\top = \begin{pmatrix} 1 & 5 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 5 & 1 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 26 & 5 \\ 5 & 2 \end{pmatrix} \tag{28}$$

To diagonalise the matrix we need to find two other matrices such that $\mathbf{M}\mathbf{M}^\top = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$. Since the matrix is symmetric we have $\mathbf{M}\mathbf{M}^\top = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^\top$ which implies $\mathbf{P}^\top\mathbf{P} = \mathbb{1}$. The matrix $\mathbf{P}$ is the matrix formed by the eigenvectors of the matrix we are asked to diagonalise while the matrix $\mathbf{\Lambda}$ is a diagonal matrix whose elements are the eigenvalues. The eigenvalues of the matrix $\mathbf{M}\mathbf{M}^\top$ are easily obtained as $\lambda_1 = 27$ and $\lambda_2 = 1$. The correspondent eigenvectors are instead are $\mathbf{P}_1 = \frac{1}{\sqrt{26}}(5,1)^\top$ and $\mathbf{P}_2 = \frac{1}{\sqrt{26}}(1,-5)^\top$. Hence, we have

$$\mathbf{X}\mathbf{X}^\top = \frac{1}{26}\begin{pmatrix} 5 & 1 \\ 1 & -5 \end{pmatrix}\begin{pmatrix} 27 & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 5 & 1 \\ 1 & -5 \end{pmatrix} \tag{29}$$

The connections with the SVD of the matrix $\mathbf{X}$ is clear. Indeed if we write $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ we can write $\mathbf{X}\mathbf{X}^\top = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^\top\mathbf{U}^\top$ where indeed $\mathbf{\Sigma}\mathbf{\Sigma}^\top$ is the diagonal matrix with the eigenvalues of $\mathbf{X}\mathbf{X}^\top$ and the left singular vectors by the definition are the correspondent eigenvectors.

(b) *We have discussed similar problems in the lectures*. The first observation is that $Trace(\mathbf{A}+\mathbf{B}) = Trace(\mathbf{A}) + Trace(\mathbf{B})$. Furthermore, $Trace(\mathbf{A}\mathbf{B}) = Trace(\mathbf{B}\mathbf{A})$. Hence, we have

$$Trace(\mathbf{M}\mathbf{M}^\top + \mathbf{M}^\top\mathbf{M}) = 2Trace(\mathbf{M}\mathbf{M}^\top) \tag{30}$$

Now, we can use the SVD of the matrix $\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$:

$$Trace(\mathbf{M}\mathbf{M}^\top) = Trace(\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top\mathbf{V}\mathbf{\Sigma}^\top\mathbf{U}^\top) = Trace(\mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^\top\mathbf{U}^\top) \tag{31}$$

using the properties of traces we can write

$$Trace(\mathbf{M}\mathbf{M}^\top) = Trace(\mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^\top\mathbf{U}^\top) = Trace(\mathbf{\Sigma}\mathbf{\Sigma}^\top\mathbf{U}^\top\mathbf{U}) = Trace(\mathbf{\Sigma}\mathbf{\Sigma}^\top) \tag{32}$$

The matrix $\mathbf{\Sigma}\mathbf{\Sigma}^\top$ is a diagonal matrix whose elements are the squared of the singular values. Hence,

$$Trace(\mathbf{M}\mathbf{M}^\top) = Trace(\mathbf{\Sigma}\mathbf{\Sigma}^\top) = \sum_{i=1}^{r}\sigma_i^2 \tag{33}$$

which implies

$$Trace(\mathbf{M}\mathbf{M}^\top + \mathbf{M}^\top\mathbf{M}) = 2\sum_{i=1}^{r}\sigma_i^2 \tag{34}$$

End of Paper.