Two hours

EXAM PAPER MUST NOT BE REMOVED FROM
THE EXAM ROOM AND MUST BE RETURNED

**UNIVERSITY OF MANCHESTER**
**SCHOOL OF COMPUTER SCIENCE**

Modelling and visualisation of high-dimensional data

|  |  |
|---|---|
| Date: | Friday 27th January 2017 |
| Time: | 14:00 - 16:00 |

**Answer ALL Questions in Section A**
**Write your answers directly on the exam paper. Only answers written in the boxes on the exam paper will be marked.**

**Answer ALL Questions in Section B, use a separate answerbook for this Section**

This is a CLOSED book examination

The use of electronic calculators is permitted provided they are not programmable and do not store text

# *Section A*

*This Section contains Multiple Choice Questions and is therefore restricted*

## Section B

Answer *all* questions from this section.

1. *Principal Component Analysis* (PCA) can be formulated from different perspectives. One of them is minimising the error of reconstructing the data from their low dimensional representations. Describe the cost function in this formulation and state what parameters are learnt with a given training data set of $N$ points, $X = \{\mathbf{x}_1, \cdots, \mathbf{x}_N\}$ . (4 marks)

2. *Multi-dimensional scaling* (MDS) can provide a visual representation of the pattern of proximities among a set of objects. You are given the similarity information among 10,000 different DNA sequences. Describe how you can apply an MDS technique to produce a 2-D map that reflects their intrinsic relationship among 10,000 DNA sequences. It is essential to give main steps in your solution. (6 marks)

3. *Locally Linear Embedding* (LLE) is a popular manifold learning method.

   (a) Give those cost functions used in LLE for manifold learning. (4 marks)

   (b) Explain why LLE can be viewed as carrying out the principle of "*think globally, fit locally*" for manifold learning. (2 marks)

4. *Linear Discriminative Analysis* (LDA) is a popular supervised dimension reduction method. For a labelled data set $D$ of $C$ classes ($C > 2$), the *within-class* scatter matrix used in LDA is defined by

$$S_W = \sum_{c=1}^{C} \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^T,$$

where $D_i$ is the subset that contains all the training examples belonging to class $i$ and $\mathbf{m}_i = \frac{1}{|D_i|} \sum_{\mathbf{x} \in D_i} \mathbf{x}$. The *between-class* scatter matrix used in LDA is defined by

$$S_B = \sum_{c=1}^{C} |D_i| (\mathbf{m} - \mathbf{m}_i)(\mathbf{m} - \mathbf{m}_i)^T,$$

where $\mathbf{m} = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \mathbf{x}$. And the *total* scatter matrix is defined by

$$S_T = \sum_{\mathbf{x} \in D} (\mathbf{x} - \mathbf{m})(\mathbf{x} - \mathbf{m})^T.$$

Prove that the total scatter matrix can be expressed as sum of the within-class and the between-class scatter matrices in LDA; i.e., $S_T = S_W + S_B$. (9 marks)

**END OF EXAMINATION**