Two hours

QUESTION PAPER MUST NOT BE REMOVED FROM THE EXAM ROOM

UNIVERSITY OF MANCHESTER
SCHOOL OF COMPUTER SCIENCE

Modelling and visualization of high-dimensional data

Date:    Monday 23rd January 2012
Time:    14:00 - 16:00

Answer ALL Questions in Sections A and B and one question from Section C

For full marks your answers should be concise as well as accurate.
Marks will be awarded for reasoning and method as well as being correct.

This is a CLOSED book examination

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

[PTO]
Section A is restricted and cannot be published
SECTION B (Answer ALL parts)

(a) In the SOM learning, not only the winner unit is moving towards the training example, but also other units in the network move along. Explain (i) what is the relationship between the winner unit and those moving along? (ii) why that is an important property of the SOM learning?

(4 marks)

(b) For a 2-D data set, after the PCA, we use only the top principal component to produce a 1-D code for data compression. Describe a draw of two different 2-D data sets such that (i) for the first data set, 2-D reconstruction from their 1-D code is nearly identical to the original form of 2-D points, i.e., nearly perfect reconstruction, and (ii) for the second data set, 2-D reconstruction from their 1-D code is significantly different from the original form of 2-D points, i.e., very poor reconstruction. It is essential to draw schematic diagrams and explain why your data sets meet the above requirements.

(6 marks)

(c) Hand-written digit recognition is a typical classification task. You are asked to apply two dimensionality reduction techniques learnt from this course unit to extract two types of digit features (representations) that tend to capture salient features and further discuss their advantages and disadvantages. It is essential to give details on how to extract such features with your chosen techniques and explain why feature extraction is likely to improve both recognition accuracy and computational efficiency.

(10 marks)
SECTION C (Answer ONE question only from this section)

C1.

(a) Briefly describe what manifold learning is. Give the main cost functions used in the Locally Linear Embedding (LLE) algorithm for manifold learning and explain what roles they play for manifold learning.

(6 marks)

(b) One of weaknesses in LLE is that the nonlinear mapping achieved from manifold learning is only applicable to the training data set. Describe an idea to extend LLE to unseen data sets of similar manifolds during learning.

(4 marks)

(c) Making use of the derivation in achieving the top principal component learned from the course unit, prove that, in PCA, the linear projection onto an $M$-dimensional subspace that maximizes the variance of the projected data is defined by the $M$ eigenvectors of the covariance matrix of an $N$-dimensional data set corresponding to the $M$ largest eigenvalues ($N > M$).

(10 marks)

C2.

(a) Briefly describe what manifold learning is. Give the main cost functions used in the Isometric Feature Mapping (ISOMAP) algorithm for manifold learning and explain what roles they play for manifold learning.

(6 marks)

(b) One of weaknesses in ISOMAP is that the nonlinear mapping achieved from manifold learning is only applicable to the training data set. Describe an idea to extend ISOMAP to unseen data sets of similar manifolds during learning.

(4 marks)

(d) Given a data set of $K$ data points represented by a data matrix $X$, find an orthogonal matrix $P$ in $Y = PX$ such that $C_Y = \frac{1}{K}YY^T$ is a diagonal matrix and prove that the rows of $P$ are the principal components of $X$. [Hint: using the fact that a symmetric matrix is diagonalised by a matrix of its orthonormal eigenvectors; i.e., if $A$ is a symmetric matrix, $A = EDE^T$ where $D$ is a diagonal matrix and $E$ is a matrix of eigenvectors of $A$ arranged as columns.]

(10 marks)

END OF EXAMINATION