A22985

Calculators may be used in this examination provided they are not capable of being used to store alphabetical information other than hexadecimal numbers

UNIVERSITY OF BIRMINGHAM

School of Computer Science

Fourth Year Undergraduate/Postgraduate

20233

Intelligent Data Analysis (Extended)

Main Summer Examinations 2019

Time allowed: 1:30

[Answer all questions]

Note

Answer ALL questions. Each question will be marked out of 20. The paper will be marked out of 60, which will be rescaled to a mark out of 100.

Question 1 Principal Components Analysis (PCA)

(a) Calculate the covariance matrix of the set

$$X = \{\begin{bmatrix} 3 \\ -1 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \end{bmatrix}, \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \begin{bmatrix} -1 \\ -3 \end{bmatrix}, \begin{bmatrix} 4 \\ -2 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix}\}.$$

[4 marks]

(b) Describe the steps that are involved in the application of Principal Components Analysis (PCA) to a set of vectors X and explain how the result should be interpreted. **[6 marks]**

(c) A set X of 5-dimensional vectors has covariance matrix C with eigenvalue decomposition $C = UDU^T$, where:

$$D = \begin{bmatrix} 0.05 & 0 & 0 & 0 & 0 \\ 0 & 52.07 & 0 & 0 & 0 \\ 0 & 0 & 0.47 & 0 & 0 \\ 0 & 0 & 0 & 4.36 & 0 \\ 0 & 0 & 0 & 0 & 78.27 \end{bmatrix}, U = \begin{bmatrix} 0.01 & 0.01 & 0.02 & -0.99 & 0.09 \\ 0.01 & -0.03 & -0.97 & 0.01 & 0.26 \\ -0.01 & -0.69 & 0.21 & 0.06 & 0.69 \\ 0.77 & 0.45 & 0.11 & 0.04 & 0.43 \\ -0.63 & 0.56 & 0.12 & 0.05 & 0.52 \end{bmatrix}.$$

Write down the projection of the vector

$$v = \begin{bmatrix} 2 \\ 1 \\ -3 \\ 1 \\ 0 \end{bmatrix}.$$

onto the first two principal components of the data set X. **[4 marks]**

(d) What are the advantages and disadvantages of Principal Components Analysis (PCA) and Linear Discriminant Analysis (LDA) for dimension reduction and visualisation of high-dimensional data? **[6 marks]**

Question 2 Mining textual data

(a) Statistical Analysis of documents

(i) A document comprises a total of 185,000 words, from a vocabulary of 15,800 different words. According to Zipf's Law, what percentage of the vocabulary words occur less than 10 times in the document? **[4 marks]**

(b) TF-IDF similarity

(i) A text corpus consists of four documents $\{d_1, d_2, d_3, d_4\}$ and (after text pre-processing, stop-word removal and stemming) six terms $\{t_1, t_2, t_3, t_4, t_5, t_6\}$. The number of times that each term occurs in each document is given in the following table:

	t_1	t_2	t_3	t_4	t_5	t_6
d_1	1	0	1	1	0	1
d_2	0	2	0	1	0	3
d_3	2	0	1	2	2	1
d_4	0	1	0	0	0	1

Calculate the TF-IDF similarity $sim(d_1, d_3)$ between documents d_1 and d_2. **[6 marks]**

(c) Vector representation of documents

(i) What is the vector representation $vec(d)$ of a document d? **[4 marks]**

(ii) Explain how Latent Semantic Analysis can be used to uncover hidden relationships between terms. **[6 marks]**

Question 3 Clustering

(a) k-means clustering

(i) Describe the steps involved in the k-means clustering algorithm. **[4 marks]**

(ii) Given a data set X, is the k-means algorithm guaranteed to find a set of centroids C that minimizes the distortion $D(C, X)$ between C and X? If not then explain why the algorithm is not optimal and what factors influence the solution that is obtained? **[4 marks]**

(b) Vector Quantization (VQ)

(i) Explain how Vector Quantization is applied to low bit rate speech coding in a CELP (Codebook Excited Linear Prediction) speech coder. What properties of speech does it exploit to achieve low bit-rates? **[6 marks]**.

(c) Topographic Maps

The update rule for a set of centroids $\{c^1, \cdots, c^J\}$ in a topographic map (self-organizing map), given the data point x is

$$c^j_{new} = c^j_{old} + h[win(x), j] \times \eta \times (x^i - c^j_{old})$$

where $win(x)$ is the index of the closest centroid to x.

(i) Describe the purpose of the function h and its practical application. **[6 marks]**

This page intentionally left blank.

Do not complete the attendance slip, fill in the front of the answer book or turn over the question paper until you are told to do so

Important Reminders

- Coats/outwear should be placed in the designated area.
- Unauthorised materials (e.g. notes or Tippex) must be placed in the designated area.
- Check that you do not have any unauthorised materials with you (e.g. in your pockets, pencil case).
- Mobile phones and smart watches **must** be switched off and placed in the designated area or under your desk. They must not be left on your person or in your pockets.
- You are not permitted to use a mobile phone as a clock. If you have difficulty seeing a clock, please alert an Invigilator.
- You are not permitted to have writing on your hand, arm or other body part.
- Check that you do not have writing on your hand, arm or other body part – if you do, you must inform an Invigilator immediately
- Alert an Invigilator immediately if you find any unauthorised item upon you during the examination.

Any students found with non-permitted items upon their person during the examination, or who fail to comply with Examination rules may be subject to Student Conduct procedures.