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

Intelligent Data Analysis

Exam including model answers!

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 66, 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]**

Model answer / LOs / Creativity:

(a) Using the standard formula for the covariance matrix the correct answer is:

$$\begin{bmatrix} 3.2 & 0.8 \\ 0.8 & 6.4 \end{bmatrix}$$

Note that this is with Bessel's correction. I deducted 1 mark for dividing by N rather than $N - 1$

(b) The steps involved in applying PCA are as follows:

(i) Calculate the covariance matrix of X. For example, calculate the mean vector and subtract it from each sample vector. Arrange the resulting vectors as the rows of a matrix Y and compute $C = \frac{1}{N-1} Y^T Y$. **[2 marks]**

(ii) Apply eigenvalue decomposition to C to get $C = UDU^T$, where U is an orthogonal matrix and D is a real diagonal matrix. **[2 marks]**

(iii) The columns of U form a new orthonormal basis for the vector space. For each column u_i of U the corresponding diagonal element d_{ii} of D indicates the variance of the data in the direction of u_i. Hence the first principal component is the vector u_i that corresponds to the biggest eigenvalue d_{ii}. The second principal component corresponds to the next biggest eigenvalue, and so on. **[2 marks]**

(c) The biggest and second biggest eigenvalues are $d_{55} = 78.27$ and $d_{22} = 52.07$. Hence the principal components are the 5^{th} and 2^{nd} columns of U. Hence the projection of v onto the first two principal components is:

$$\begin{bmatrix} v \cdot u_5 \\ v \cdot u_2 \end{bmatrix} = \begin{bmatrix} -1.2 \\ 2.51 \end{bmatrix}$$

(d) The answer should include the following (4 marks for (i) and two marks for either (ii) or (iii))

(i) In cases where a m-dimensional object is embedded in an n-dimensional space V, with $n > m$, both PCA and LDA will only reveal the underlying dimension of the object if the embedding is approximately linear so that the object can be shifted by mean normalization to lie in an m-dimensional vector subspace of V. If the embedding is non-linear then something like a Self-Organizing Map would be better.

(ii) LDA is only applicable in cases where the data points are assigned to classes. Where PCA finds the m dimensions which contain as much as possible of the variance of the data, LDA finds the m dimensions which best separate the classes of data. It it easy to construct examples to show that these are not the same.

(iii) The covariance matrix is a symmetric $n \times n$ matrix. If the dimension n of the vector space is large and the number of samples is relatively small, then it may not be possible to accurately estimate the covariance matrix in PCA or the average within-class covariance matrix in LDA.

(iv) For visualization the data is projected onto the 2-dimensional subspace spanned by the eigenvectors corresponding to the two largest eigenvalues, corresponding to the two dimensions in which the variance of the data is greatest. In some cases, for example where the data belongs to distinct classes and the objective is to separate the classes, LDA may provide a more useful and informative visualization than PCA.

Learning outcomes: Parts (a), (b) and (c) test the student's understanding of how PCA is implemented. Part (d) examines creativity by allowing the student to demonstrate his or her understanding of the strengths and weaknesses of the technique.

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]**

Model answer / LOs / Creativity:

(a) Zipf's Law says the the rank-frequency distribution $F(r)$ for words in a large document is approximately:

$$F(r) = \frac{C}{r^\alpha}$$

where $C \approx 0.1$ and $\alpha \approx 1$. Taking these values of C and α, the number of occurrences of the r-ranked word in the vocabulary is approximately

$$\frac{185,000 \times 0.1}{r}$$

For this to be less than 10 we need $r \geq 1851$. Hence the percentage of vocabulary words which occur 10 times or less is

$$\frac{15800 - 1851}{15800} = 0.8828$$

Hence the answer is approximately 88%.

(b) From the table the inverse document frequencies of the terms are:

$$\begin{aligned} IDF(t_1) &= log_e(\frac{N}{N_{t_1}}) = log_e(\frac{4}{2}) = 0.69 \\ IDF(t_2) &= IDF(t_3) = 0.69 \\ IDF(t_4) &= 0.29 \\ IDF(t_5) &= 1.39 \\ IDF(t_6) &= 0 \end{aligned} \tag{1}$$

Hence the TF-IDF weights, given by $w_{t,d} = f_{f,d} \times IDF(t)$ are:

	t_1	t_2	t_3	t_4	t_5	t_6
d_1	0.69	0	0.69	0.29	0	0
d_2	0	1.39	0	0.29	0	0
d_3	1.39	0	0.69	0.58	2.77	0
d_4	0	0.69	0	0	0	0

and the document lengths are:

$$||d_1|| = 1.02, ||d_2|| = 1.42, ||d_3|| = 3.23, ||d_4|| = 0.69$$

Hence,

$$sim(d_1, d_3) = \frac{\sum_{t \in d_1 \cap d_3} w_{t,d_1} \times w_{t,d_3}}{||d_1|| \times ||d_3||} = 0.487$$

Because of the typo in the question I also full marks for

$$sim(d_1, d_2) = \frac{\sum_{t \in d_1 \cap d_2} w_{t,d_1} \times w_{t,d_2}}{||d_1|| \times ||d_2||} = 0.057.$$

(c) (i) Suppose that the corpus has J significant terms $\{t_1, \cdots, t_J\}$. By "significant term" I mean the different vocabulary items that remain after text pre-processing, but I gave full marks for "Suppose the corpus has a vocabulary of J distinct terms". The vector representation $vec(d)$ of a document d is the J dimensional vector whose j^{th} element is $w_{t_j,d}$.

(ii) i. In LSA Singular Value Decomposition (SVD) is applied to the word-document matrix W to decompose W into $W = USV^T$, where U is a $N \times N$ orthogonal matrix, S is a $N \times J$ real-valued matrix such the $s_{ij} = 0 if i \neq j$ and $s_{ii} \geq s_{jj}$ if $i \geq j$, and V is a $J \times J$ orthogonal matrix.

ii. The columns of the matrix V are J-dimensional "document vectors" which describe important hidden semantic classes in the corpus. The singular value s_{ii} indicates the importance of the semantic class corresponding to the i^{th} column of V. In fact, $\frac{s_{ii}^2}{N-1}$ is the variance of the set of document vectors in the direction of the unit vector corresponding to the i^{th} column of V.

iii. Each term t can be represented as a "one hot" document vector $vec(t)$ with 1 in the entry corresponding to t and all other entries 0. The vector $top(t) = V^T vec(t)$ is the projection of $vec(t)$ onto the LSA basis vectors (the "hidden semantic classes"). Intuitively, if two terms t_1 and t_2 are semantically related then the vectors $top(t_1)$ and $top(t_2)$ will point in similar directions. Thus the cosine similarity between $top(t_1)$ and $top(t_2)$ is a measure of the similarity of meaning of the two terms.

Learning outcomes: Parts (a) and (b) demonstrate the student's understanding of word frequency in texts and the basic calculations involved in computing the TF-IDF similarity between two documents. The more creative part is part (c) which tests deeper understanding of LSA.

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]**

Model answer / LOs / Creativity:

(a) k-means clustering for a set of samples $X = \{x_1, \cdots, x_N\}$

(i) i. Choose the number of centroids K and an initial set of centroids $\{c^{(0)}_1, \cdots, c^{(0)}_K\}$

ii. Given centroids $\{c^{(i)}_1, \cdots, c^{(i)}_K\}$, calculate the distance $d(x_n, c_k)$ for each sample x_n and each centroid c_k.

iii. For each centroid $c^{(i)}_k$ identify the set $X(k) = \{x_n \in X | d(x_n, c^{(i)}_k) \leq d(x_n, c^{(i)}_j) \forall j\}$. This is the set of samples for which $c^{(i)}_k$ is the closest centroid.

iv. Define

$$c^{(i+1)}_k = \frac{1}{|X(k)|} \sum_{x_n \in X(n)} x_n$$

in other words $c^{(i+1)}_k$ is the average of the set of samples which are closest to $c^{(i)}_k$.

v. Set $i = i + 1$. The process stops if i reaches the maximum number of permitted iterations or the difference between the i^{th} and $(i+1)^{th}$ centroid sets falls below some threshold. Otherwise return to step (ii)

(ii) The k-means algorithm is a gradient descent algorithm. Given a set of centroids $C^{(i)} = \{c_1^{(i)}, \cdots, c_K^{(i)}\}$ it will find a new set of centroids $C^{(i+1)} = \{c_1^{(i+1)}, \cdots, c_K^{(i+1)}\}$ such that $D(C^{(i)}, X) \geq D(C^{(i+1)}, X)$. Thus the solution found by the k-means algorithm will be a local minimum close to the initial centroid set $C^{(0)}$. A different choice of initial centroid set $C^{(0)}$ with result, potentially, in a different minimum.

(b) Low bit-rate speech coding

(i) CELP coding relies on the fact that the vocal tract moves relatively slowly and therefore its properties only need to be sample relatively infrequently - say 100 sample per second instead of $16,000$ samples per second for a raw speech waveform - and that the vocal tract filter can be encoded using a small number (approximately 10) integers.

(ii) CELP coding uses the source-filter model of speech production, where a speech signal is created by passing an excitation (either the signal from the vocal cords for a voiced sound, or noise for an unvoiced sound) through a filter (determined by the shape of the vocal tract)

(iii) The vocal tract filter is estimated using a technique called Linear Prediction (LP). The results of LP can be encoded in approximately 10 measurements.

(iv) Good quality speech requires a good estimate of the exitation signal. This is achieved by constructing a Vector Quantizer from a set of example excitation signals. At the start of coding the codebook is transmitted to the receiver. Subsequently, each excitation signal is replaced by the closest VQ symbol, and the index of that symbol is transmitted alongside the 10 LP coefficients.

(v) This is Codebook Excited Linear Prediction (CELP). It is the basis of speech coders in mobile phones.

(c) SOMs

(i) The function h defines the neighbourhood structure on the set of centroids. When a sample x_n is presented to the SOM during training, the closest centroid $c_{i(n)}$ to x_n is identified and $c_{i(n)}$ is moved towards x_n. In addition, a "neighbour" c_j of $c_{i(n)}$ is also moved towards x_n by an amount proportional to $h(i(n), j)$. As the iterative learning process proceeds the width of these neighbourhoods is decreased. In the example in class,

$$h(i(n), j) = e^{-\frac{|i(n)-j|}{\sigma}}$$

where σ is the neighbourhood width. The neighbourhood width is made smaller during training by setting

$$\sigma(t) = \sigma(0) \times e^{-\frac{t}{\nu}}$$

, where $\nu > 0$ is the time scale and $\sigma(0)$ is the initial neighbourhood width.

Learning outcomes: Part (a) tests the student's knowledge of how the k-means clustering algorithm works. and part (b) tests the student's understanding of a real application of VQ to

speech coding that he or she uses every day. Part (c) tests the student's deeper understanding of how a Self-Organizing Map works.

End of Paper

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.