UNIVERSITY OF BRISTOL

May/June 2021 Examination Period

FACULTY OF ENGINEERING

Second Year Examination for the Degree of
Bachelor of Science and Master of Engineering

COMS20011J
Data-Driven Computer Science

TIME ALLOWED:
2 Hours

This paper contains 15 questions.
Each question has exactly one correct answer.
All answers will be used for assessment.
The maximum for this paper is 100 marks.

Other Instructions:

TURN OVER ONLY WHEN TOLD TO START WRITING

Help Formulas:

Minkowski distance:

$$D(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

One-dimensional Gaussian/Normal probability density function:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Multi-dimensional Gaussian/Normal probability density function:

$$p(\mathrm{x}) = \frac{1}{\sqrt{(2\pi)^M |\Sigma|}} e^{-\frac{1}{2}(\mathrm{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathrm{x}-\boldsymbol{\mu})}$$

Least Squares Matrix Form:

$$\mathrm{a}_{LS} = (\mathrm{X}^T \mathrm{X})^{-1}\ \mathrm{X}^T\ \mathrm{y}$$

Matrix inversion:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Matrix Determinant:

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

Q1. The eigenvalues of a dataset are: [40, 21, 11, 5, 1, 0.88, 0.33]. Approximately what variance in the dataset do the first 3 eigenvalues represent?

A. 90.1%

B. 90.9%

C. 96.5%

D. 95.6%

E. 92.3%

[6 marks]

Q2. A 5x5 spatial filter has all its elements set to -0.75, except for the central element which is set to 5.0. It must then have a:

A. normalisation factor of $\frac{1}{23}$

B. normalisation factor of $\frac{1}{18}$

C. normalisation factor of $\frac{-1}{18}$

D. normalisation factor of $\frac{1}{13}$

E. normalisation factor of $\frac{-1}{13}$

[4 marks]

Q3. Which is the correct 2-norm distance (L2) between datapoints A and B, where A=(2,4,3,6,8) and B=(3,5,5,7,1):

A. L2(A,B) = $\sqrt{56}$

B. L2(A,B) = 2

C. L2(A,B) = 56

D. L2(A,B) = $\sqrt{2}$

E. None of the above

[4 marks]

Q4. Two eigenvalues of the covariance matrix below are approximately -0.873 and 1.646:

$$\begin{bmatrix} 3.5 & -6.0 & 2.0 \\ -6.0 & 8.25 & -4.5 \\ 2.0 & -4.5 & 3.5 \end{bmatrix}$$

What is the third eigenvalue?

A. 12.731

B. 14.477

C. 3.50

(cont.)

D. 4.5

E. 17.769

[6 marks]

Q5. You are given a three-dimensional data set, where each sample is a three-dimensional vector $\mathbf{x} = (x_1, x_2, x_3)$, with the following covariance matrix:

$$\begin{bmatrix} 4 & -1.7 & 0.6 \\ -1.7 & 4 & -2.5 \\ 0.6 & -2.5 & 5 \end{bmatrix}$$

Which of the following conclusions cannot definitively be demonstrated by the covariance matrix?

A. x_3 has the highest variance

B. x_1 has a stronger correlation with x_2 than x_3

C. x_2 has a negative correlation with x_3

D. x_1 and x_2 have an equal mean

E. x_3 has a positive correlation with x_1

[6 marks]

Q6. Figure 1 shows handwritten graffiti type letters B, M, and V which are correspondingly labelled (B, M, V).

B *M* *V*

Figure 1: Handwritten images of the letters B, M, and V

Below in Figure 2, there are three results, labelled $(1, 2, 3)$ that represent, in an arbitrary order, the FFT of the images in Figure 1. Select the choice that correctly states which FFT image corresponds to which graffiti image, using the image labels.

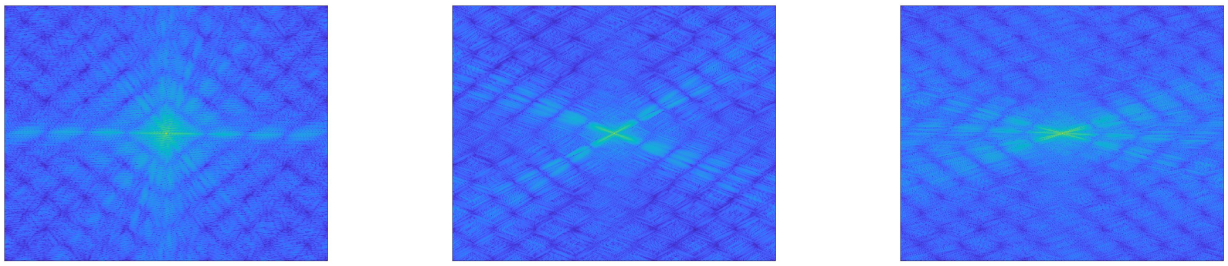
A. $(1, 2, 3)$ corresponds to (B, M, V)

B. $(1, 2, 3)$ corresponds to (M, V, B)

C. $(1, 2, 3)$ corresponds to (M, B, V)

D. $(1, 2, 3)$ corresponds to (B, V, M)

E. $(1, 2, 3)$ corresponds to (V, B, M)



1 2 3

Figure 2: FFT results

[8 marks]

Q7. Using the {Delete,Insert,Substitute} operations, what is the minimum Edit Distance between the words "INTENTION" and "EXECUTION"?

A. 4

B. 6

C. 5

D. 7

E. 3

[8 marks]

Q8. Figure 3 shows an image of Einstein and its Fourier Transform output.

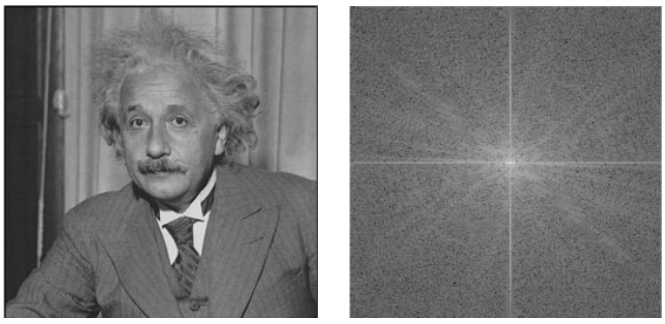
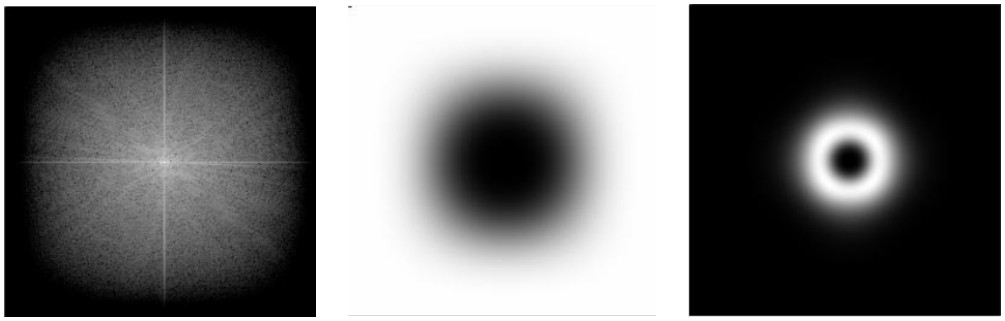


Figure 3: Einstein image and its FFT space.

Below in Figure 4, the top row shows three differently filtered versions of the Einstein FFT space, labelled ($F1, F2, F3$). The three images in the bottom row labelled ($R1, R2, R3$) show in an arbitrary order, the inverse FFT results of those filtered Fourier outputs. Select the choice that correctly states which inerse FFT image corresponds to which filtered version of the original FFT space of the Einstein image.

(cont.)



$F1$ $F2$ $F3$

$R1$ $R2$ $R3$

Figure 4: (top row) Filters that were applied to the FFT result of the Einstein image, and (bottom row) Inverse FFT results in arbitrary order.

A. $(F1, F2, F3)$ corresponds to $(R1, R2, R3)$

B. $(F1, F2, F3)$ corresponds to $(R2, R1, R3)$

C. $(F1, F2, F3)$ corresponds to $(R1, R3, R2)$

D. $(F1, F2, F3)$ corresponds to $(R3, R2, R1)$

E. $(F1, F2, F3)$ corresponds to $(R3, R1, R2)$

[8 marks]

Q9. For training data displayed in the table, and a model of the form $y = w_0 + w_1 x$, compute the maximum-likelihood straight line,

x	y
-1.0	10.3
-0.3	5.3
0.3	-0.2
1.0	-5.3

A. $y = 1.32 - 5.76x$

B. $y = 2.53 - 7.91x$

C. $y = 3.57 - 9.32x$

D. $y = 3.85 - 8.56x$

E. $y = 3.62 - 8.94x$

[7 marks]

Q10. Given the test set in the table, which model has the lowest test sum-squared-error?

x	y
-2.0	-0.3
0.0	4.2
2.0	7.5

A. $y = 4$

B. $y = 2x + 4$

C. $y = 3x + 2$

D. $y = -0.3x^2 + 2x + 2$

E. $y = x^3 - 0.3x^2 + 2x + 2$

Q11. Which statement is FALSE?

A. Regularisation penalises very large values of the weights

B. Regularisation mitigates overfitting

C. Regularisation is particularly useful when fitting a complex function

D. Regularisation always improves test performance

E. We can use cross-validation to choose the strength of regularisation

Q12. When is overfitting least likely to be a serious issue?

A. when fitting a high-order polynomial

B. when fitting a complex nonlinear function with many parameters

(cont.)

C. when input data, x, is a high-dimensional vector

D. when little training data is available

E. when fitting a straight line (i.e. $y = w_0 x + w_1$) with lots of training data

Q13. Classify the points $x_1 = 1.1$, $x_2 = 0.0$ using K nearest neighbour, where $K = 1$, and using the training data in the table.

x	y
-1.0	0
-1.9	0
1.3	0
0.5	1
2.7	1
2.3	1

A. $y_1 = 0; y_2 = 0$

B. $y_1 = 0; y_2 = 1$

C. $y_1 = 1; y_2 = 0$

D. $y_1 = 1; y_2 = 1$

Q14. In K-means, consider initializing the algorithm with two cluster centers at -3 and 3, and data at,

x
-4.2
-3.6
-3.9
1.8
2.4
1.7

Compute updated cluster centers under a full K-means update

A. -3.90 and 1.97

B. -4.13 and 1.85

C. -2.73 and 1.53

D. -2.32 and 1.45

E. -2.54 and 2.32

Q15. Consider a classifying documents as relating to financial news using Naive Bayes with

$$P(\text{"results"}|\text{financial}) = 0.8 \quad (1)$$
$$P(\text{"revenue"}|\text{financial}) = 0.7 \quad (2)$$
$$P(\text{"profit"}|\text{financial}) = 0.9 \quad (3)$$

$$P(\text{"results"}|\text{not financial}) = 0.7 \quad (4)$$
$$P(\text{"revenue"}|\text{not financial}) = 0.2 \quad (5)$$
$$P(\text{"profit"}|\text{not financial}) = 0.3 \quad (6)$$

$$P(\text{financial}) = 0.5 \quad (7)$$
$$P(\text{not financial}) = 0.5 \quad (8)$$

Compute $P(\text{financial}|\text{"results", "revenue""profit"})$

A. 0.90

B. 0.92

C. 0.95

D. 0.97

E. 0.99

END OF PAPER