UNIVERSITY OF BRISTOL

May/June 2022 Examination Period

FACULTY OF ENGINEERING

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

COMS20011
Data-Driven Computer Science

TIME ALLOWED:

Answers to COMS20011: Data-Driven Computer Science

Intended Learning Outcomes:

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:

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

Least Squares Matrix Form:

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

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. Choose which word is best suited to fill in the blank in this sentence:

[———] *refers to a model that can neither represent the training data nor generalise to new data.*

A. Decision Boundary

B. Poor Correlation

C. Underfitting

D. Overfitting

E. None of the above

[2 marks]

Solution: C - An underfit model will have high training and high testing error as it does not produce a good decision boundary and does not fit the data well

Q2. You are given a three-dimensional data set, where each sample is a three-dimensional vector $\mathbf{v} = (v_1, v_2, v_3)$, with the following covariance matrix:

$$\begin{bmatrix} 7 & 2.4 & -3.3 \\ 2.4 & 3 & 0.5 \\ -3.3 & 0.5 & 7 \end{bmatrix}$$

Which of the following conclusions can definitively be demonstrated by the covariance matrix?

A. v_1 has stronger correlation with v_2 than v_3

B. v_3 has a positive correlation with v_2

C. v_2 has a negative correlation with v_1

D. v_1 and v_3 have the same mean

E. v_2 has the highest variance

[7 marks]

Solution: B - is the only fact that is demonstrated by the covariance matrix.

Q3. What does the term 'Outlier' mean?

A. A value that is left out of the analysis because of missing data

B. When two data points are the same and one can be discarded

C. A type of variable that is not easy to quantify

(cont.)

D. An extreme value at either end of a distribution

E. None of the above

[2 marks]

Solution: D - an outlier is a point with a value significantly different from that of the other points

Q4. The eigenvalues of a dataset are: [23.0, 18.0, 10.0, 8.0, 6.0, 2.0, 0.77, 0.15]. Approximately what variance in the dataset do the first 4 eigenvalues represent?

A. 90.1%

B. 89.2%

C. 91.0%

D. 59.0%

E. 86.9%

[6 marks]

Solution: E - Sum of the first 4 eigenvalues divided by the sum of all the eigenvalues, multiplied by 100 and rounded to 1 decimal point.

Q5. Which of the following statements is TRUE:

A. The central regions of the Fourier space represent the contrast in the image and are used for smoothing.

B. The outer regions of the Fourier space represent the detail in the image and are used for smoothing.

C. The outer regions of the Fourier space represent the contrast in the image and are used for sharpening.

D. The central regions of the Fourier space represent the detail in the image and are used for sharpening.

E. The central regions of the Fourier space represent the detail in the image and are used for smoothing.

[3 marks]

Solution: A

Q6. Compute the Hamming and Edit distances between these two strings: 'datascience' and 'datedscience'

A. Hamming = 2, Edit distance = 3

B. Hamming = 2, Edit distance = 2

C. Hamming = 1, Edit distance = 3

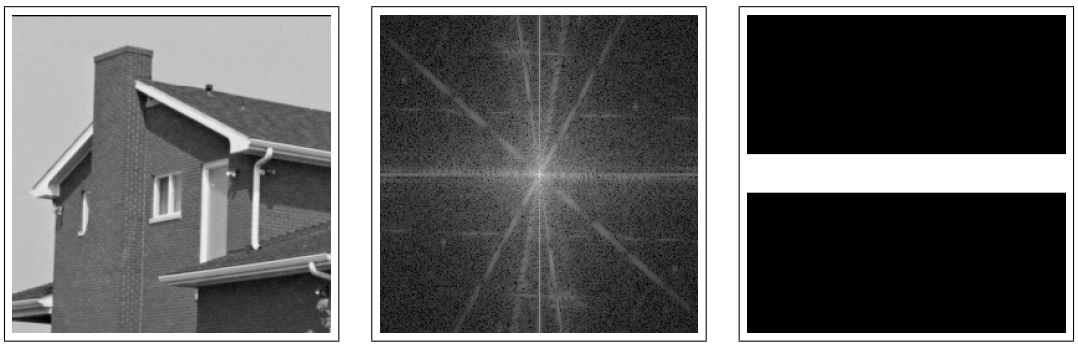
D. Hamming = 1, Edit distance = 2

E. None of the above

[4 marks]

Solution: E - For Hamming distance, the strings must be of the same length. For the Edit Distance, just 1 insertion would be the only operation needed.

Q7. Figure 1 shows the original image of a House on the left. The Fast Fourier Transform was applied to generate the Fourier space shown in the middle. The right image shows a mask where the white area is 1 and the black area is 0.



House Fourier Space Mask

Figure 1: (top) Original House image, its Fourier Space, and the mask used to apply to the Fourier Space.

Which of the images in Figure 2 is the correct result AFTER applying the mask in Figure 1 to the Fourier Space of the House image in Figure 1 and then performing an inverse Fourier Transform?

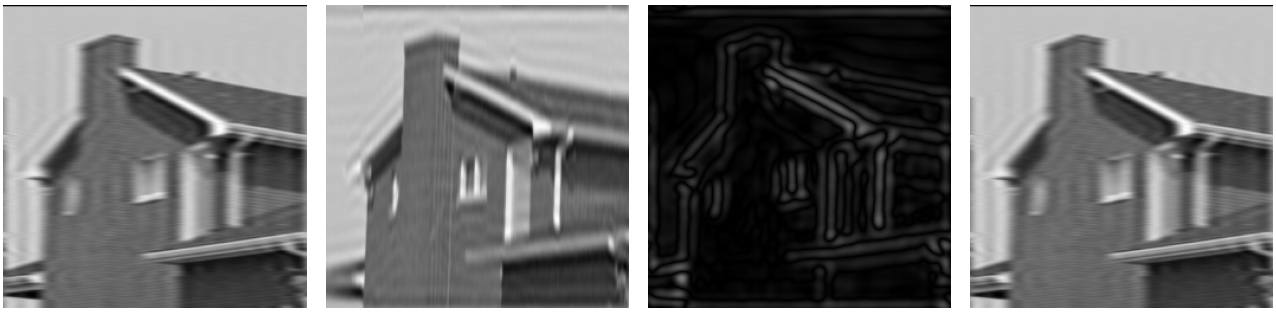
A. Image *A*

B. Image *B*

C. Image *C*

D. Image *D*

E. None of these images.



A *B* *C* *D*

Figure 2

[7 marks]

Solution: B - there is stronger evidence that horizontal and near horizontal lines have been affected in this image while the rest of the information is well preserved. For example see top ledge of the window.

Q8. Matrix $M = \begin{pmatrix} 8 & -6 & 2 \\ -6 & 7 & -4 \\ 2 & -4 & 3 \end{pmatrix}$, has three eigenvalues, two of which are 3 and 0. What is the third eigenvalue?

A. 15

B. 18

C. 21

D. 12

E. 0

[5 marks]

Solution: A - Sum of variances = sum of eigenvalues, so 8+7+3=3+0+15

Q9. The following is a simple audio signal:

$$s = \begin{pmatrix} 0 & 0 & 10 & 20 & 30 & 25 & 15 & 10 & 0 & 0 \end{pmatrix}.$$

Ignoring normalisation, which of these below is the correct result if the signal is convolved with the filter $g = (-1 \quad 5 \quad -1)$?

A. $s * g = \begin{pmatrix} 0 & 0 & 10 & 30 & 60 & 105 & 80 & 40 & 35 & 10 & 0 & 0 \end{pmatrix}$

B. $s * g = \begin{pmatrix} 0 & 0 & 10 & 30 & 60 & 105 & 80 & 40 & 35 & -10 & 0 & 0 \end{pmatrix}$

C. $s * g = \begin{pmatrix} 0 & 0 & -10 & 30 & 60 & 105 & 80 & 40 & 35 & 10 & 0 & 0 \end{pmatrix}$

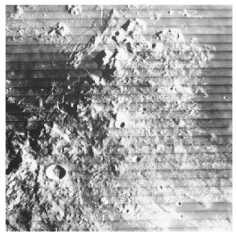
D. $s * g = \begin{pmatrix} 0 & 0 & -10 & 30 & 60 & 105 & 80 & 40 & 35 & -10 & 0 & 0 \end{pmatrix}$

E. $s * g = \begin{pmatrix} 0 & 0 & -10 & 35 & 60 & 105 & 80 & 40 & 30 & -10 & 0 & 0 \end{pmatrix}$

[7 marks]

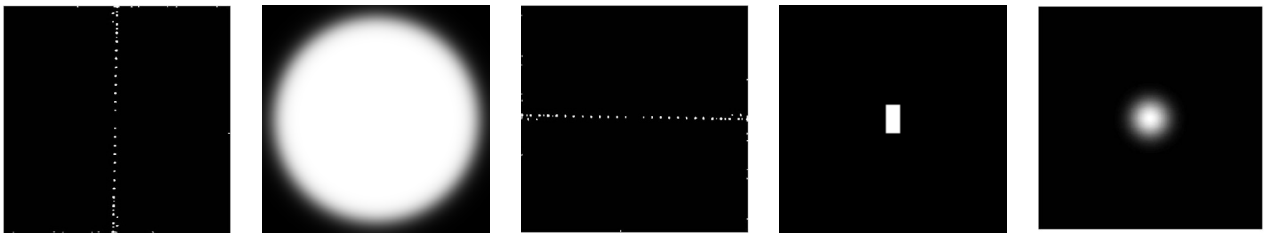
Solution: D

Q10. Consider the satellite image below (Moon Surface), of a small area of the surface of the Moon, which has suffered from some linear noise during transmission back to Earth. The image below-right (Cleaned-Image) is a cleaned-up version after the Fourier space of the original image was altered using a special mask. The second row of the figure shows 5 masks, labelled (A, B, C, D, E), one of which was used to produce the cleaned-up image.



Moon Surface

Cleaned Image



A *B* *C* *D* *E*

Figure 3: (top) Original Moon Surface image and its cleaned version, (bottom) Five possible masks applied to the Fourier space of the Moon Surface image.

Which of the above masks was used to clean the image?

A. Mask *A*

B. Mask *B*

C. Mask *C*

D. Mask *D*

E. Mask *E*

[7 marks]

(cont.)

Solution: A - the range of frequencies masked out correspond to the horizontal noise lines in the image.

Q11. We have N datapoints, $x_1, \ldots, x_N$, distributed according to,

$$P(x_i|\alpha,\beta) \propto \beta^{\alpha} x_i^{\alpha-1} e^{-\beta x_i} \tag{1}$$

What is the maximum-likelihood solution for β for known α?

A. $\frac{1}{\alpha}(\frac{1}{N}\sum_{i=1}^{N} x_i)$

B. $\frac{\alpha}{N}\sum_{i=1}^{N} x_i$

C. $\frac{1}{\alpha}(\frac{1}{N}\sum_{i=1}^{N} \log x_i)$

D. $\frac{\alpha}{\frac{1}{N}\sum_{i=1}^{N} x_i}$

E. $\frac{\alpha}{N}\sum_{i=1}^{N} e^{x_i}$

[6 marks]

Solution:

$$\log P(x|\alpha,\beta) = \sum_{i=1}^{N}(\alpha \log\beta + (\alpha-1)\log x_i - \beta x_i) \tag{2}$$

$$= N\alpha\log\beta + (\alpha-1)\sum_{i=1}^{N}\log x_i - \beta\sum_{i=1}^{N} x_i \tag{3}$$

$$0 = \frac{\partial}{\partial\beta}\log P(x_i|\alpha,\beta) = N\frac{\alpha}{\beta} - \sum_{i=1}^{N} x_i \tag{4}$$

$$\frac{\alpha}{\beta} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{5}$$

$$\beta = \frac{\alpha}{\frac{1}{N}\sum_{i=1}^{N} x_i} \tag{6}$$

Q12. We have N datapoints, $x_1, \ldots, x_N$, distributed according to,

$$P(x_i|\mu,\theta) = \frac{1}{2\theta}\exp(-\tfrac{|x_i-\mu|}{\theta}) \tag{7}$$

What is the maximum-likelihood solution for θ?

A. $\theta = \frac{1}{N}\sum_{i=1}^{N}\text{sign}(x_i-\mu)$.

B. $\theta = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$.

C. $\theta = \frac{1}{2}\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2}$.

D. $\theta = \frac{1}{2N}\sum_{i=1}^{N}|x_i-\mu|$.

(cont.)

E. $\theta = \frac{1}{N}\sum_{i=1}^{N}|x_i - \mu|$.

[6 marks]

Solution:

$$\log P(x|\mu,\theta) = -N\log 2 - N\log\theta - \tfrac{1}{\theta}\sum_i |x_i - \mu| \tag{8}$$

$$0 = \frac{\partial}{\partial\theta}\log P(x_i|\mu,\theta) = -\tfrac{N}{\theta} + \tfrac{1}{\theta^2}\sum_i |x_i - \mu| \tag{9}$$

$$\theta = \tfrac{1}{N}\sum_i |x_i - \mu|. \tag{10}$$

Q13. For the data in the table, fit a model of the form $\hat{y} = w_0 + w_1 x$

x	y
-2.1	-4.2
-0.9	-2.3
0.2	-0.1
1.2	2.1
2.4	3.9

[4 marks]

A. $w_0 = -0.40,\ w_1 = 1.89$.

B. $w_0 = -0.39,\ w_1 = 1.87$.

C. $w_0 = -0.36,\ w_1 = 1.91$.

D. $w_0 = -0.42$**,** $w_1 = 1.85$**.**

E. $w_0 = -0.32,\ w_1 = 1.99$.

Q14. For the data in the table, fit a model of the form $\hat{y} = w_1 x + w_2 x^2$

x	y
-2.1	-4.2
-0.9	-2.3
0.2	-0.1
1.2	2.1
2.4	3.9

[4 marks]

A. $w_1 = 1.83,\ w_2 = -0.12$.

B. $w_1 = 1.85,\ w_2 = -0.06$.

C. $w_1 = 1.81,\ w_2 = -0.03$.

D. $w_1 = 1.87$**,** $w_2 = -0.10$**.**

E. $w_1 = 1.89,\ w_2 = 0.01$.

Q15. Which of these is the key, defining feature of overfitting?

A. Better train than test performance.

B. Predictions vary rapidly as a function of x.

C. Complex function class.

D. No regularisation.

E. Poor train and test performance.

[5 marks]

Q16. Compute $\sum_i \log P(y_i|x_i)$ for binary classification, where

$$P(y_i = 1|x_i = 1) = \sigma(-3 + x_i + 2x_i^2) \tag{11}$$

with data,

x	y
-2.1	0
-0.9	0
0.2	0
1.2	1
2.4	1

[5 marks]

A. -4.18

B. -4.20

C. -4.22

D. -4.24

E. -4.26

Q17. For the following data,

x	y
-2.1	0
-0.9	0
0.2	0
1.2	1
2.4	1

Fit a Gaussian distribution to each class, and compute the posterior probability that $x = 1.3$ is in class 1, given a prior of $P(y = 1) = 0.4$.

A. 0.93

B. 0.95

C. 0.97

D. 0.99

(cont.)

E. 1.00

[5 marks]

Q18. You get a sample of spam and not spam messages, and count the number times various words appear:

word	frequency in spam	frequency in not-spam
send	123	13
money	325	12
work	32	102

Use maximum likelihood to find the best model for words in spam and not spam messages, then compute the probability of spam for this message, "send money send money work". Assume a uniform prior.

A. 0.959

B. 0.961

C. 0.964

D. 0.971

E. 0.975

[5 marks]

Q19. In K-means clustering, for the following data, x and cluster assignments, z, perform a (hard) M-step to give cluster centres and perform an E-step to compute the resulting cluster-assignments.

x	z
-2.1	0
-0.9	1
0.2	0
1.2	1
2.4	0

A. $0, 1, 1, 1, 1$

B. $1, 1, 1, 0, 0$

C. $0, 0, 0, 1, 1$

D. $0, 0, 1, 1, 1$

E. $1, 1, 0, 0, 0$

[5 marks]

Q20. Use Bayesian inference to compute $P(A = 1|B = 1)$, where $A = 0$ or 1, and $B = 0$, 1, or 2. We have a uniform prior over A, and $P(B|A)$ is given by the table,

	$B = 0$	$B = 1$	$B = 2$
$A = 0$	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{6}$
$A = 1$	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{1}{4}$

A. $\frac{2}{5}$
B. $\frac{4}{5}$
C. $\frac{3}{5}$
D. $\frac{3}{4}$
E. $\frac{2}{3}$

[5 marks]

Solution: The normalizing constant is,

$$P(B=1) = \sum_A P(B|A)P(A) \tag{12}$$
$$= \tfrac{1}{2}(P(B=1|A=0) + P(B=1|A=1)) \tag{13}$$
$$= \tfrac{1}{2}(\tfrac{1}{3} + \tfrac{1}{2}) \tag{14}$$
$$= \tfrac{1}{2}(\tfrac{2}{6} + \tfrac{3}{6}) \tag{15}$$
$$= \tfrac{1}{2}\tfrac{5}{6} \tag{16}$$
$$= \tfrac{5}{12} \tag{17}$$

Thus, the result is,

$$P(A=1|B=1) = \frac{P(B=1|A=1)P(A=1)}{P(B=1)} \tag{18}$$
$$P(A=1|B=1) = \frac{\frac{1}{2} \times \frac{1}{2}}{\frac{5}{12}} \tag{19}$$
$$P(A=1|B=1) = \tfrac{1}{4} \times \tfrac{12}{5} \tag{20}$$
$$P(A=1|B=1) = \tfrac{3}{5}. \tag{21}$$

END OF PAPER