**UNIVERSITY OF BRISTOL**

**May/June 2023 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:**
**2 Hours**

# 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:

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

2D Convolution:

$$g(x,y) = \sum_{m=-1}^{1} \sum_{n=-1}^{1} h(m,n) f(x-m, y-n)$$

Least Squares Matrix Form:

$$\mathbf{w}^* = (\mathbf{X}^T \mathbf{X})^{-1} \ \mathbf{X}^T \ \mathbf{y}$$

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**. Consider the pixel values of a small image below:

| | | | | | |
|---|---|---|---|---|---|
| 10 | 2 | 2 | 2 | 2 | 2 |
| 10 | 2 | 1 | 14 | 2 | 2 |
| 10 | 2 | 2 | 13 | 0 | 2 |
| 10 | 2 | 1 | 15 | 2 | 2 |
| 10 | 2 | 2 | 15 | 1 | 2 |
| 10 | 2 | 2 | 2 | 2 | 2 |

Using the position of the pixel with value 13 as the centre pixel to be convolved, apply the following convolution filter once using 8-connectivity and once using 4-connectivity:

$$\begin{pmatrix} -1 & 1 & -1 \\ -1 & 3 & -1 \\ -1 & 1 & -1 \end{pmatrix}$$

Which of the options below is the correct answer for the new pixel value in each connectivity case?

A. For 8-connectivity it is 76, and for 4-connectivity it is 70

**B. For 8-connectivity it is 60, and for 4-connectivity it is 66**

C. For 8-connectivity it is 60, and for 4-connectivity it is 76

D. For 8-connectivity it is 66, and for 4-connectivity it is 60

E. For 8-connectivity it is 70, and for 4-connectivity it is 76

*[6 marks]*

**Solution:** B - convolve or in fact correlate (since the filter is symmetric) with 8 neighbours or 4 neighbours.

**Q2**. Which of the following statements is TRUE:

A. The outer regions of the Fourier space represent the detail 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 sharpening.**

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

D. Options A, B, and C are all true.

(cont.)

E. Options A, B, and C are all false.

*[3 marks]*

**Solution:** B

**Q3**. Naive Bayes Classifier - The table below shows the probability of certain words from amongst a large selection of spam and not spam emails received at a university. The occurrence of the words and their probabilities are independent of each other.

| Word | $p(word\|spam)$ | $p(word\|\neg spam)$ |
|---|---|---|
| Ink | 0.80 | 0.30 |
| Term | 0.02 | 0.93 |
| Summer | 0.40 | 0.65 |
| Printer | 0.18 | 0.75 |
| Bulk | 0.70 | 0.10 |

Making a Naive Bayes assumption, compute the probability of sentence $S1$ below being spam and the probability of sentence $S2$ below not being spam:

$S1$- *Buy printer ink in bulk at prices seen last Summer.*
$S2$- *The Summer term notes are by the printer.*

Choose the correct option for $P(S1|spam)$ and $P(S2|not\ spam)$:

A. $P(S1|spam) = 0.0403$ and $P(S2|not\ spam) = 0.0146$

B. $P(S1|spam) = 0.0146$ and $P(S2|not\ spam) = 0.4534$

C. $P(S1|spam) = 0.0403$ and $P(S2|not\ spam) = 0.0014$

D. $P(S1|spam) = 0.0146$ and $P(S2|not\ spam) = 0.4095$

**E.** $P(S1|spam) = 0.0403$ **and** $P(S2|not\ spam) = 0.4534$

*[6 marks]*

**Solution:** E - $P(S1|spam) = 0.18\times0.80\times0.70\times0.40 = 0.0403$ and $P(S2|not\ spam) = 0.65 \times 0.93 \times 0.75 = 0.4095$

**Q4**. For a digitised sample acquired using 4Hz sampling and 4 quantisation levels, the following file has been provided:

0010101101010010011010111010 10

The first sample collected after the first second of recording has passed is equal to:

A. 0110

B. 1011

(cont.)

**C. 01**

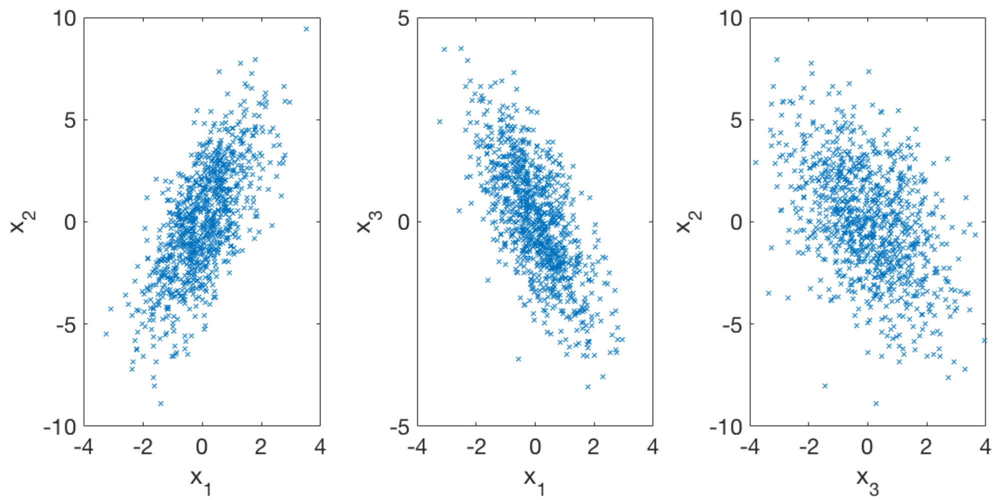
D. 10

E. 0010

*[3 marks]*

**Solution:** C - 4 quantisation levels requires 2 binary digits. 4Hz is 1/4 seconds, then at 1 second we have 00101011, and the first sample collected next is 01.

**Q5**. For three-dimensional data $X = (x_1, x_2, x_3)$, we plot each variable against the other as shown below:




Given these plots, determine which of the following is a reasonable estimate of the co-variance matrix $\Sigma$ of dataset $X$?

**A.** $\Sigma = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 7 & -2 \\ -1 & -2 & 2 \end{bmatrix}$

B. $\Sigma = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 1 & -2 \\ -1 & -2 & 7 \end{bmatrix}$

C. $\Sigma = \begin{bmatrix} 1 & -2 & 1 \\ -2 & 1 & 2 \\ 1 & 2 & 7 \end{bmatrix}$

D. $\Sigma = \begin{bmatrix} 1 & 0 & -1 \\ 0 & 7 & -2 \\ -1 & -2 & 2 \end{bmatrix}$

E. $\Sigma = \begin{bmatrix} 1 & 2 & -1 \\ 2 & 7 & -2 \\ -1 & -2 & 0 \end{bmatrix}$

*[5 marks]*

**Solution:** A

**Q6**. Which of the following 2D matrices are NOT separable? Ignore normalisation factors which are not stated here.

$$M_1 = \begin{pmatrix} -1 & 3 & -1 \\ -3 & 9 & -3 \\ 1 & -3 & 1 \end{pmatrix} \qquad M_2 = \begin{pmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ -1 & 0 & 1 \end{pmatrix} \qquad M_3 = \begin{pmatrix} -1 & 4 & -1 \\ -1 & 8 & -1 \\ -1 & 4 & -1 \end{pmatrix}$$

$$M_4 = \begin{pmatrix} 1 & 2 & -1 & 2 & 4 \\ 2 & 4 & -2 & 4 & 8 \\ -1 & -2 & 1 & -2 & -4 \\ 2 & 4 & -2 & 4 & 8 \\ 4 & 8 & -4 & 8 & 16 \end{pmatrix} \qquad M_5 = \begin{pmatrix} 1 & 0 & 0 & -1 \\ 2 & 1 & 1 & -2 \\ 1 & 0 & 0 & -1 \\ 2 & 1 & 1 & -2 \end{pmatrix}$$

Choose the correct option:

A. $M_1$ and $M_5$

B. $M_2$ and $M_3$ and $M_5$

C. $M_1$ and $M_4$

D. $M_2$ and $M_3$ and $M_4$

**E. $M_3$ and $M_5$**

*[5 marks]*

**Solution:** E - Only $M_3$ and $M_5$ are not separable - the others can be arrived at with one column and one row vector. Verify using Python or Matlab.

**Q7**. Two eigenvalues of the matrix below are 1 and 8:

$$\begin{bmatrix} 1 & -2 & 0 & 5 \\ 0 & 7 & 1 & 5 \\ 0 & 4 & 4 & 0 \\ 0 & 0 & 0 & 2 \end{bmatrix}$$

What are the other two eigenvalues?

A. 3 and 4

B. 4 and 2

**C. 2 and 3**

D. 1 and 3

E. 0 and 4

*[4 marks]*

(cont.)

**Solution:** C - Sum of the variances (main diagonal elements = 14) = sum of the eigenvalues, so answer has to be 2 and 3 for all the eigenvalues to sum up to 14 too

**Q8**. Fig. 1 shows an image of 'lines & numbers' and its Fourier Transform output.

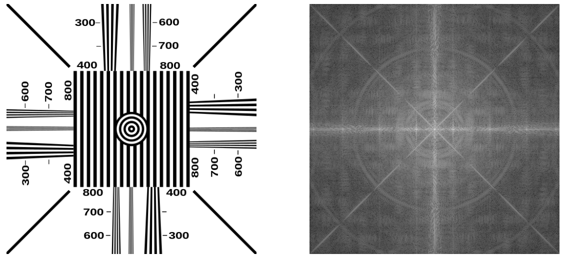


Figure 1: An image and its FFT space.

The top row in Fig. 2 shows 4 filtered versions of the 'lines & numbers' FFT space: ($F1, F2, F3, F4$). The 4 images in the bottom row, ($W, X, Y, Z$), show *in a random order*, the inverse FFT results of those filtered FFT outputs. Select the choice that correctly states which filtered FFT image corresponds to which inverse filtered FFT image.

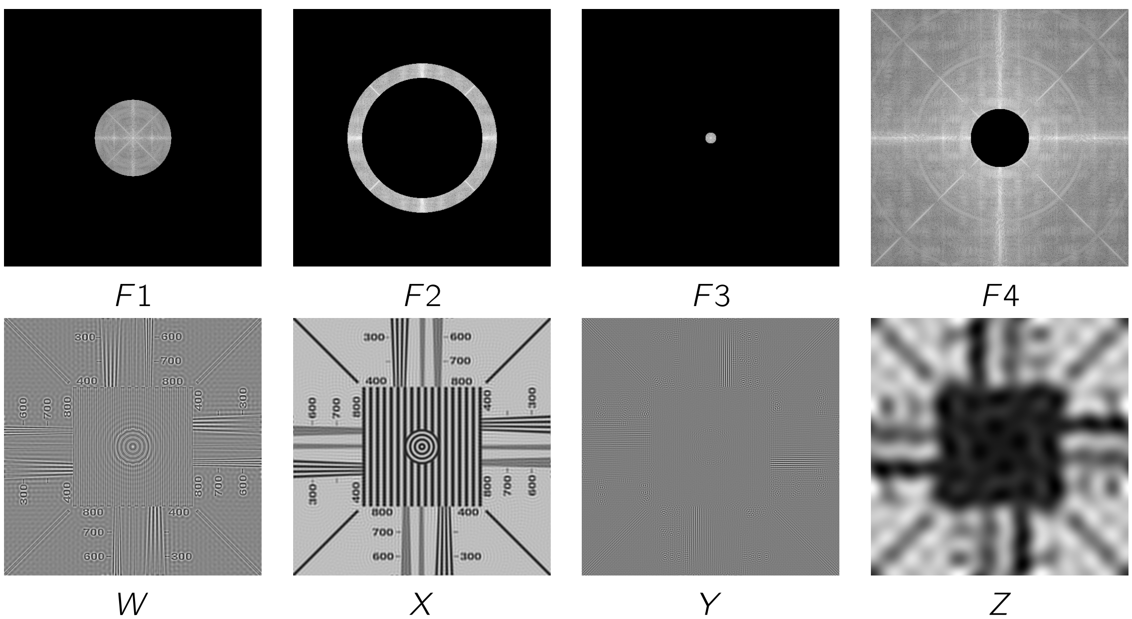



Figure 2: (top row) Filtered versions of the FFT result of the 'lines & numbers' image, and (bottom row) Inverse FFT results of tose in the top row but in a random order.

A. ($F1, F2, F3, F4$) correspond to ($Y, X, Z, W$)

B. ($F1, F2, F3, F4$) correspond to ($Y, X, W, Y$)

C. ($F1, F2, F3, F4$) correspond to ($Z, W, X, Y$)

**D. ($F1, F2, F3, F4$) correspond to ($X, Y, Z, W$)**

E. ($F1, F2, F3, F4$) correspond to ($X, Y, W, Z$)

*[10 marks]*

**Solution:** D - the filters are high pass ($F4 \Rightarrow W$), low pass ($F1 \Rightarrow X$), band pass ($F2 \Rightarrow Y$), and very low pass ($F2 \Rightarrow Z$).

**Q9**. The eigenvalues of a dataset are: [26.0, 16.0, 13.0, 5.0, 4.0, 3.0, 1.95, 0.85, 0.60]. Approximately what variance in the dataset do the first 4 eigenvalues represent?

A. 91.2%

B. 88.6%

C. 93.3%

D. 77.3%

**E. 85.2%**

*[5 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.

**Q10**. What is the minimum Edit Distance between the words "Sunday" and "Saturday"?

A. 4

B. 6

C. 7

**D. 3**

E. 2

*[3 marks]*

**Solution:** D - We would need to convert "un" to "atur" using 3 operations: substitute 'n' with 'r', insert 'a', insert 't'

**Q11**. Which of these is NOT a potential cause of overfitting?

A. Choosing a function class that is too complex

**B. Choosing a function class that is too simple**

C. No regularisation.

D. Too few datapoints.

E. Datapoints only cover a small region in the input space.

*[5 marks]*

**Q12**. Which of these statements about cross-validation is FALSE:

A. Cross-validation can be used to assess overfitting.

**B. Cross-validation reports performance on the training data used to fit the function.**

C. Cross-validation can be used to choose the function class.

D. Cross-validation can be used to choose the amount of regularisation.

E. Cross-validation can be computationally expensive if we have more than one or two hyperparameters.

*[5 marks]*

**Q13**. Which of these statements about the logarithm and its use in data-science is FALSE:

A. The logarithm converts products into sums, i.e. $\log ab = \log a + \log b$.

B. The logarithm converts powers into products, i.e. $\log a^b = b \log a$.

**C. The gradient of the logarithm is $\frac{\partial \log p}{\partial p} = 2p^{-1}$.**

D. Using log-probabilities rather than "raw" probabilities helps us avoid numerical under/overflow.

E. When doing maximum-likelihood fitting, the parameters with the highest log-likelihood are the same as the parameters with the highest "raw" likelihood.

*[5 marks]*

**Q14**. Find the value of:

$$\sum_{i=1}^{5} (\delta_{i2} i^3 + \delta_{i5} i^2)$$

where $\delta$ is the Kronecker-delta.

A. 30

**B. 33**

C. 34

(cont.)

D. 36

E. 40

*[5 marks]*

**Solution:**

$$\sum_{i=1}^{5}(\delta_{i2}i^3 + \delta_{i5}i^2) = 2^3 + 5^2 = 8 + 25 = 33. \tag{1}$$

**Q15**. For the data in the table, fit a model of the form $\hat{y} = w_1 + w_2 x$

| x | y |
|---|---|
| 0 | -4.2 |
| 1 | -2.3 |
| 2 | -0.1 |
| 3 | 2.1 |
| 4 | 3.9 |

*[5 marks]*

A. $w_1 = -4.20$, $w_2 = 2.12$.

**B.** $w_1 = -4.24$, $w_2 = 2.06$**.**

C. $w_1 = -4.30$, $w_2 = 2.12$.

D. $w_1 = -4.32$, $w_2 = 2.06$.

E. $w_1 = -4.35$, $w_2 = 2.12$.

**Q16**. For the data in the table, fit a model of the form $\hat{y} = w_1 x + w_2 x^2$

| x | y |
|---|---|
| 0 | -4.2 |
| 1 | -2.3 |
| 2 | -0.1 |
| 3 | 2.1 |
| 4 | 3.9 |

*[5 marks]*

A. $w_1 = -1.42$, $w_2 = 0.603$.

B. $w_1 = -1.55$, $w_2 = 0.654$.

**C.** $w_1 = -1.60$, $w_2 = 0.674$**.**

D. $w_1 = -1.78$, $w_2 = 0.687$.

E. $w_1 = -1.82$, $w_2 = 0.742$.

**Q17**. For the data in the table, fit a model of the form $\hat{y}_i = w_1 X_{i1} + w_2 X_{i2}$

| $X_{i1}$ | $X_{i2}$ | $y_i$ |
|---|---|---|
| -1 | -1 | -2.6 |
| -1 | 1 | -0.2 |
| 1 | -1 | 0.5 |
| 1 | 1 | 2.4 |

*[5 marks]*

A. $w_1 = 1.275$, $w_2 = 0.925$

B. $w_1 = 1.305$, $w_2 = 0.970$

C. $w_1 = 1.350$, $w_2 = 1.025$

D. $w_1 = 1.400$, $w_2 = 1.505$

**E.** $w_1 = 1.425$, $w_2 = 1.075$

**Q18**. Compute $\sum_i \log P(y_i|x_i)$ for binary classification, where

$$P(y_i = 1|x_i) = \sigma(1 + x_i - 2x_i^2)$$

with data,

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

*[5 marks]*

A. -9.32

B. -9.56

C. -9.61

**D. -9.69**

E. -9.83

**Q19**. We have $N$ datapoints, $x_1, \ldots, x_N$, distributed according to,

$$P(x_i|\mu) \propto \frac{1}{x_i} e^{-(\log x_i - \mu)/2}$$

What is the maximum-likelihood solution for $\mu$?

A. $\mu = \frac{1}{N} \sum_{i=1}^N x_i$.

B. $\mu = \frac{1}{2N} \sum_{i=1}^N \log x_i$

**C.** $\mu = \frac{1}{N} \sum_{i=1}^N \log x_i$

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

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

*[5 marks]*

**Solution:**

$$\log P(x|\mu) = \sum_i (-\log x_i - (\log x_i - \mu)^2/2) \tag{2}$$

$$\log P(x|\mu) = \sum_i (-\log x_i - (\log x_i)^2/2 + \mu \log x_i - \mu^2/2) \tag{3}$$

$$0 = \frac{\partial}{\partial \mu} \log P(x|\mu) = \sum_i (\log x_i - \mu) \tag{4}$$

$$0 = \sum_i (\log x_i) - N\mu \tag{5}$$

$$N\mu = \sum_i (\log x_i) \tag{6}$$

$$\mu = \tfrac{1}{N} \sum_i (\log x_i) \tag{7}$$

**Q20**. We have $N$ datapoints, $x_1, \ldots, x_N$, distributed according to,

$$P(x_i|\beta) \propto \beta^2 \tfrac{1}{x_i^3} e^{-\beta/x_i}$$

What is the maximum-likelihood solution for $\beta$?

A. $\frac{1}{2}(\frac{1}{N}\sum_{i=1}^N x_i)$

B. $\frac{1/2}{\frac{1}{N}\sum_{i=1}^N x_i}$

C. $\frac{1/2}{\frac{1}{N}\sum_{i=1}^N (1/x_i)}$

D. $\frac{2}{\frac{1}{N}\sum_{i=1}^N x_i}$

**E.** $\frac{2}{\frac{1}{N}\sum_{i=1}^N (1/x_i)}$

*[5 marks]*

**Solution:**

$$\log P(x|\beta) = \sum_{i=1}^{N}(2\log\beta - 3\log x - \beta/x_i) \tag{8}$$

$$= N(2\log\beta - 3\log x) - \beta\sum_{i=1}^{N}(1/x_i) \tag{9}$$

$$0 = \frac{\partial}{\partial\beta}\log P(x|\beta) = 2N\tfrac{1}{\beta} - \sum_{i=1}^{N}(1/x_i) \tag{10}$$

$$\sum_{i=1}^{N}(1/x_i) = 2N\tfrac{1}{\beta} \tag{11}$$

$$\beta = \frac{2N}{\sum_{i=1}^{N}(1/x_i)} \tag{12}$$

$$\beta = \frac{2}{\frac{1}{N}\sum_{i=1}^{N}(1/x_i)} \tag{13}$$



**END OF PAPER**

Left intentionally blank for workings out.

Left intentionally blank for workings out.

Left intentionally blank for workings out.

Left intentionally blank for workings out.

Left intentionally blank for workings out.

Left intentionally blank for workings out.