University of Glasgow

Tuesday 12 May 2020
09:00 BST
(24 hour open online assessment – Indicative duration: 1 hour 30 minutes)

DEGREES of MSci, MEng, BEng, BSc, MA and MA (Social Sciences)

DATA FUNDAMENTALS (H)

COMPSCI 4073

Answer any two of three questions

This examination paper is worth a total of 50 marks.

24 hour open online assessment.
Adhere to the stated maximum word counts.

1. You are tasked with working with a radio telescope, which scans the sky for radio transmissions. The telescope has a fixed angle to the ground (fixed elevation), scanning as it rotates continuously around its vertical axis.

- Every night this telescope runs for 12 hours, making 500 complete revolutions each hour.
- For analysis purposes, the sky is divided into 112 angular zones, denoted ϕ, numbered $0 \ldots 111$. Zone 111 is therefore adjacent to zone 0 and zone 110.
- The data for one night is stored in a sequentially ordered `(6000, 112)` NumPy array `scan_data`.
- The function `llik(x)` is defined to take an `(n, 112)` array of data `x` and returns the log-likelihood $\log(P(\phi|\text{signal})$ (also as an `(n, 112)` array) of observing a signal in each zone under some model.
- A prior function `prior(phi)` is defined, which returns the prior probability $P(\phi)$ of a observing signal in a given angular zone `phi`. It can accept array-valued `phi`.

(a) Write a *vectorised* NumPy function that computes the zone with the maximum posterior probability of a signal given an angular zone, $P(\text{signal}|\phi)$, *for each time step*, producing a shape `(6000,)` element array. **Comment your solution in detail, explaining each step of the process.**

Maximum 40 lines of code, including comments

[8]

(b) An approximation could be fitted using **maximum likelihood estimation (MLE)**, instead of the Bayesian procedure in part (a). Why is MLE *not* a Bayesian approach to inference? Discuss what would be required to utilise a Bayesian approach, and what advantages it disadvantages this would bring.

Maximum 300 words

[6]

(c) The probability of observing a radio signal in the sky varies radically across different areas of the sky in an unpredictable spiky pattern; some zones have $P(\text{signal}|\phi) = 0.000001$; others have $P(\text{signal}|\phi) = 0.9999$. You are asked to come up with a numerical display that shows the probability of a signal being detected in a particular zone to a human operator. Argue a case for using one of these representations for *display*:

- probabilities
- log probabilities
- odds
- logits

Maximum 200 words

[4]

(d) A visualisation like the one below is used to show the probability of signals as a function of angle.

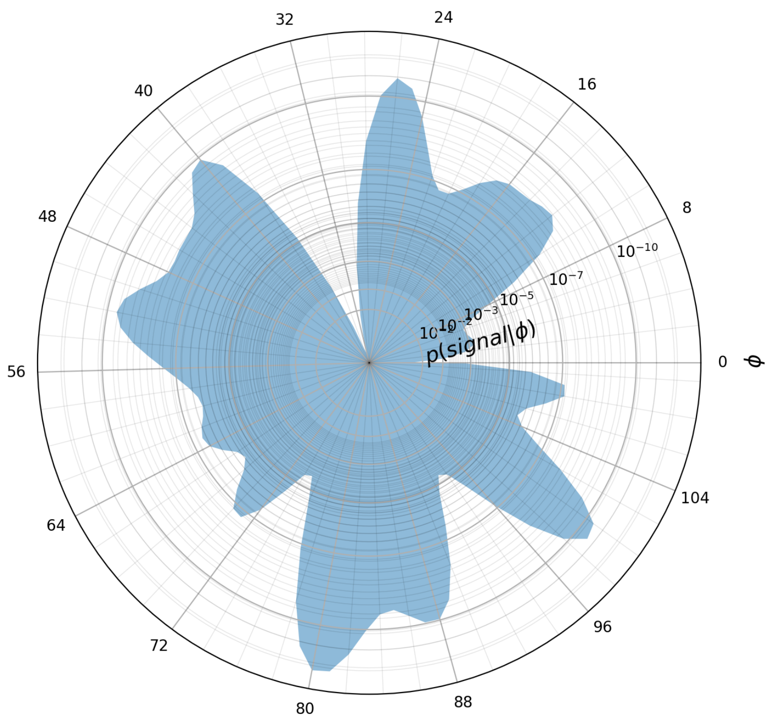

Figure 1: Signal probability plot

(i) What type of **coords** does this plot use and what benefit might they have in this application? For each of the four probability representations above (probabilities, log-probabilities, logits and odds) discuss the issues in using these representations in this plot.
Maximum 200 words

[7]

2. You are tasked with implementing the calibration of a projector projecting a digital image onto a flat screen. This is a problem of mapping a 2D point $\mathbf{x} \in \mathbb{R}^2$ in screen pixel coordinates, onto a 2D point $\mathbf{y} \in \mathbb{R}^2$ in physical world coordinates. Assume that this transformation is **linear**, and defined by the equation:

$$\mathbf{y} = A\mathbf{x}$$

A is not known.

(a) You are asked to analyse the transformation that the projector applies.

A camera is pointed at the projector. This transforms from physical coordinates $\mathbf{y}$ to a third vector space $\mathbf{z} \in \mathbb{R}^2$ (camera sensor coordinates) via a transformation matrix B, $\mathbf{z} = B\mathbf{y}$.

- You are able to display a bright pixel at any position $\mathbf{x}$.
- You are able to determine the location of a bright pixel on the camera sensor $\mathbf{z}$.

From this, you are asked to devise a procedure to estimate *one* eigenvector x_1 and corresponding eigenvalue λ_1 of the *unknown* transformation matrix *A from pixel space* $\mathbf{x}$ *to physical space* $\mathbf{y}$. Write down detailed steps of this estimation procedure.

- Assume B is known and is non-singular.
- Ignore any constraints as to the resolution, projectable area or camera field of view.
- Assume the camera and projector are aligned such that [0,0] in pixel $\mathbf{x}$ = [0,0] in physical $\mathbf{y}$ space = [0,0] in $\mathbf{z}$ space.

Maximum 600 words

[15]

(b) Instead of a linear model, the projector design team decide to fit a **nonlinear** function to approximate the transform $\mathbf{z} = f(\mathbf{x}; \theta)$, where θ is a 200 dimensional parameter vector.

They intend to adjust θ by comparing a large set of test calibration image vectors $\mathbf{x_i}$ with simulated camera image vectors $\mathbf{z'_i}$ against the true camera image vectors $\mathbf{z_i}$.

The optimisation will therefore take the form of:

$$\theta^* = \text{argmin}_\theta \sum_i \|f(\mathbf{x_i}; \theta) - \mathbf{y_i}\|$$

(i) Discuss the advantages and disadvantages of a **nonlinear model** versus the **linear model** above for this problem.
Maximum 200 words

[4]

(ii) The team intend optimise for θ numerically. It is suggested that **stochastic gradient descent** might be applied. Discuss whether this would be suitable in this context, and argue for or against its use.
Maximum 300 words

[6]

3. You are employed by a double glazing company to analyse the effects of acoustic isolating windows.

(a) Your are tasked with measuring the effectiveness of acoustic glass in decreasing urban noise, which is concentrated between 30-300Hz. Your boss is deciding between microphones to measure noise levels.

- Microphone A: sampling rate 600Hz; £35/unit
- Microphone B: sampling rate 1400Hz; £138/unit
- Microphone C: sampling rate 80000Hz; £160/unit
- Microphone D: sampling rate 100Hz; £15/unit

Which microphone should be chosen? Justify your answer, and explain what would happen if an inappropriate sample rate were chosen.

Maximum 200 words

[6]

(b) You are asked to *empirically* measure the (sampled) convolution kernel $f[t]$ of a real glass panel using the microphone sensing unit so that the effect can be simulated to demonstrate to customers. Explain how to do this, and the relevance of the Dirac delta function to this process. Discuss any potential issues with practically applying this idea to replicate the acoustic response of the glass.

Maximum 300 words

[7]

(c) You are given the table below, which has information on the frequency response of three different window types. For each type of glass, the relative dampening at that frequencies is shown. At each frequency, the measurement error for that frequency is also given.

Freq.	Error	Glass types		
		Acoustic-A	R-5	Standard
10Hz	± 20dB	-30dB	-10dB	-5dB
50Hz	± 10dB	-50dB	-8dB	-15dB
90Hz	± 2dB	-60dB	-9dB	-25dB
130Hz	± 2dB	-65dB	-11dB	-26dB
200Hz	± 2dB	-40dB	-18dB	-26dB
300Hz	± 5dB	-20dB	-25dB	-27dB

(i) The graph below visualises these results. Criticise this graph in detail, *using the correct terminology from the Layered Grammar of Graphics* in your discussion. For each criticism you make, suggest how you would improve the visualisation.
Maximum 600 words

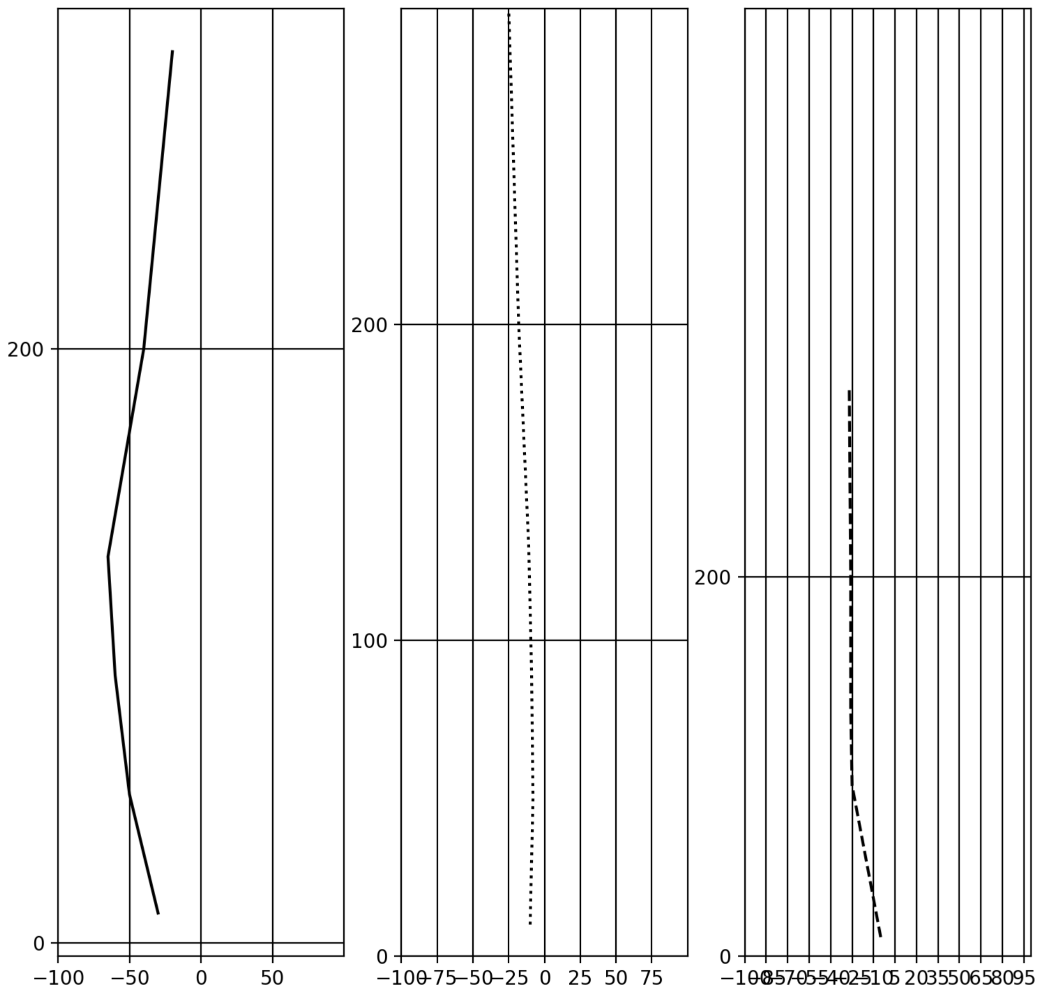

Figure 2: A summary of the acoustic response of three acoustic isolation panels. Each panel exhibits a different frequency response.

[12]

END OF QUESTION PAPER