Assessed Exercise #1 CS4061 / CS5014 ML

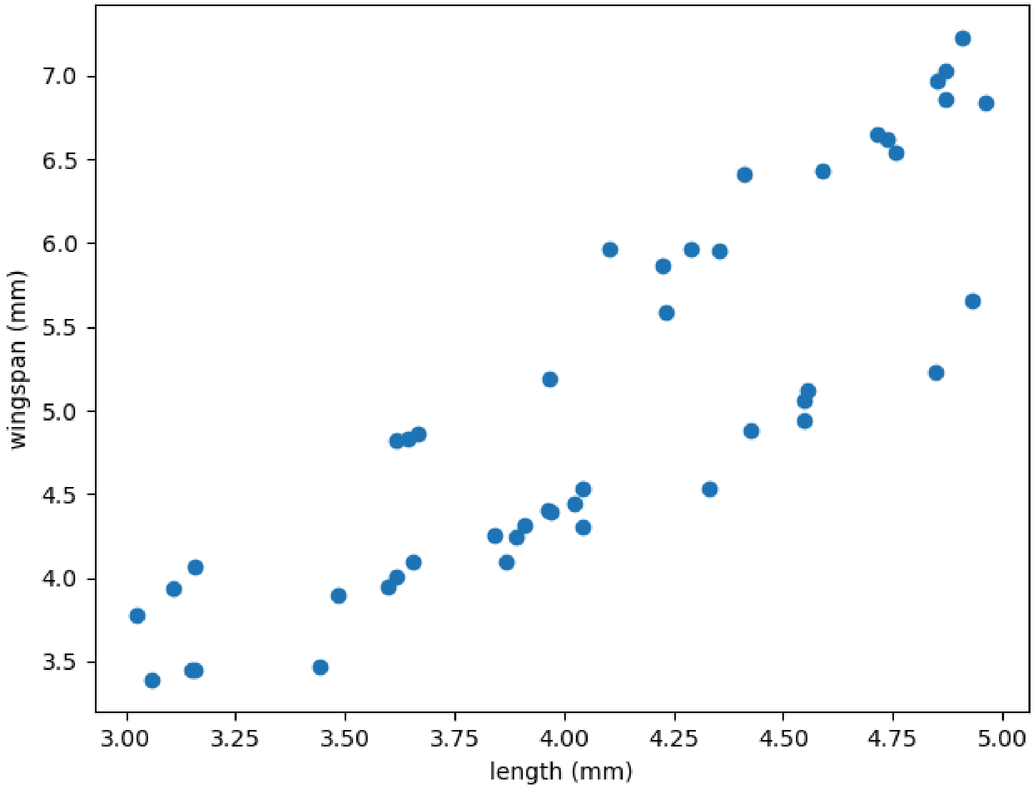
Instructions

- This is a solo coursework; it is expected to take you no more than five hours
- See the Moodle page for submission information and deadline
- You should answer all parts of both questions, and submit your answers together in the Moodle response text-box
- For each part of each question, you should write one short paragraph in response
- The two questions carry equal marks (a maximum of 25 marks each – one mark per correct/insightful and *relevant* comment, regardless of how these are split over parts)
- The overall length (total for the two questions combined) is limited to **1250 words**. This is plenty to gain full marks provided you write crisp, precise text
- Justify and explain your answers. If you think an answer is subjective, dependent on certain assumptions, etc., then explain this
- **You do not need to implement / code anything for this exercise!**
- The questions are intended to make you think, and to extend your understanding. They use ideas from the lectures, but you will not be able to look everything up in the slides
- Feel free to discuss questions with your colleagues
- Feel free to research whatever you need online
- **However: make sure the answers you submit are your own ideas in your own words!** I'll use Turnitin to check for plagiarism, and I know which parts ChatGPT gets right/wrong
- You may include images/diagrams/formulae in your responses if you want, but this is not required

Question 1

Insects of a certain species hatch either in autumn or in spring. Individuals exhibit slightly different developmental characteristics (rate of growth, etc.) depending which season they hatched in. However, there is no distinctive identifying feature that shows unambiguously when a given individual hatched.

Suppose we want to predict an insect's wingspan from its length, using a linear regression model. We collect random insects from the environment, and observe measurements as shown in the plot below.



The differing developmental characteristics result in two different noisy trends being present in the data (the steeper and shallower bands visible above).

continues overleaf…

Answer the following questions…

a. What would the result of fitting a simple linear model maximising a Gaussian log-likelihood on this data look like? Why? Would the resulting model be useful for predicting wingspan from length, (i) for lengths between 3–5mm, (ii) for lengths <3mm, and (iii) for lengths >5mm?

b. Would using non-linear features (e.g. RBFs) calculated from the length improve the quality of predictions? Why?

c. Would a deep neural network inputting length and trained to minimise the MSE between true and predicted wingspans perform better than the above linear models?

d. Consider a fully Bayesian linear regression with Gaussian likelihood and a Gaussian prior. Would the posterior be multi-modal? Why? Would the resulting model be more/less useful than the maximum-likelihood fit?

e. What alternative choice(s) of density for the likelihood might work well (in the fully Bayesian setting)? Why? Would different choices of likelihood have any impact on the difficulty of finding the posterior? How might you fit a linear regression model with these likelihoods?

Question 2

In the lectures, we used a linear model to predict the winning time of the men's Olympic sprint, given the year as input. However, measurements of the winning time are likely to be imprecise (i.e. noisy/uncertain). In practice it seems likely that the winning time was measured more accurately in more recent games (high-frame-rate videos of the finish line are more accurate than a human with a mechanical stopwatch). Perhaps the uncertainty (noise) has lessened linearly over time; perhaps there is some more complex relationship (e.g. if new technologies were introduced in particular years causing a sharp decrease in noise).

Answer the following questions…

a. Would a linear model minimising the mean square error still be appropriate when the noise is of varying amplitude (assuming such a model is reasonable in the case of fixed noise variance)?

b. How would the output (predictions) from a linear model be affected by the fact that measurements were less accurate historically? Does this effect depend on the dataset size (e.g. what if there were Olympic Games every year instead of once every four years)?

c. How could a maximum-likelihood fit be adapted to explicitly handle the varying noise amplitude in this situation? What optimization method could be used to fit it to the data? Would there be any benefit in doing so (versus assuming fixed noise amplitude)?

d. Would this model be more or less prone to overfitting than one assuming fixed noise amplitude?

e. Would a Bayesian model handle this situation better?