THE UNIVERSITY OF MANCHESTER

One Hour

COURSE TITLE: STATISTICS AND MACHINE LEARNING I

Date: ???
Time: ???

Answer all questions
In total you may achieve **FORTY** (40) marks.

Please enter your answers in the Answer Book provided.

Electronic calculators may be used, provided that they cannot store text

---

P.T.O

Answer **ALL** questions.
For maximal points answers have to be correct and well-motivated

1.

(a) Explain the idea behind the Ordinary Least Squares method of estimation and illustrate your answer with a stylised figure. (5 marks)

A student of Sociology has estimated the parameters of a logistic regression model that explains receiving a promotion at work in the last year (1 if a promotion was obtained, 0 if not), with the sex of the individual (variable *fem,* taking values 0 for males and 1 for females), their level of education (three levels: 1 for lower education *l_ed*, 2 for middle education *m_ed* and 3 for higher education *h_ed*) and their interactions as the explanatory variables. "Lower education" was used as a reference category.

(b) Interpret the estimates of the model parameters in the table below, showing your working in detail. (6 marks)

Parameter θ	*intercept*	*fem*	*m_ed*	*h_ed*	*fem*m_ed*	*fem*h_ed*
e^{θ}	1	0.6	1.2	1.5	0.95	0.9

(c) Explain the results you would expect if you specified a model without the main effects for level of education. (4 marks)

2. The first two parts of this question concern a kernel density estimate (KDE)

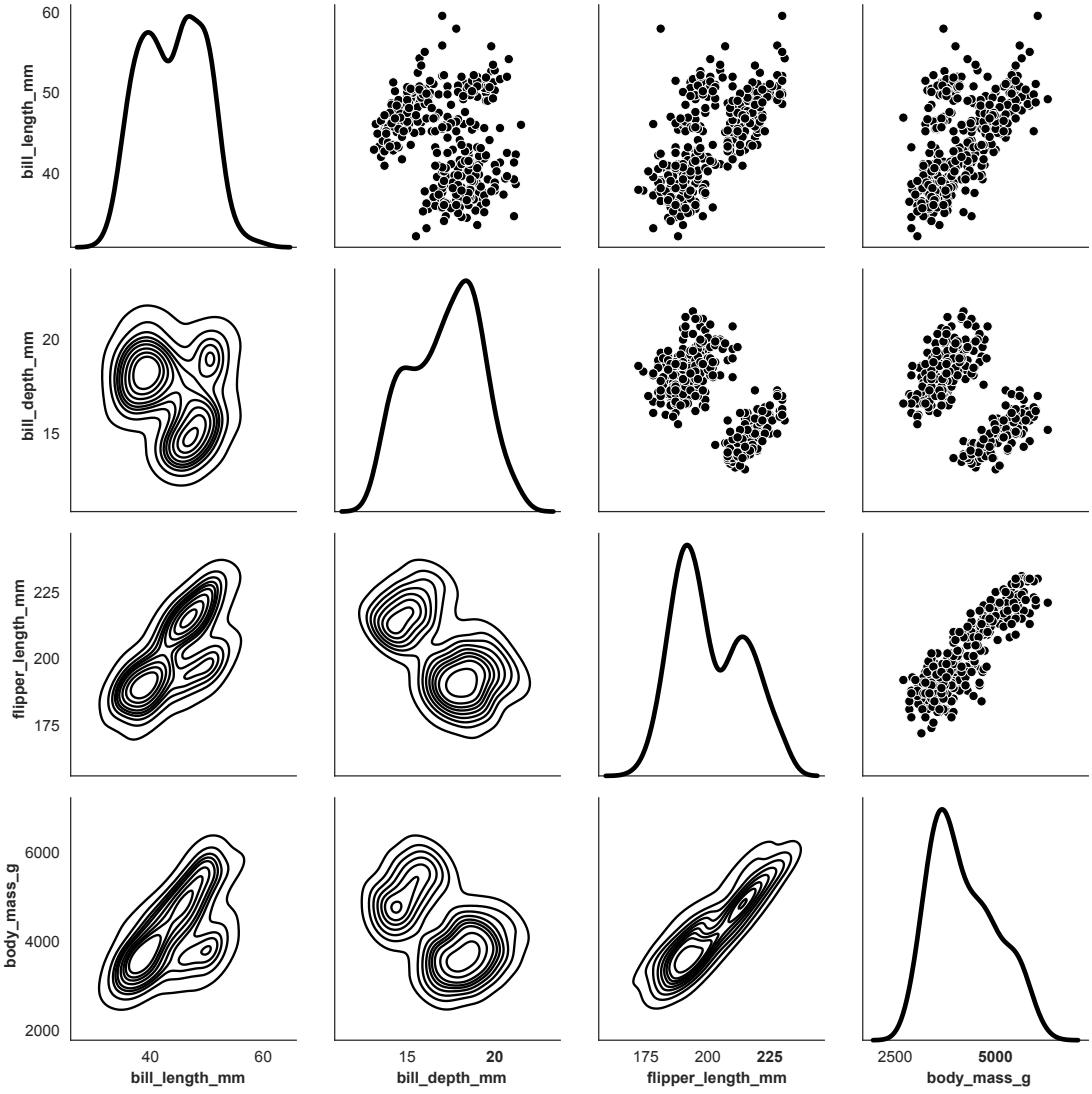
$$f(s) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{w}K\left(\frac{s - x_j}{w}\right)$$

constructed from a data set $\{x_1, \ldots, x_n\}$, where the x_j are ordinary real numbers.

(a) Explain how a KDE is used. Your answer should include a discussion of all the elements in the formula above. (5 marks)

(b) In the case where the data are $\{0, 0.5, 1\}$, the kernel is rectangular and $w = 1$, sketch $f(s)$. (5 marks)

The figure below is part of an exploratory data analysis on a data set consisting of measurements on several species of penguins.



(c) What insights can you draw from this figure? How many species do you think are involved, and why? (5 marks)

3.

(a) You are asked to use a training data set, $\mathcal{Z} = \{(x_i, y_i)\}_{i=1}^{n}$ to learn an approximation of the unknown non-linear mapping with polynomial regression models. Describe two different methods that can find out an optimal polynomial model in terms of generalisation. In your answer, you need to give the main steps in each method chosen by you. (4 marks)

(b) According to the *no free lunch* theorem in machine learning, there does not exist one single learning algorithm that always outperforms others in all circumstances. For linear regression, you have learned three learning algorithms in this course unit: *ordinary least squares* (OLS), *ridge regression* and *LASSO*. State that in what circumstance,

1) OLS is likely to outperform ridge regression and LASSO? (2 marks)
2) Ridge regression is likely to outperform OLS and LASSO? (2 marks)
3) LASSO is likely to outperform OLS and ridge regression? (2 marks)

It is essential to justify your answers to questions 1) – 3).

END OF EXAMINATION