Examiners' commentary 2021

ST3189 Machine learning

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2020–21. The format and structure of the examination may change in future years, and any such changes will be publicised on the virtual learning environment (VLE).

Information about the subject guide and the Essential reading references

Unless otherwise stated, all cross-references will be to the latest version of the course (2019). You should always attempt to use the most recent edition of any Essential reading textbook, even if the commentary and/or online reading list and/or subject guide refer to an earlier edition. If different editions of Essential reading are listed, please check the VLE for reading supplements – if none are available, please use the contents list and index of the new edition to find the relevant section.

This included all the material in the VLE as well as references therein such as the main textbook:

- James G., Witten D., Hastie T. and R. Tibshirani, *An introduction to Statistical Learning: with Applications in R*, Springer (2013), ISBN 9781461471387

that will be referred to as James *et al.* throughout.

General remarks

Learning outcomes

At the end of the course and having completed the essential reading and activities you should be able to:

- develop an understanding of the process to learn from data
- be familiar with a wide variety of algorithmic and model-based methods to extract information from data
- apply and evaluate suitable methods to various datasets by model selection and predictive performance evaluation.

Planning your time in the examination

You have two hours to complete this paper, which consists of four compulsory questions. Remember that each of these questions is likely to cover more than one topic. This means that it is really important that you make sure you have a reasonable idea of what topics are covered before you start work on the paper! We suggest you divide your time as follows during the examination:

- Spend the first 10 minutes annotating the paper. Note the topics covered in each question and subquestion.
- Allow yourself 25 minutes for each question. Do not allow yourself to get stuck on any one question, but do not just give up after two minutes!
- This leaves you with 10 minutes. Do not leave the examination hall at this point! Check over any questions you may not have completely finished. Make sure you have labelled and given a title to any tables or diagrams which were required.

What are the examiners looking for?

The examiners are looking for very simple demonstrations from you. They want to be sure that you:

- have covered the syllabus as described and explained in the course material
- know the basic concepts given and, more importantly, when and how to use them
- understand and answer the questions set.

You are *not expected to write long essays* with lengthy explanations. However, clear and accurate language, both mathematical and written, is expected and marked. The explanations below and in the specific commentary for the examination paper should make these requirements clear.

Key steps to improvement

The most important thing you can do is answer the question set! This may sound very simple, but these are some of the things that candidates often do not do, though asked! Remember:

- If you are asked to label a diagram (which is almost always the case!), please do so. What do the data describe? What are the units? What are the x and y axes?
- Do not waste time calculating things which are not required by the examiners.
- When making calculations try to use as many decimal places as possible to reach the most accurate solution. It is advised to have at least two decimal places in general and at least three decimal places when calculating probabilities.

How should you use the specific comments on each question given in the *Examiners′ commentaries*?

We hope that you find these useful. For each question and subquestion, they give:

- further guidance for each question on the points made in the last section
- the answers, or keys to the answers, which the examiners were looking for
- where appropriate, suggested activities from the course material which should help you to prepare, as well as similar questions.

Any further references you might need are given in the part of the course to which you are referred for each answer.

Memorising from the *Examiners′ commentaries*

It is generally noted in similar examination papers that a small number of candidates appear to be memorising answers from previous years' *Examiners' commentaries* – for example, plots – and, therefore, produce the exact same image of them without looking at the examination questions at all! Note that this is very easy to spot. The *Examiners' commentaries* should be used as a guide to practise on sample examination questions and it is pointless to attempt to memorise them.

Examination revision strategy

Many candidates are disappointed to find that their examination performance is poorer than they expected. This may be due to a number of reasons, but one particular failing is '**question spotting**', that is, confining your examination preparation to a few questions and/or topics which have come up in past papers for the course. This can have serious consequences.

We recognise that candidates might not cover all topics in the syllabus in the same depth, but you need to be aware that examiners are free to set questions on **any aspect** of the syllabus. This means that you need to study enough of the syllabus to enable you to answer the required number of examination questions.

The syllabus can be found in the Course information sheet available on the VLE. You should read the syllabus carefully and ensure that you cover sufficient material in preparation for the examination. Examiners will vary the topics and questions from year to year and may well set questions that have not appeared in past papers. Examination papers may legitimately include questions on any topic in the syllabus. So, although past papers can be helpful during your revision, you cannot assume that topics or specific questions that have come up in past examinations will occur again.

If you rely on a question-spotting strategy, it is likely you will find yourself in difficulties when you sit the examination. We strongly advise you not to adopt this strategy.

Examiners' commentary 2021

ST3189 Machine learning

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2020–21. The format and structure of the examination may change in future years, and any such changes will be publicised on the virtual learning environment (VLE).

Information about the subject guide and the Essential reading references

Unless otherwise stated, all cross-references will be to the latest version of the course (2019). You should always attempt to use the most recent edition of any Essential reading textbook, even if the commentary and/or online reading list and/or subject guide refer to an earlier edition. If different editions of Essential reading are listed, please check the VLE for reading supplements – if none are available, please use the contents list and index of the new edition to find the relevant section.

Comments on specific questions – Zone A

Candidates should answer all **FOUR** questions. All questions carry equal marks.

Answer **all** parts of the following questions.

Question 1

(a) **Indicate whether the following statements are true or false. Briefly justify your answers.**

i. **The maximum likelihood estimates for α_1, α_2 in the model $y_i = x_i^{\alpha_1} e^{\alpha_2} + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$ are i.i.d. noise, can be obtained using linear regression.**

(3 marks)

ii. **In random forests, for each tree a random selection of internal nodes are discarded in order to decorrelate the trees.**

(3 marks)

iii. **The partitioning of the input space shown in Figure 1 could be generated by recursive binary splitting.**

(3 marks)

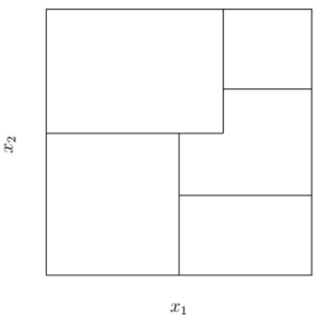


Figure 1: For Question 1. (a) iii.

Reading for this question

This question refers to several techniques and machine learning concepts of the course requiring some basic understanding of them. More specifically, part i. is on linear regression that can be found in several sections of the James *et al.* textbook but mainly in Chapter 3. For parts ii. and iii. the content of tree-based methods (Chapter 8 of James *et al.*) such as the procedure to obtain a tree (part iii.) and random forests (part ii.).

Approaching the question

Remember that the justification has to be one sentence so organise your thoughts accordingly and avoid lengthy answers. Some 'good answers' are provided below. Note that there can be more than one 'correct' answer in some of these questions.

i. False. y_i is not linear in α_1, α_2, and no simple transformation will make it linear ($\log(x_i^{\alpha_1\alpha_2} + \varepsilon_i) \neq \alpha_1 \log x_i + \alpha_i + \varepsilon_i$).

ii. False. In order to decorrelate the trees, a random selection of input variables at each split (not internal nodes) are discarded.

iii. False. Recursive binary splitting ensures axis-alighed splits, so the top-left split is impossible.

Overall, candidates did okay on parts i. and iii. but not so good on part ii.

(b) **Consider a learning problem with two features. How are the decision tree and 1-nearest neighbour decision boundaries related? Specifically, discuss the similarities and dissimilarities.**

(4 marks)

Reading for this question

This question examines the main ideas of tree-based methods and K-nearest neighbours, that can be found in Chapter 8 and Section 2.2 of James *et al.*, respectively.

Approaching the question

In both cases, the decision boundary is piecewise linear. Decision trees do axis-aligned splits while 1-nearest neighbour gives a voronoi diagram.

(c) i. **Provide a 2-dimensional dataset where 1-nearest neighbour has lower Leave-One-Out Cross Validation (LOOCV) error than linear support vector machines.**

(3 marks)

ii. Provide a 2-dimensional dataset where 1-nearest neighbour has higher LOOCV error than linear support vector machines.

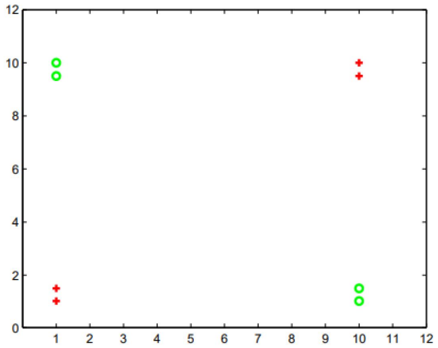
(3 marks)

Reading for this question

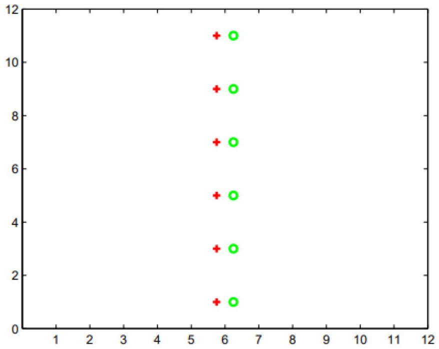
This question contains material regarding the $K-$nearest neighbours technique that can be found in Section 2.2 of the James *et al.* textbook. The content of support vector machines is also relevant and can be found in Chapter 9.

Approaching the question

i. For example:



ii. For example:



(d) **Consider the K-nearest neighbours approach using Euclidean distance on the dataset shown in Figure 2. What are the LOOCV errors for the following cases? Briefly justify your answers.**

i. **1-nearest neighbour.**

(3 marks)

ii. **3-nearest neighbours.**

(3 marks)

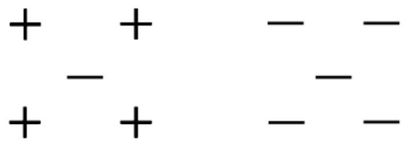


Figure 2: For Question 1. (d)

Reading for this question

As before the content on K-nearest neighbours and support vector machines is relevant but this time knowledge on the leave-one-out cross-validations is required that can be found in Chapter 5 of the James *et al.* textbook.

Approaching the question

i. The left-hand points are misclassified and the error is 5/10.

ii. The left '–' point is misclassified and the error is 1/10.

Question 2

Consider a linear regression setting where the response variable is $y = (y_1, \ldots, y_n)$ and there is one feature, or else predictor, $x = (x_1, \ldots, x_n)$. We are interested in fitting the following model

$$y_i = \beta \exp x_i + \epsilon_i, \quad i = 1, \ldots, n,$$

where the error terms ϵ_i's are independent and distributed according to the normal distribution with mean 0 and known variance σ^2. Equivalently, we can write that given x each y_i is independent and distributed according to the normal distribution with mean $\beta \exp x_i$ and known variance σ^2.

(a) Derive the likelihood function for the unknown parameter β.

(2 marks)

(b) Derive the Jeffreys prior for β. Use it to obtain the corresponding posterior distribution.

(5 marks)

(c) Consider the normal distribution prior for β with zero mean and variance ω^2. Use it to obtain the corresponding posterior distribution.

(5 marks)

(d) Consider the least squares criterion

$$\sum_{i=1}^{n} (y_i - \beta \exp x_i)^2, \tag{1}$$

and show that the estimator of β that minimises equation (1), also maximises the likelihood function derived in part (a). Derive this estimator and, in addition, consider the following penalised least squares criterion

$$\left\{ \sum_{i=1}^{n} (y_i - \beta \exp x_i)^2 \right\} + \lambda \beta^2, \tag{2}$$

given a $\lambda > 0$. Derive the estimator of β that minimises equation (2) and compare with the one that minimises equation (1).

(4 marks)

(e) Provide a Bayes estimator for each of the posteriors in parts (b) and (c) and compare them with the estimators of part (d).

(5 marks)

(f) Let y_{n+1} represent a future observation from the same model given the corresponding value of the predictor x_{n+1}. Find the posterior predictive distribution of y_{n+1} for one of the posteriors in parts (b) or (c).

(4 marks)

Reading for this question

This question is examining Bayesian inference which can be found in Block 4 of the VLE section of the course. Read the parts on 'Bayesian Inference Essentials' and 'Bayesian Inference Examples'. Exercises 1–6 as well as Exercise 2 of the mock examination are relevant for practice (try to do a few of them). Also the part on least squares and the part on shrinkage methods, in Chapter 3 and Section 6.2 of James *et al.*, respectively, are relevant for part (d).

Approaching the question

(a) The likelihood can be written as:

$$\begin{aligned} L(\beta \,|\, x) = f(x \,|\, \beta) &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(y_i - \beta\exp(x_i))^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{\sum\limits_{i=1}^{n}(y_i - \beta\exp(x_i))^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{\beta^2\sum\limits_{i=1}^{n}\exp(2x_i) - 2\beta\sum\limits_{i=1}^{n} y_i\exp(x_i)}{2\sigma^2}\right). \end{aligned}$$

(b) The log-likelihood can be written as:

$$\ell(\beta \,|\, x) = -\frac{\beta^2\sum\limits_{i=1}^{n}\exp(2x_i) - 2\beta\sum\limits_{i=1}^{n} y_i\exp(x_i)}{2\sigma^2}.$$

In order to find Jeffreys prior we calculate:

$$\frac{\partial\ell(\beta \,|\, x)}{\partial\theta} = -\frac{\beta\sum\limits_{i=1}^{n}\exp(2x_i) - \sum\limits_{i=1}^{n} y_i\exp(x_i)}{\sigma^2}$$

also:

$$\frac{\partial^2\ell(\beta \,|\, x)}{\partial\beta^2} = -\frac{\sum\limits_{i=1}^{n}\exp(2x_i)}{\sigma^2}$$

and the Fisher informatio:

$$\mathcal{I}(\beta) = -\mathrm{E}\left(-\frac{\sum\limits_{i=1}^{n}\exp(2x_i)}{\sigma^2}\right) = \frac{\sum\limits_{i=1}^{n}\exp(2x_i)}{\sigma^2}.$$

The Jeffreys prior is:

$$\pi^J(\beta) \propto \mathcal{I}(\beta)^{1/2} \propto 1.$$

Using the Jeffreys prior, the corresponding posterior $\pi^J(\beta \,|\, x)$ is proportional to:

$$\pi^J(\beta \,|\, x) \propto \exp\left(-\frac{\beta^2 \sum\limits_{i=1}^n \exp(2x_i) - 2\beta \sum\limits_{i=1}^n y_i \exp(x_i)}{2\sigma^2}\right)$$

$$= \exp\left(-\frac{\beta^2 - 2\beta \frac{\sum_{i=1}^n y_i \exp(x_i)}{\sum_{i=1}^n \exp(2x_i)}}{2\frac{\sigma^2}{\sum_{i=1}^n \exp(2x_i)}}\right)$$

$$\stackrel{\mathcal{D}}{=} N\left(\frac{\sum\limits_{i=1}^n y_i \exp(x_i)}{\sum\limits_{i=1}^n \exp(2x_i)}, \frac{\sigma^2}{\sum\limits_{i=1}^n \exp(2x_i)}\right).$$

(c) The prior of β can be written as:

$$\pi(\beta) = (2\pi\sigma^2\omega^2)^{-1/2} \exp\left(-\frac{\beta^2}{2\sigma^2\omega^2}\right) \propto \exp\left(-\frac{\beta^2}{2\sigma^2\omega^2}\right).$$

Hence the posterior will be proportional to:

$$\pi(\beta \,|\, x) \propto \exp\left(-\frac{\beta^2 \sum\limits_{i=1}^n \exp(2x_i) - 2\beta \sum\limits_{i=1}^n y_i \exp(x_i)}{2\sigma^2}\right) \exp\left(-\frac{\beta^2}{2\sigma^2\omega^2}\right)$$

$$= \exp\left(-\frac{\beta^2\omega^2 \sum\limits_{i=1}^n \exp(2x_i) - 2\beta\omega^2 \sum\limits_{i=1}^n y_i \exp(x_i) - \beta^2}{2\sigma^2\omega^2}\right)$$

$$= \exp\left(-\frac{\beta^2 - 2\beta \frac{\omega^2 \sum_{i=1}^n y_i \exp(x_i)}{1+\omega^2 \sum_{i=1}^n \exp(2x_i)}}{2\frac{\sigma^2\omega^2}{1+\omega^2 \sum_{i=1}^n \exp(2x_i)}}\right)$$

$$\stackrel{\mathcal{D}}{=} N\left(\frac{\omega^2 \sum\limits_{i=1}^n y_i \exp(x_i)}{1+\omega^2 \sum\limits_{i=1}^n \exp(2x_i)}, \frac{\sigma^2\omega^2}{1+\omega^2 \sum\limits_{i=1}^n \exp(2x_i)}\right).$$

(d) For the least squares criterion in equation (1) the derivative is equal to:

$$2\left(\beta \sum_{i=1}^n \exp(2x_i) - \sum_{i=1}^n y_i \exp(x_i)\right).$$

Setting it equal to 0 and solving, we get that the least squares estimator $\widehat{\theta}^{LS}$ is:

$$\widehat{\theta}^{LS} = \frac{\sum\limits_{i=1}^n y_i \exp(x_i)}{\sum\limits_{i=1}^n \exp(2x_i)}.$$

For the penalised least squares criterion in equation (2) the derivative is:

$$2\beta\sum_{i=1}^{n}\exp(2x_i) - 2\sum_{i=1}^{n} y_i\exp(x_i) + 2\lambda\beta.$$

Setting it equal to 0 and solving, we get that the least squares estimator $\widehat{\theta}^{PLS}$ is:

$$\widehat{\theta}^{PLS} = \frac{\sum_{i=1}^{n} y_i\exp(x_i)}{\lambda + \sum_{i=1}^{n}\exp(2x_i)}.$$

(e) A reasonable Bayes estimator is the posterior mean, which is the same as the posterior mode and median, in each of these cases since the posterior is normal. For part (b) this Bayes estimator is:

$$\frac{\sum_{i=1}^{n} y_i\exp(x_i)}{\sum_{i=1}^{n}\exp(2x_i)}$$

which is the same as $\widehat{\theta}^{LS}$.

For part (c) this Bayes estimator is:

$$\frac{\omega^2\sum_{i=1}^{n} y_i\exp(x_i)}{1+\omega^2\sum_{i=1}^{n}\exp(2x_i)} = \frac{\sum_{i=1}^{n} y_i\exp(x_i)}{\frac{1}{\omega^2}+\sum_{i=1}^{n}\exp(2x_i)}.$$

Note that setting $\omega^2 = 1/\lambda$ gives the same as $\widehat{\theta}^{PLS}$.

(f) Since y_{n+1} is from the same model, we get that:

$$y_{n+1}\,|\,\beta \sim N(\beta\exp(x_{n+1}), \sigma^2)$$

or else, using standard properties of the normal distribution:

$$\frac{y_{n+1}}{\exp(x_{n+1})}\,\Big|\,\beta \sim N\left(\beta, \frac{\sigma^2}{\exp(2x_{n+1})}\right).$$

We also obtained in, say, part (b) that:

$$\beta\,|\,y, x \sim N\left(\frac{\sum_{i=1}^{n} y_i\exp(x_i)}{\sum_{i=1}^{n}\exp(2x_i)}, \frac{\sigma^2}{\sum_{i=1}^{n}\exp(2x_i)}\right).$$

Combining these and using standard properties of the normal distribution, we get that:

$$\frac{y_{n+1}}{\exp(x_{n+1})}\,\Big|\,y, x \sim N\left(\frac{\sum_{i=1}^{n} y_i\exp(x_i)}{\sum_{i=1}^{n}\exp(2x_i)}, \frac{\sigma^2}{\exp(2x_{n+1})} + \frac{\sigma^2}{\sum_{i=1}^{n}\exp(2x_i)}\right)$$

or else:

$$y_{n+1}\,|\,y, x \sim N\left(\exp(x_{n+1})\frac{\sum_{i=1}^{n} y_i\exp(x_i)}{\sum_{i=1}^{n}\exp(2x_i)}, \sigma^2 + \frac{\sigma^2\exp(2x_{n+1})}{\sum_{i=1}^{n}\exp(2x_i)}\right).$$

Question 3

(a) **Consider a model with one dimensional y_i, x_i and data $(y_i, x_i)_{i=1}^n$, where y_i's are binary random variables, taking values 0, 1, and x_i's are continuous random variables. Assume that $x_i \sim N(\mu_0, \sigma_0^2)$ when $y_i = 0$ and that $x_i \sim N(\mu_1, \sigma_1^2)$ when $y_i = 1$, and that the x_i's are independent given the y_i's. Further assume that each y_i is a Bernoulli(π) random variable and that the y_i's are independent.**

i. **Describe why the likelihood function for a pair (y_i, x_i) can be written as**

$$[\pi f(x_i \mid y_i = 1)]^{y_i}[(1-\pi)f(x_i \mid y_i = 0)]^{1-y_i}.$$

ii. **Provide the maximum likelihood estimators for π, μ_0, μ_1, σ_0^2 and σ_1^2 based on all the data.**

iii. **Suppose that logistic regression performs better in your data than the model in parts (a) i. and (a) ii., and suppose you want to predict a future y_i. Provide an example where it would be preferable not to use logistic regression despite its better performance.**

(14 marks)

Reading for this question

This question covers the generative models for classification such as the linear and quadratic discriminant analysis. These can be found in Section 4.4 of the James *et al.* textbook. Additionally, some standard operations using maximum likelihood are needed, that should have been provided by the prerequisites of this course.

Approaching the question

i. If $y_i = 1$ the likelihood be comes $\pi f(x_i \mid y_i = 1)$ which is equal to the probability of $y_i = 1$ times the pdf of x_i given $y_i = 1$. Similarly, for $y_i = 0$.

ii. The likelihood for $\theta = (\pi, \mu_1, \mu_2, \sigma_0^2, \sigma_1^2)$ based on $(y_i, x_i)_{i=1}^n$ can be written as:

$$f(x, y \mid \theta) = \prod_{i=1}^n [\pi N(\mu_1, \sigma_1^2)]^{y_i}[(1-\pi)N(\mu_0, \sigma_0^2)]^{1-y_i}.$$

To maximise with respect to π we write the log-likelihood keeping the terms that involve π:

$$\log f(x, y \mid \pi) = c + \sum_{i=1}^n \{y_i \log \pi + (1-y_i)\log(1-\pi)\}.$$

After differentiating the above w.r.t. π, setting equal to 0 and solving the equation we get:

$$\widehat{\pi} = \frac{1}{n}\sum_{i=1}^n y_i = \frac{n_1}{n} = \frac{n_1}{n_0 + n_1}.$$

To maximise with respect to μ_0 we write the log-likelihood keeping the terms that involve μ_0:

$$\log f(x, y \mid \mu_1) = c + \sum_{i=1}^n y_i \log N(x_i \mid \mu_0, \sigma_0^2) = c - \frac{1}{2}\frac{\sum\limits_{i=1}^n y_i(x_i - \mu_0)^2}{\sigma_0^2}.$$

After differentiating the above w.r.t. π, setting equal to 0 and solving the equation we get:

$$\widehat{\mu}_1 = \frac{\sum\limits_{i=1}^n y_i x_i}{\sum\limits_{i=1}^n y_i} = \frac{\sum\limits_{i=1}^n y_i x_i}{n_1}.$$

Similarly, we obtain:

$$\widehat{\mu}_0 = \frac{\sum\limits_{i=1}^{n}(1-y_i)x_i}{\sum\limits_{i=1}^{n}(1-y_i)} = \frac{\sum\limits_{i=1}^{n}(1-y_i)x_i}{n_0}.$$

For the variance σ_0^2:

$$\log f(x, y \mid \sigma_0^2) = c + \sum_{i=1}^{n}(1-y_i)\log N(x_i \mid \mu_0, \sigma_0^2)$$

$$= c - \frac{n}{2}\log\sigma_0^2 - \frac{1}{2}\frac{\sum\limits_{i=1}^{n}(1-y_i)(x_i-\mu_0)^2}{\sigma_0^2}.$$

After differentiating the above w.r.t. σ_0^2, setting equal to 0 and solving the equation we get:

$$\widehat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}\{(1-y_i)(x_i-\mu_0)^2\}.$$

Given that $\widehat{\mu}_0$ does not depend on $\widehat{\sigma}_0^2$, we can just use:

$$\widehat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}\{(1-y_i)(x_i-\widehat{\mu}_0)^2\}.$$

Finally, for the variance σ_1^2 we get:

$$\log f(x, y \mid \sigma_1^2) = c + \sum_{i=1}^{n} y_1 \log N(x_i \mid \mu_1, \sigma_1^2)$$

$$= c - \frac{n}{2}\log\sigma_0^2 - \frac{1}{2}\frac{\sum\limits_{i=1}^{n} y_i(x_i-\mu_1)^2}{\sigma_1^2}.$$

After differentiating the above w.r.t. σ_1^2, setting equal to 0 and solving the equation we get:

$$\widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}\{y_i(x_i-\mu_1)^2\}.$$

Given that $\widehat{\mu}_1$ does not depend on $\widehat{\sigma}_1^2$, we can just use:

$$\widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}\{y_i(x_i-\widehat{\mu}_1)^2\}.$$

iii. There may be situations where we do not know the future value of x. With LDA we can generate it. With logistic we cannot.

(b) Consider a binary classification problem with two continuous inputs x_1 and x_2 and class label y of 'Red' and 'Blue'. We are given the following training set with 7 cases, in which we would like to explore the maximal margin classifier.

Case	x_1	x_2	y
1	2	4	Red
2	2	2	Red
3	4	4	Red
4	1	3	Red
5	2	1	Blue
6	4	3	Blue
7	3	1	Blue

Provide a graph of the data labelling the class of each point. Also give the equation of a separating hyperplane for the classifier and add it to the graph.

(7 marks)

Reading for this question

This part targets support vector machines presented in Chapter 9 of the James *et al.* textbook. More specifically, the question is on the maximal margin classifier which is covered in Section 9.1.3.

Approaching the question

The maximal margin classifier has to be inbetween observations $(2, 2)$, $(4, 4)$, $(2, 1)$ and $(4, 3)$. Taking points in the middle, we get the two points:

$$(2, 1.5) \quad \text{and} \quad (4, 3.5).$$

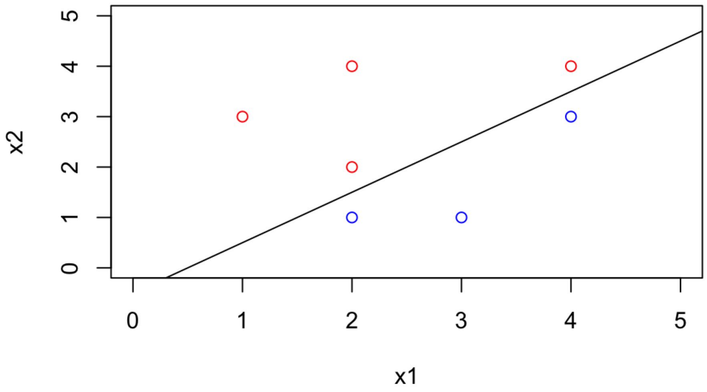
Hence slope is:

$$b = \frac{3.5 - 1.5}{4 - 2} = 1.$$

Checking against the points above, the constant is:

$$a = -0.5.$$

So the line becomes $0.5 - X_1 + X_2 > 0$. A graph of the points and the maximal margin classifier is given below.



(c) **Consider Bagging and suppose that 10 bootstrapped samples have been generated from a data set containing blue and red classes. A classification tree is then applied to each bootstrapped sample and, for a future value of the input $\boldsymbol{X}$, 10 estimates of the probability for the class being blue are produced:**

$$\boldsymbol{P(\text{class is blue} \mid X): \quad 0.25, 0.35, 0.40, 0.45, 0.51, 0.55, 0.60, 0.60, 0.65, 0.70.}$$

Provide the final classification based on the majority voting approach. Repeat for the average probability.

(4 marks)

Reading for this question

This part targets regression and classification trees presented in Chapter 8 of the James *et al.* textbook. More specifically, the concepts of majority vote and average probability in boosting which are presented in Section 8.2.1.

Approaching the question

In 6 out of 10 samples the suggested classification is blue, so this is also the majority approach classification.

The average probability is 0.506 so it also classifies the new sample as blue.

Question 4

(a) **Suppose that we have five observed points, each with three features. We present the correlation between any two observations with measurements on these three features in the following correlation matrix. We use the correlation-based distance to perform the hierarchical clustering.**

	1	2	3	4	5
1	1.00	0.10	0.41	0.55	0.35
2	0.10	1.00	0.64	0.47	0.98
3	0.41	0.64	1.00	0.44	0.85
4	0.55	0.47	0.44	1.00	0.76
5	0.35	0.98	0.85	0.76	1.00

i. **Draw the corresponding Dendrogram using complete linkage and correlation-based distance.**

(9 marks)

ii. **Suppose that we cut the Dendrogram obtained in part (a) i. such that two clusters result. Which observations are in each cluster?**

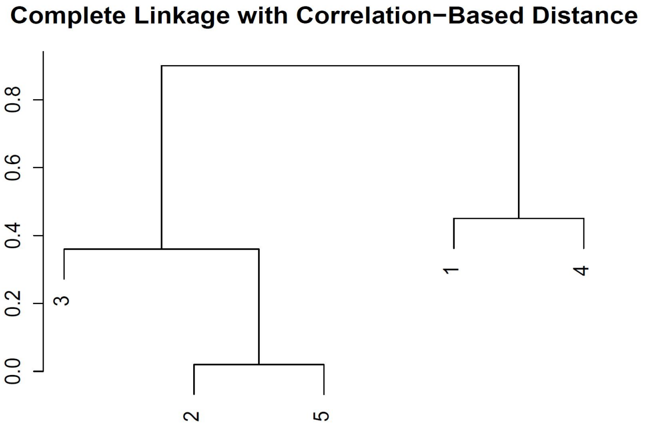
(3 marks)

Reading for this question

This is a question on the technique of hierarchical clustering which is covered in Section 10.2 of the James *et al.* textbook.

Approaching the question

i Note we need use one minus the correlation matrix to obtain the correlation-based distance matrix. The dendrogram is:



ii. We obtain $\{3, 2, 5\}$ in the first cluster and $\{1, 4\}$ in the second cluster.

(b) i. **Show that the variance of the average of B identically distributed variables (each with variance σ^2 and all with positive pairwise correlation ρ) is**

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

(5 marks)

ii. **Using part (b) i.'s result, discuss how the changes of m (number of randomly selected variables during each split) in the random forests and B (number of bootstrap samples) improve the variance reduction of the bagging.**

(4 marks)

Reading for this question

This is an exercise related to the boostrap resampling methods in Section 5.2 of the James *et al.* textbook. This is mainly for part i., but this idea is expanded in part ii. to random forests and bagging that are covered in Section 8.2 of the same textbook.

Approaching the question

i. Let $\mathrm{Var}(X_i) = \sigma^2$ and $\mathrm{Cov}(X_i, X_j) = \rho\sigma^2$ for $1 \le i \ne j \le B$, then:

$$\mathrm{Var}\left(\sum_{i=1}^{B} \frac{X_i}{B}\right) = \sum_{i=1}^{B} \frac{\mathrm{Var}(X_i)}{B^2} + \sum_{i \ne j} \frac{\mathrm{Cov}(X_i, X_j)}{B^2} = \frac{B\sigma^2}{B^2} + \frac{B(B-1)\rho\sigma^2}{B^2} = \rho\sigma^2 + \frac{1-\rho}{B}\,\sigma^2.$$

ii As B increases, the second term and the variance of the average will decrease. As m decreases, ρ will decrease, so the variance of the average will decrease as B is sufficiently large.

(c) **Consider the following binary classification problem with $\boldsymbol{Y = k}$, $\boldsymbol{k \in \{0, 1\}}$. At a data point $\boldsymbol{x}$, the conditional probability of $\boldsymbol{Y = 0}$ is $\boldsymbol{P(Y = 0 \mid X = x) = 0.7}$. Let $\boldsymbol{x'}$ be the nearest neighbour of $\boldsymbol{x}$ and $\boldsymbol{P(Y = 0 \mid X = x') = 0.6}$. What is the 1-neighbour error at $\boldsymbol{x}$?**

(4 marks)

Reading for this question

This is an exercise related to the K-nearest neighbours technique, covered in Section 2.2 of the James *et al.* textbook.

Approaching the question

1-nearest neighbour error at x is:

$$\begin{aligned} & P(Y = 0 \mid X = x)\, P(Y = 1 \mid X = x') + P(Y = 0 \mid X = x')\, P(Y = 1 \mid X = x) \\ & \quad = 0.7 \times (1 - 0.6) + 0.6 \times (1 - 0.7) \\ & \quad = 0.46. \end{aligned}$$

Examiners’ commentary 2021

ST3189 Machine learning

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2020–21. The format and structure of the examination may change in future years, and any such changes will be publicised on the virtual learning environment (VLE).

Information about the subject guide and the Essential reading references

Unless otherwise stated, all cross-references will be to the latest version of the course (2019). You should always attempt to use the most recent edition of any Essential reading textbook, even if the commentary and/or online reading list and/or subject guide refer to an earlier edition. If different editions of Essential reading are listed, please check the VLE for reading supplements – if none are available, please use the contents list and index of the new edition to find the relevant section.

Comments on specific questions – Zone B

Candidates should answer all **FOUR** questions. All questions carry equal marks.

Answer **all** parts of the following questions.

Question 1

(a) **Suppose that we have five observed points, each with three features. We present the correlation between any two observations with measurements on these three features in the following correlation matrix. We use the correlation-based distance to perform the hierarchical clustering.**

	1	2	3	4	5
1	1.00	0.10	0.64	0.55	0.35
2	0.10	1.00	0.41	0.47	0.96
3	0.64	0.41	1.00	0.44	0.85
4	0.55	0.47	0.44	1.00	0.76
5	0.35	0.96	0.85	0.76	1.00

i. **Draw the corresponding Dendrogram using complete linkage and correlation-based distance.**

(9 marks)

ii. **Suppose that we cut the Dendrogram obtained in part (a) i. such that two clusters result. Which observations are in each cluster?**

(3 marks)

Reading for this question

This is a question on the technique of hierarchical clustering which is covered in Section 10.2 of the James *et al.* textbook.

Approaching the question

i Note we need use one minus the correlation matrix to obtain the correlation-based distance matrix.

	1	2	3	4	5
1	0.00	0.90	0.36	0.45	0.65
2	0.90	0.00	0.59	0.53	0.04
3	0.36	0.59	0.00	0.56	0.15
4	0.45	0.53	0.56	0.00	0.24
5	0.65	0.04	0.15	0.24	0.00

The minimum values is 0.04 so we first group $(2, 5)$ together.

Then, we recompute the above matrix with $(2, 5)$ as one cluster. The new elements are coming from the 'distances':

$$D[(2,5),1] = \max[D(2,1), D(5,1)] = \max[0.9, 0.65] = 0.9$$

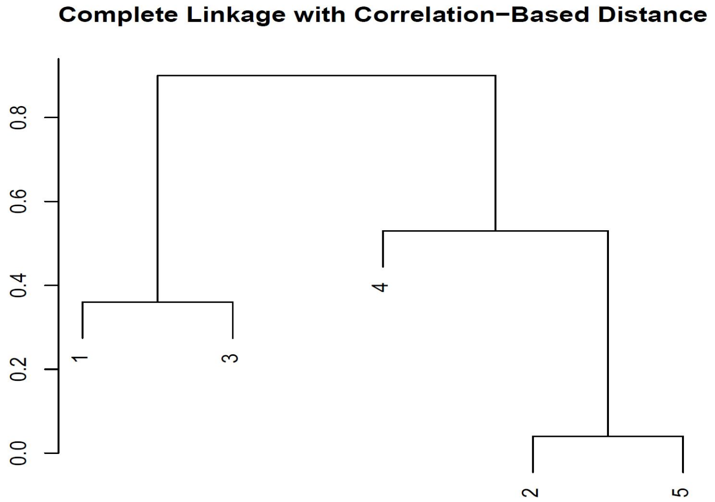
$$D[(2,5),3] = \max[D(2,3), D(5,3)] = \max[0.59, 0.15] = 0.59$$

$$D[(2,5),4] = \max[D(2,4), D(5,4)] = \max[0.53, 0.24] = 0.53$$

and the 'distances' $D(1,3) = 0.36$, $D(1,4) = 0.45$ and $D(3,4) = 0.56$. The smallest distance is 0.36, so we group $(1,3)$ together.

Now we have the clusters $(1,3)$, 4 and $(2,5)$. We have $D[(2,5),4] = 0.53$ from the previous step and calculate $D[(1,3),4] = 0.56$ and $D[(1,3),(2,5)] = 0.9$, so we merge 4 and $(2,5)$ together.

The final step just merges together the clusters $(1,3)$ and $(2,4,5)$. All of these are summarised in the figure below that plots the Dendrogram using complete linkage.



ii. We obtain $\{1, 3\}$ in the first cluster and $\{4, 2, 5\}$ in the second cluster.

(b) i. Show that the variance of the average of B identically distributed variables (each with variance σ^2 and all with positive pairwise correlation ρ) is

$$\rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

(5 marks)

ii. **Using part (b) i.'s result, discuss how the changes of m (number of randomly selected variables during each split) in the random forests and B (number of bootstrap samples) improve the variance reduction of the bagging.**

(4 marks)

Reading for this question

This is an exercise related to the boostrap resampling methods in Section 5.2 of the James *et al.* textbook. This is mainly for part i., but this idea is expanded in part ii. to random forests and bagging that are covered in Section 8.2 of the same textbook.

Approaching the question

i. Let $\text{Var}(X_i) = \sigma^2$ and $\text{Cov}(X_i, X_j) = \rho\sigma^2$ for $1 \leq i \neq j \leq B$, then:

$$\text{Var}\left(\sum_{i=1}^{B} \frac{X_i}{B}\right) = \sum_{i=1}^{B} \frac{\text{Var}(X_i)}{B^2} + \sum_{i \neq j} \frac{\text{Cov}(X_i, X_j)}{B^2} = \frac{B\sigma^2}{B^2} + \frac{B(B-1)\rho\sigma^2}{B^2} = \rho\sigma^2 + \frac{1-\rho}{B}\sigma^2.$$

ii As B increases, the second term and the variance of the average will decrease. As m decreases, ρ will decrease, so the variance of the average will decrease as B is sufficiently large.

(c) **Consider the following binary classification problem with $Y = k$, $k \in \{0, 1\}$. At a data point x, the conditional probability of $Y = 0$ is $P(Y = 0 \,|\, X = x) = 0.5$. Let x' be the nearest neighbour of x and $P(Y = 0 \,|\, X = x') = 0.4$. What is the 1-neighbour error at x?**

(4 marks)

Reading for this question

This is an exercise related to the K-nearest neighbours technique, covered in Section 2.2 of the James *et al.* textbook.

Approaching the question

Given x, an error is committed if $Y = 0$ and $Y = 1$ is chosen from 1-NN algorithm (based on x') or, if $Y = 1$ and $Y = 0$ is chosen from 1-NN algorithm (based on x').

The above occurs with probability:

$$\begin{aligned} & P(Y = 0 \,|\, X = x)\, P(Y = 1 \,|\, X = x') + P(Y = 0 \,|\, X = x')\, P(Y = 1 \,|\, X = x) \\ & \quad = 0.5 \times (1 - 0.4) + 0.4 \times (1 - 0.5) \\ & \quad = 0.5. \end{aligned}$$

Question 2

Consider a linear regression setting where the response variable is $y = (y_1, \ldots, y_n)$ and there is one feature, or else predictor, $x = (x_1, \ldots, x_n)$. We are interested in fitting the following model

$$y_i = \beta x_i^2 + \epsilon_i, \quad i = 1, \ldots, n,$$

where the error terms ϵ_i's are independent and distributed according to the normal distribution with mean 0 and known variance σ^2. Equivalently, we can write that given x each y_i is independent and distributed according to the normal distribution with mean βx_i^2 and known variance σ^2.

(a) **Derive the likelihood function for the unknown parameter β.**

(2 marks)

(b) **Derive the Jeffreys prior for β. Use it to obtain the corresponding posterior distribution.**

(5 marks)

(c) **Consider the normal distribution prior for β with zero mean and variance ω^2. Use it to obtain the corresponding posterior distribution.**

(5 marks)

(d) **Consider the least squares criterion**

$$\sum_{i=1}^{n}(y_i - \beta x_i^2)^2, \tag{1}$$

and show that the estimator of β that minimises equation (1), also maximises the likelihood function derived in part (a). Derive this estimator and, in addition, consider the following penalised least squares criterion

$$\left\{\sum_{i=1}^{n}(y_i - \beta x_i^2)^2\right\} + \lambda\beta^2, \tag{2}$$

given a $\lambda > 0$. Derive the estimator of β that minimises equation (2) and compare with the one that minimises equation (1).

(4 marks)

(e) **Provide a Bayes estimator for each of the posteriors in parts (b) and (c) and compare them with the estimators of part (d).**

(5 marks)

(f) **Let y_{n+1} represent a future observation from the same model given the corresponding value of the predictor x_{n+1}. Find the posterior predictive distribution of y_{n+1} for one of the posteriors in parts (b) or (c).**

(4 marks)

Reading for this question

This question is examining Bayesian inference which can be found in Block 4 of the VLE section of the course. Read the parts on 'Bayesian Inference Essentials' and 'Bayesian Inference Examples'. Exercises 1–6 as well as Exercise 2 of the mock examination are relevant for practice (try to do a few of them). Also the part on least squares and the part on shrinkage methods, in Chapter 3 and Section 6.2 of James *et al.*, respectively, are relevant for part (d).

Approaching the question

(a) The likelihood can be written as:

$$\begin{aligned} L(\beta\,|\,x) = f(x\,|\,\beta) &= \prod_{i=1}^{n}(2\pi\sigma^2)^{-1/2}\exp\left(-\frac{(y_i-\beta x_i^2)^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{\sum\limits_{i=1}^{n}(y_i-\beta x_i^2)^2}{2\sigma^2}\right) \\ &\propto \exp\left(-\frac{\beta^2\sum\limits_{i=1}^{n}x_i^4 - 2\beta\sum\limits_{i=1}^{n}y_ix_i^2}{2\sigma^2}\right). \end{aligned}$$

(b) The log-likelihood can be written as:

$$\ell(\beta \,|\, x) = -\frac{\beta^2 \sum_{i=1}^{n} x_i^4 - 2\beta \sum_{i=1}^{n} y_i x_i^2}{2\sigma^2}.$$

In order to find Jeffreys prior we calculate:

$$\frac{\partial \ell(\beta \,|\, x)}{\partial \theta} = -\frac{\beta \sum_{i=1}^{n} x_i^4 - \sum_{i=1}^{n} y_i x_i^2}{\sigma^2}$$

also:

$$\frac{\partial^2 \ell(\beta \,|\, x)}{\partial \beta^2} = -\frac{\sum_{i=1}^{n} x_i^4}{\sigma^2}$$

and the Fisher informatio:

$$\mathcal{I}(\beta) = -\mathrm{E}\left(-\frac{\sum_{i=1}^{n} x_i^4}{\sigma^2}\right) = \frac{\sum_{i=1}^{n} x_i^4}{\sigma^2}.$$

The Jeffreys prior is:

$$\pi^J(\beta) \propto \mathcal{I}(\beta)^{1/2} \propto 1.$$

Using the Jeffreys prior, the corresponding posterior $\pi^J(\beta \,|\, x)$ is proportional to:

$$\begin{aligned}\pi^J(\beta \,|\, x) &\propto \exp\left(-\frac{\beta^2 \sum_{i=1}^{n} x_i^4 - 2\beta \sum_{i=1}^{n} y_i x_i^2}{2\sigma^2}\right)\\ &= \exp\left(-\frac{\beta^2 - 2\beta \frac{\sum_{i=1}^{n} y_i x_i^2}{\sum_{i=1}^{n} x_i^4}}{2\frac{\sigma^2}{\sum_{i=1}^{n} x_i^4}}\right)\\ &\overset{\mathcal{D}}{=} N\left(\frac{\sum_{i=1}^{n} y_i x_i^2}{\sum_{i=1}^{n} x_i^4}, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^4}\right).\end{aligned}$$

(c) The prior of β can be written as:

$$\pi(\beta) = (2\pi\omega^2)^{-1/2} \exp\left(-\frac{\beta^2}{2\omega^2}\right) \propto \exp\left(-\frac{\beta^2}{2\omega^2}\right).$$

Hence the posterior will be proportional to:

$$\pi(\beta \,|\, x) \propto \exp\left(-\frac{\beta^2 \sum\limits_{i=1}^{n} x_i^4 - 2\beta \sum\limits_{i=1}^{n} y_i x_i^2}{2\sigma^2}\right) \exp\left(-\frac{\beta^2}{2\omega^2}\right)$$

$$= \exp\left(-\frac{\beta^2\omega^2 \sum\limits_{i=1}^{n} x_i^4 - 2\beta\omega^2 \sum\limits_{i=1}^{n} y_i x_i^2 + \beta^2\sigma^2}{2\sigma^2\omega^2}\right)$$

$$= \exp\left(-\frac{\beta^2 - 2\beta \frac{\omega^2 \sum_{i=1}^{n} y_i x_i^2}{\sigma^2 + \omega^2 \sum_{i=1}^{n} x_i^4}}{2\frac{\sigma^2\omega^2}{\sigma^2 + \omega^2 \sum_{i=1}^{n} x_i^4}}\right)$$

$$\overset{\mathcal{D}}{=} N\left(\frac{\omega^2 \sum\limits_{i=1}^{n} y_i x_i^2}{\sigma^2 + \omega^2 \sum\limits_{i=1}^{n} x_i^4}, \frac{\sigma^2\omega^2}{\sigma^2 + \omega^2 \sum\limits_{i=1}^{n} x_i^4}\right).$$

(d) For the least squares criterion in equation (1) the derivative is equal to:

$$2\left(\beta \sum_{i=1}^{n} x_i^4 - \sum_{i=1}^{n} y_i x_i^2\right).$$

Setting it equal to 0 and solving, we get that the least squares estimator $\widehat{\theta}^{LS}$ is:

$$\widehat{\theta}^{LS} = \frac{\sum\limits_{i=1}^{n} y_i x_i^2}{\sum\limits_{i=1}^{n} x_i^4}.$$

For the penalised least squares criterion in equation (2) the derivative is:

$$2\beta \sum_{i=1}^{n} x_i^4 - 2\sum_{i=1}^{n} y_i x_i^2 + 2\lambda\beta.$$

Setting it equal to 0 and solving, we get that the least squares estimator $\widehat{\theta}^{PLS}$ is:

$$\widehat{\theta}^{PLS} = \frac{\sum\limits_{i=1}^{n} y_i x_i^2}{\lambda + \sum\limits_{i=1}^{n} x_i^4}.$$

(e) A reasonable Bayes estimator is the posterior mean, which is the same as the posterior mode and median, in each of these cases since the posterior is normal. For part (b) this Bayes estimator is:

$$\frac{\sum\limits_{i=1}^{n} y_i x_i^2}{\sum\limits_{i=1}^{n} x_i^4}$$

which is the same as $\widehat{\theta}^{LS}$.

For part (c) this Bayes estimator is:

$$\frac{\omega^2 \sum_{i=1}^{n} y_i x_i^2}{\sigma^2 + \omega^2 \sum_{i=1}^{n} x_i^4} = \frac{\sum_{i=1}^{n} y_i x_i^2}{\frac{\sigma^2}{\omega^2} + \sum_{i=1}^{n} x_i^4}.$$

Note that setting $\omega^2 = \sigma^2/\lambda$ gives the same as $\widehat{\theta}^{PLS}$.

(f) Since y_{n+1} is from the same model, we get that:

$$y_{n+1} \,|\, \beta \sim N(\beta x_{n+1}^2, \sigma^2)$$

or else, using standard properties of the normal distribution:

$$\frac{y_{n+1}}{x_{n+1}^2} \,\bigg|\, \beta \sim N\left(\beta, \frac{\sigma^2}{x_{n+1}^4}\right).$$

We also obtained in, say, part (b) that:

$$\beta \,|\, y, x \sim N\left(\frac{\sum_{i=1}^{n} y_i x_i^2}{\sum_{i=1}^{n} x_i^4}, \frac{\sigma^2}{\sum_{i=1}^{n} x_i^4}\right).$$

Combining these and using standard properties of the normal distribution, we get that:

$$\frac{y_{n+1}}{x_{n+1}^2} \,\bigg|\, y, x \sim N\left(\frac{\sum_{i=1}^{n} y_i x_i^2}{\sum_{i=1}^{n} x_i^4}, \frac{\sigma^2}{x_{n+1}^4} + \frac{\sigma^2}{\sum_{i=1}^{n} x_i^4}\right)$$

or else:

$$y_{n+1} \,|\, y, x \sim N\left(x_{n+1}^2 \frac{\sum_{i=1}^{n} y_i x_i^2}{\sum_{i=1}^{n} x_i^4}, \sigma^2 + \frac{\sigma^2 x_{n+1}^4}{\sum_{i=1}^{n} x_i^4}\right).$$

Question 3

(a) Consider a model with one dimensional y_i, x_i and data $(y_i, x_i)_{i=1}^n$, where y_i's are binary random variables, taking values 0, 1, and x_i's are continuous random variables. Assume that $x_i \sim N(\mu_0, \sigma_0^2)$ when $y_i = 0$ and that $x_i \sim N(\mu_1, \sigma_1^2)$ when $y_i = 1$, and that the x_i's are independent given the y_i's. Further assume that each y_i is a Bernoulli(π) random variable and that the y_i's are independent.

i. Describe why the likelihood function for a pair (y_i, x_i) can be written as

$$[\pi f(x_i \,|\, y_i = 1)]^{y_i} [(1 - \pi) f(x_i \,|\, y_i = 0)]^{1-y_i}.$$

ii. Provide the maximum likelihood estimators for π, μ_0, μ_1, σ_0^2 and σ_1^2 based on all the data.

iii. Suppose that logistic regression performs better in your data than the model in parts (a) i. and (a) ii., and suppose you want to predict a future y_i. Provide an example where it would be preferable not to use logistic regression despite its better performance.

(14 marks)

Reading for this question

This question covers the generative models for classification such as the linear and quadratic discriminant analysis. These can be found in Section 4.4 of the James *et al.* textbook. Additionally, some standard operations using maximum likelihood are needed, that should have been provided by the prerequisites of this course.

Approaching the question

i. If $y_i = 1$ the likelihood be comes $\pi f(x_i \,|\, y_i = 1)$ which is equal to the probability of $y_i = 1$ times the pdf of x_i given $y_i = 1$. Similarly, for $y_i = 0$.

ii. The likelihood for $\theta = (\pi, \mu_1, \mu_2, \sigma_0^2, \sigma_1^2)$ based on $(y_i, x_i)_{i=1}^n$ can be written as:

$$f(x, y \,|\, \theta) = \prod_{i=1}^n [\pi N(\mu_1, \sigma_1^2)]^{y_i} [(1-\pi) N(\mu_0, \sigma_0^2)]^{1-y_i}.$$

To maximise with respect to π we write the log-likelihood keeping the terms that involve π:

$$\log f(x, y \,|\, \pi) = c + \sum_{i=1}^n \{y_i \log \pi + (1-y_i)\log(1-\pi)\}.$$

After differentiating the above w.r.t. π, setting equal to 0 and solving the equation we get:

$$\widehat{\pi} = \frac{1}{n}\sum_{i=1}^n y_i = \frac{n_1}{n} = \frac{n_1}{n_0 + n_1}.$$

To maximise with respect to μ_0 we write the log-likelihood keeping the terms that involve μ_0:

$$\log f(x, y \,|\, \mu_1) = c + \sum_{i=1}^n y_i \log N(x_i \,|\, \mu_0, \sigma_0^2) = c - \frac{1}{2}\frac{\sum\limits_{i=1}^n y_i (x_i - \mu_0)^2}{\sigma_0^2}.$$

After differentiating the above w.r.t. π, setting equal to 0 and solving the equation we get:

$$\widehat{\mu}_1 = \frac{\sum\limits_{i=1}^n y_i x_i}{\sum\limits_{i=1}^n y_i} = \frac{\sum\limits_{i=1}^n y_i x_i}{n_1}.$$

Similarly, we obtain:

$$\widehat{\mu}_0 = \frac{\sum\limits_{i=1}^n (1-y_i) x_i}{\sum\limits_{i=1}^n (1-y_i)} = \frac{\sum\limits_{i=1}^n (1-y_i) x_i}{n_0}.$$

For the variance σ_0^2:

$$\begin{aligned}\log f(x, y \,|\, \sigma_0^2) &= c + \sum_{i=1}^n (1-y_i) \log N(x_i \,|\, \mu_0, \sigma_0^2)\\ &= c - \frac{n}{2}\log \sigma_0^2 - \frac{1}{2}\frac{\sum\limits_{i=1}^n (1-y_i)(x_i-\mu_0)^2}{\sigma_0^2}.\end{aligned}$$

After differentiating the above w.r.t. σ_0^2, setting equal to 0 and solving the equation we get:

$$\widehat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^n \{(1-y_i)(x_i - \mu_0)^2\}.$$

Given that $\widehat{\mu}_0$ does not depend on $\widehat{\sigma}_0^2$, we can just use:

$$\widehat{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^{n}\{(1-y_i)(x_i-\widehat{\mu}_0)^2\}.$$

Finally, for the variance σ_1^2 we get:

$$\log f(x, y \mid \sigma_1^2) = c + \sum_{i=1}^{n} y_1 \log N(x_i \mid \mu_1, \sigma_1^2)$$

$$= c - \frac{n}{2}\log\sigma_0^2 - \frac{1}{2}\frac{\sum_{i=1}^{n} y_i(x_i-\mu_1)^2}{\sigma_1^2}.$$

After differentiating the above w.r.t. σ_1^2, setting equal to 0 and solving the equation we get:

$$\widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}\{y_i(x_i-\mu_1)^2\}.$$

Given that $\widehat{\mu}_1$ does not depend on $\widehat{\sigma}_1^2$, we can just use:

$$\widehat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}\{y_i(x_i-\widehat{\mu}_1)^2\}.$$

iii. There may be situations where we do not know the future value of x. With LDA we can generate it. With logistic we cannot.

(b) Consider a binary classification problem with two continuous inputs x_1 and x_2 and class label y of 'Red' and 'Blue'. We are given the following training set with 7 cases, in which we would like to explore the maximal margin classifier.

Case	x_1	x_2	y
1	**3**	**5**	**Red**
2	**2**	**2**	**Red**
3	**4**	**4**	**Red**
4	**1**	**2**	**Red**
5	**2**	**1**	**Blue**
6	**4**	**3**	**Blue**
7	**4**	**2**	**Blue**

Provide a graph of the data labelling the class of each point. Also give the equation of a separating hyperplane for the classifier and add it to the graph.

(7 marks)

Reading for this question

This part targets support vector machines presented in Chapter 9 of the James *et al.* textbook. More specifically, the question is on the maximal margin classifier which is covered in Section 9.1.3.

Approaching the question

The maximal margin classifier has to be inbetween observations $(2, 2)$, $(4, 4)$, $(2, 1)$ and $(4, 3)$. Taking points in the middle, we get the two points:

$$(2, 1.5) \quad \text{and} \quad (4, 3.5).$$

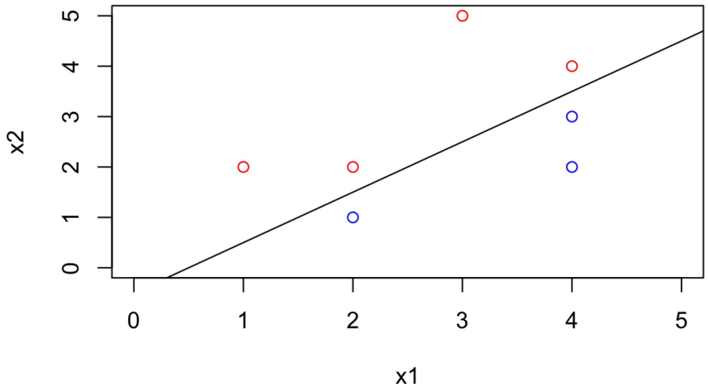
Hence slope is:

$$b = \frac{3.5 - 1.5}{4 - 2} = 1.$$

Checking against the points above, the constant is:

$$a = -0.5.$$

So the line becomes $0.5 - X_1 + X_2 > 0$. A graph of the points and the maximal margin classifier is given below.



(c) **Consider Bagging and suppose that 10 bootstrapped samples have been generated from a data set containing blue and green classes. A classification tree is then applied to each bootstrapped sample and, for a future value of the input $\boldsymbol{X}$, 10 estimates of the probability for the class being green are produced:**

$$\boldsymbol{P(\text{class is green} \mid X) : \quad 0.30, 0.35, 0.40, 0.45, 0.55, 0.60, 0.60, 0.65, 0.70, 0.75.}$$

Provide the final classification based on the majority voting approach. Repeat for the average probability.

(4 marks)

Reading for this question

This part targets regression and classification trees presented in Chapter 8 of the James *et al.* textbook. More specifically, the concepts of majority vote and average probability in boosting which are presented in Section 8.2.1.

Approaching the question

In 6 out of 10 samples the suggested classification is green, so this is also the majority approach classification.

The average probability is 0.535 so it also classifies the new sample as green.

Question 4

(a) **Indicate whether the following statements are true or false. Briefly justify your answers.**

i. **The partitioning of the input space shown in Figure 1 could be generated by recursive binary splitting.**

(3 marks)

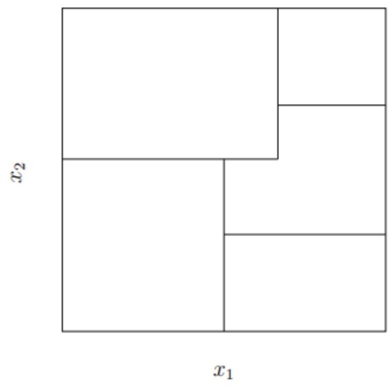


Figure 1: For Question 4. (a) i.

ii. **In random forests, for each tree a random selection of observations are discarded in order to decorrelate the trees.**

(3 marks)

iii. **The maximum likelihood estimates for α_1, α_2 in the model $y_i = e^{\alpha_1 x_i + \alpha_2} + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$ are i.i.d. noise, can be obtained using linear regression.**

(3 marks)

Reading for this question

This question refers to several techniques and machine learning concepts of the course requiring some basic understanding of them. More specifically, part iii. is on linear regression that can be found in several sections of the James *et al.* textbook but mainly in Chapter 3. For parts i. and ii. the content of tree-based methods (Chapter 8 of James *et al.*) such as the procedure to obtain a tree (part i.) and random forests (part ii.).

Approaching the question

Remember that the justification has to be one sentence so organise your thoughts accordingly and avoid lengthy answers. Some 'good answers' are provided below. Note that there can be more than one 'correct' answer in some of these questions.

i. False. Recursive binary splitting ensures axis-alighed splits, so the top-left split is impossible.

ii. False. In order to decorrelate the trees, a random selection of input variables at each split (not observations) are discarded.

iii. True. The maximum likelihood estimates for α_1, α_2 in the model $y_i = e^{\alpha_1 x_i + \alpha_2} + \varepsilon_i$, where $\varepsilon_i \sim N(0, \sigma^2)$ are i.i.d. noise, can be obtained using linear regression by way of a logarithmic transformation.

Overall, candidates did okay on parts i. and iii. but not so good on part ii.

(b) **Consider a learning problem with two features. How are the decision tree and 1-nearest neighbour decision boundaries related? Specifically, discuss the similarities and dissimilarities.**

(4 marks)

Reading for this question

This question examines the main ideas of tree-based methods and K-nearest neighbours, that can be found in Chapter 8 and Section 2.2 of James *et al.*, respectively.

Approaching the question

In both cases, the decision boundary is piecewise linear. Decision trees do axis-aligned splits while 1-nearest neighbour gives a voronoi diagram.

(c) i. **Provide a 2-dimensional dataset where 1-nearest neighbour has lower Leave-One-Out Cross Validation (LOOCV) error than linear support vector machines.**

(3 marks)

ii. **Provide a 2-dimensional dataset where 1-nearest neighbour has higher LOOCV error than linear support vector machines.**

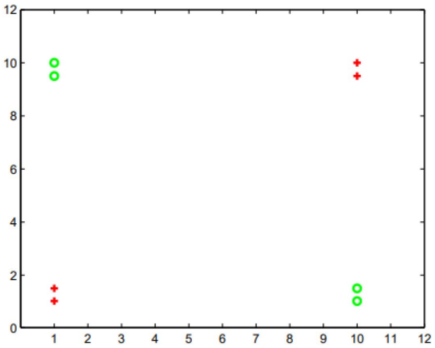
(3 marks)

Reading for this question

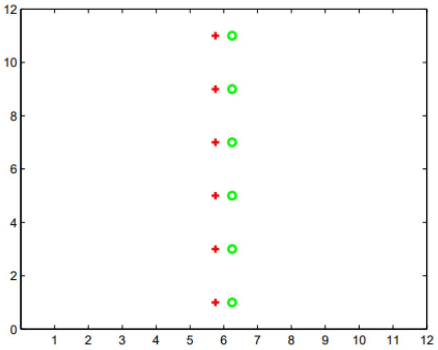
This question contains material regarding the $K-$nearest neighbours technique that can be found in Section 2.2 of the James *et al.* textbook. The content of support vector machines is also relevant and can be found in Chapter 9.

Approaching the question

i. For example:



ii. For example:



(d) **Consider the K-nearest neighbours approach using Euclidean distance on the dataset shown in Figure 2. What are the LOOCV errors for the following cases? Briefly justify your answers.**

i. **3-nearest neighbours.**

(3 marks)

ii. **1-nearest neighbour.**

(3 marks)

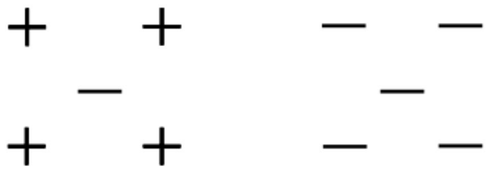


Figure 2: For Question 4. (d)

Reading for this question

As before the content on K-nearest neighbours and support vector machines is relevant but this time knowledge on the leave-one-out cross-validations is required that can be found in Chapter 5 of the James *et al.* textbook.

Approaching the question

i. The left '–' point is misclassified and the error is 1/10.

ii. The left-hand points are misclassified and the error is 5/10.