Examiners' commentary 2022

ST3189 Machine learning

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2021–22. 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.

Online examination

Over the last years, due to the measures taken to ensure the health and safety of candidates and everyone involved in light of the Covid-19 pandemic, the examinations were converted into online assessments. In these types of assessment candidates are typically given an extra amount of time to prepare and upload their examination script. Some notes of good practice for such circumstances are given below:

- Do not leave uploading for the last minute. Upload a file as soon as you can; remember you can always change it at any time before the deadline.
- Use the extra time to make sure your writing is legible. If needed rewrite a few parts; you can always pick up small things in that way.
- For graphs do use some graph paper if you have it handy. It is fine if you do not, but there is certainly no harm to use graph paper; it is also helpful to you in order to draw your graph.
- Make sure your candidate number is legible.

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 2022

ST3189 Machine learning

Important note

This commentary reflects the examination and assessment arrangements for this course in the academic year 2021–22. 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

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

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

Question 1

(a) **The lasso and best subset selection can be used for variable selection. Discuss the main advantage and disadvantage of the lasso compared with best subset selection.**

(4 marks)

Reading for this question

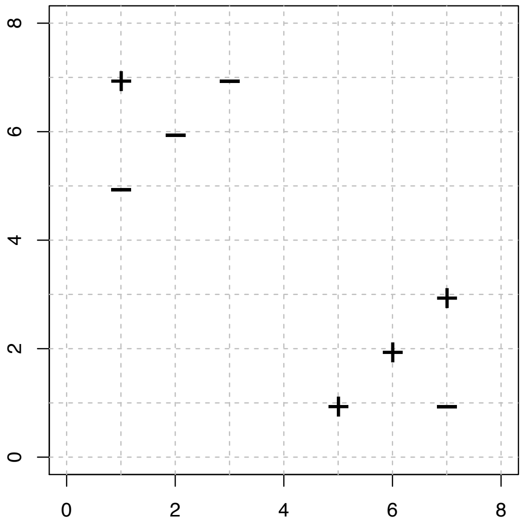
This question refers to variable selection and shrinkage methods in linear regression covered in the relevant block. You are advised to go over the notes and conduct the corresponding exercises and activities to get a good understanding of the concepts.

Approaching the question

The justification does not need to be long and it would be good to organise your thoughts accordingly and provide a brief answer such as the following:

The lasso is computationally more efficient than best subset selection. The lasso estimates suffer from the bias of large coefficients, while the estimates by best subset selection are unbiased.

(b) **Consider the k-nearest neighbours classification using the Euclidean distance on the dataset shown in Figure 1.**



For Question 1 (b).

i. **Sketch the 1-nearest neighbour decision boundary and identify regions classified as '+' and '−', respectively.**

(6 marks)

ii. **What is the Leave-One-Out Cross Validation (LOOCV) error when using 3-nearest neighbours?**

(3 marks)

iii. **What is the LOOCV error when using 5-nearest neighbours?**

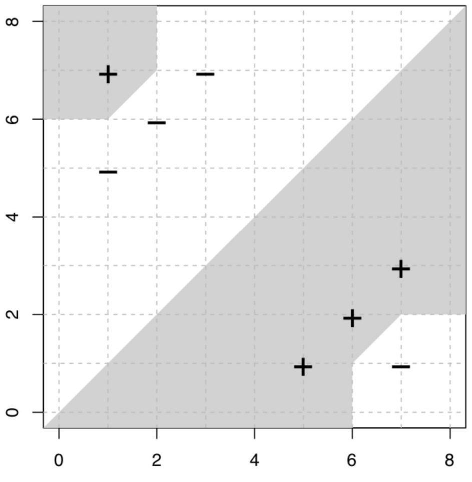
(3 marks)

Reading for this question

This question examines the main ideas of k-nearest neighbours, that can be found in Chapter 8 and Section 2.2 of the James *et al.* textbook, respectively. It is essential to practise similar exercises; see also other *Examiners' commentaries* to reinforce your understanding of the k-nearest neighbours approach.

Approaching the question

i. The decision boundary is shown below with white from class '−' and grey from class '+'.



ii. The top-left and bottom-right points are misclassified and the error is 1/4.

iii. All points are misclassified and the error is 1.

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

i. If the sensitivity of a classifier increases, so does its specificity.

(3 marks)

ii. Quadratic discriminant analysis can only produce a quadratic decision boundary.

(3 marks)

iii. If we train a linear regression estimator on only half the data, the variance of the estimator will be larger than training it on the entire dataset.

(3 marks)

Reading for this question

The questions here target different topics of the syllabus. Sensitivity and specificity are about assessing predictions in classification and can be found in the relevant chapter, specifically on page 149 of the James *et al.* textbook. The second part targets quadratic discriminant analysis which is covered in Section 4.4.3 of the same textbook. Finally, part iii. is cross-validation; see Section 5.1 of the same textbook.

Approaching the question

The justifications do not need to be long and it would be good to organise your thoughts accordingly and provide brief answers such as the following:

i. No, the specificity actually tends to decrease.

ii. No, if a 3rd-order polynomial of a covariate is treated as a new covariate, then the decision boundary will be non-quadratic.

iii. Yes, the variance will be larger since halving the data will result in loss of information and therefore higher uncertainty.

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, $x = (x_1, \ldots, x_n)$, where $x_i > 0$ for all $i = 1, \ldots, n$. We are interested in fitting the following model

$$y_i = \beta\sqrt{x_i} + \epsilon_i, \quad i = 1, \ldots, n,$$

where the error terms ϵ_is 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 $\beta\sqrt{x_i}$ and known variance σ^2.

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

(3 marks)

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

(6 marks)

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

(6 marks)

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

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

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

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

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

(5 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)

Reading for this question

This question is examining Bayesian inference that 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 the James *et al.* textbook, respectively, are relevant for part (d).

Approaching the question

(a) The likelihood can be written as:

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

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

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

In order to find Jeffreys prior we calculate:

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

also:

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

and the Fisher information:

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

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:

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

(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:

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

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

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

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

$$\widehat{\theta}^{LS} = \frac{\sum_{i=1}^{n} y_i \sqrt{x_i}}{\sum_{i=1}^{n} x_i}.$$

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

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

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

$$\widehat{\theta}^{PLS} = \frac{\sum_{i=1}^{n} y_i \sqrt{x_i}}{\lambda + \sum_{i=1}^{n} x_i}.$$

(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) 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:

$$\frac{\sum_{i=1}^{n} y_i \sqrt{x_i}}{\sum_{i=1}^{n} x_i},$$

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

For part (c) this Bayes estimator is:

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

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

Question 3

(a) **Consider the regression task of predicting the variable Y based on the variable X given the following training sample:**

Y	X
7	8
6	9
8	7
3	1
4	0

Apply the recursive binary splitting algorithm to produce a regression tree. The objective is to minimise the residual sum of squares (RSS)

$$RSS = \sum_m \sum_{i:i\in R_m} (Y_i - c_m)^2,$$

where c_m is the prediction for Y_i corresponding to the region R_m of the tree. The stopping criterion, in order to find the regions R_m of the tree, requires all nodes to have less than 4 observations. Provide the splitting rules, the regions R_m and a diagram of the tree as well as your calculations in detail.

(13 marks)

Reading for this question

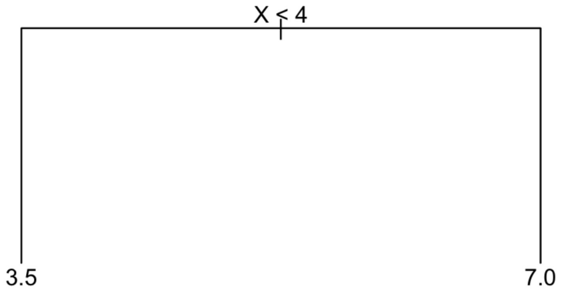
The question asks you to apply the recursive binary splitting algorithm to a simple numerical example. You can read about this algorithm in Section 8.1.1 of the James *et al.* textbook, mainly on pages 330 and 331.

Approaching the question

The steps of applying recursive binary splitting to these data are shown below:

- To start let's compute the RSS in case of no splits. This is simply $\sum_i (Y_i - \bar{Y})^2$ which is equal to 17.2.
- Now let's look for the first split. The cutpoint for X can be either:
  - ∗ between $X = 0$ and $X = 1$ – say 0.5
  - ∗ between $X = 1$ and $X = 7$ – say 4
  - ∗ between $X = 7$ and $X = 8$ – say 7.5
  - ∗ between $X = 8$ and $X = 9$ – say 8.5.
- Let us now see the RSS corresponding to the above cutpoints:
  - ∗ we get a prediction 4 for $Y = 0$ and $(7 + 6 + 8 + 3)/4 = 6$ for the other Ys. The RSS will be $0^2 + 1^2 + 0^2 + 2^2 + (-3)^2 = 14$.
  - ∗ we get a prediction of 3.5 for $Y = 3, 4$ and 7 for $Y = 6$, 7, 8. The RSS will be $(-0.5)^2 + 0.5^2 + (-1)^2 + 0^2 + 1^2 = 2.5$
  - ∗ we get a prediction of 5 for $Y = 3, 4, 8$ and 6.5 for $Y = 6$, 7. The RSS will be $(-2)^2 + (-1)^2 + 3^2 + (-0.5)^2 + 0.5^2 = 14.5$
  - ∗ we get a prediction 5.5 for $Y = 3$, 4, 8, 7 and 6 for $Y = 6$. The RSS will be $(-2.5)^2 + (-1.5)^2 + 2.5^2 + 1.5^2 + 0^2 = 10.75$.
- From the above the split with the greatest reduction in RSS is between $X = 1$ and $X = 7$, i.e. 4. The two regions are $X < 4$ and $X \geq 4$.
- Both nodes now have less than 4 observations so the tree is complete.

The tree can be drawn as:



(b) Suppose we wish to perform k-means clustering with $k = 2$ on the following data set containing five observations and one variable: $X = (-3, -4, 2, 3, 5)$. Suppose that our random initialisation ends up with two cluster centres at the following locations: Cluster Centre 1: $X = 1$; Cluster Centre 2: $X = 4$.

i. Show how the k-means algorithm will work from this point on. You need to indicate what the initial cluster assignments will be, how the cluster centres and assignments change at each step, as well as the final cluster assignments and centres. Note that you should only need to do this for a few iterations before you get the final solution.

(8 marks)

ii. What would happen in the k-means algorithm if the observation 2 was actually recorded wrong and its correct value was 1?

(4 marks)

Reading for this question

This part targets k-means clustering which is covered in Section 12.4.1 of the James *et al.* textbook. Make sure to read this section and practise on its examples.

Approaching the question

The requested steps are shown below:

i. The k-means clustering can be performed via the following steps:

- If we work through this you will see that initially observations -3, -4 and 2 are closest to cluster centre 1 and observations 3 and 5 are closest to cluster centre 2.
- The new cluster centres will then be the average of observations -3, -4 and 2, which is $-5/3$, and the average of observations 3 and 5, which is 4.
- Now observations -3 and -4 are closest to centre 1 and the other observations are closest to centre 2.
- The new cluster centre for cluster 1 will then be the average of observations -3 and -4, which is -3.5, and the new cluster centre for cluster 2 will be the average of observations 2, 3 and 5, which is $10/3$ (3.33).
- Now observations 1 and 2 are closest to centre 1 and the other observations are closest to centre 2.
- Since there has been no change in the clusters k-means stops at this point with final cluster assignments of $\{-3, -4\}$ and $\{2, 3, 5\}$ and centres of -3.5 and 3.33.

ii. Same as above but now the observation 1 is not allocated to any cluster. k-means stops at this point with final cluster assignments of $\{-3, -4\}$ and $\{3, 5\}$ and centres of -3.5 and 4.

Question 4

(a) **Suppose that we have five observed points, each with four features. We present the Euclidean distance between any two observations with measurements on these four features in the following matrix.**

	1	**2**	**3**	**4**	**5**
1	**0.00**	**0.90**	**0.16**	**0.45**	**0.60**
2	**0.90**	**0.00**	**0.55**	**0.50**	**0.04**
3	**0.16**	**0.55**	**0.00**	**0.57**	**0.35**
4	**0.45**	**0.50**	**0.57**	**0.00**	**0.30**
5	**0.60**	**0.04**	**0.35**	**0.30**	**0.00**

Use the matrix with Euclidean distances to perform hierarchical clustering, using simple linkage.

(13 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.* question.

Approaching the question

The steps of the hierararchical clustering method to obtain the requested dendrogram are listed below:

1. In the beginning all 5 points form their own cluster.
2. Next, we need to calculate the minimum distances between the pairs of each cluster. The minimum distance is 0.04 so we first group $(2, 5)$ together.
3. Then, we recompute the above matrix with $(2, 5)$ as one cluster. The new elements are coming from the 'distances':

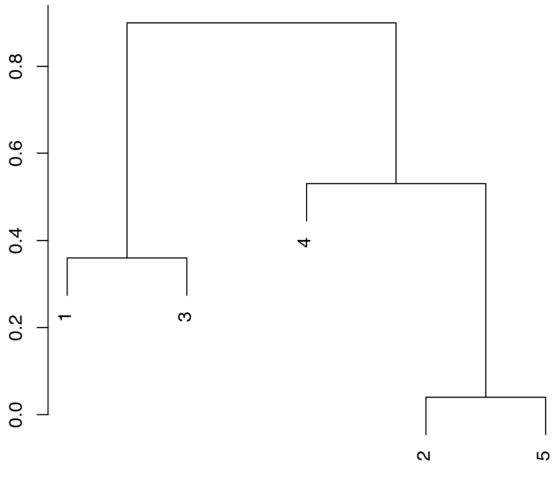
$$D[(2, 5), 1] = \min[D(2, 1), D(5, 1)] = \min[0.90, 0.60] = 0.60$$

$$D[(2, 5), 3] = \min[D(2, 3), D(5, 3)] = \min[0.55, 0.35] = 0.35$$

$$D[(2, 5), 4] = \min[D(2, 4), D(5, 4)] = \min[0.50, 0.30] = 0.30$$

   and the distances $D(1, 3) = 0.16$, $D(1, 4) = 0.45$ and $D(3, 4) = 0.57$. The smallest distance is 0.16, so we group $(1, 3)$ together.
4. Now we have the clusters $(1, 3)$, 4 and $(2, 5)$. We have $D[(2, 5), 4] = 0.30$ from the previous step and calculate $D[(1, 3), 4] = \min[0.45, 0.57] = 0.45$ and $D[(1, 3), (2, 5)] = \min[(0.9 + 0.6 + 0.55 + 0.35] = 0.35$, so we merge 4 and $(2, 5)$ together.
5. The final step just merges together the clusters $(1, 3)$ and $(2, 4, 5)$.

All of these are summarised in the figure below.



(b) **Assume that we take a data set, divide it into equally-sized training and test sets, and then try out two different classification procedures. First we use linear discriminant analysis and get an error rate of 20% on the training data and 15% on the test data. Next we use 1-nearest neighbours (i.e. $k = 1$) and get an average error rate (averaged over both test and training data sets) of 10%. Based on these results, which method should we prefer to use for classification of new observations? Why?**

(6 marks)

Reading for this question

This question examines the main ideas of k-nearest neighbours, that can be found in Chapter 8 and Section 2.2 of the James *et al.* textbook, respectively. Also the content of train and test error is quite relevant; see, for example, Section 5.1.

Approaching the question

For k-nearest neighbours with $k = 1$, the training error rate is 0% because for any training observation, its nearest neighbour will be the response itself. So, k-nearest neighbours has a test error rate of 20%. We would choose linear discriminant analysis because of its lower test error rate of 15%.

(c) **Consider the following binary classification problem with $Y = k$, $k \in \{1, 2\}$. At a data point x, $P(Y = 1 \mid X = x) = 0.4$. Let x' be the nearest neighbour of x and $P(Y = 1 \mid X = x') = p > 0$. What are the values of p such that the 1-neighbour error at x is at least 0.5?**

(6 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:

$$\begin{aligned} &P(Y = 1 \mid X = x)\, P(Y = 2 \mid X = x') + P(Y = 1 \mid X = x')\, P(Y = 2 \mid X = x) \\ &\quad = 0.4 \times (1 - p) + p \times (1 - 0.4) \\ &\quad = 0.4 + 0.2p \\ &\quad \geq 0.5 \end{aligned}$$

which implies that $p \geq 0.5$.