Queen Mary University of London

Main Examination period 2020 – May – Semester B

MTH6101: Introduction to Machine Learning

Duration: 2 hours

Student number | | | | | | | | | **Desk number** | | |

Make and model of calculator used ____________________

Apart from this page, you are not permitted to read the contents of this question paper until instructed to do so by an invigilator.

Write your solutions in the spaces provided in this exam paper. If you need more paper, ask an invigilator for an additional booklet and attach it to this paper at the end of the exam.

You should attempt ALL questions. Marks available are shown next to the questions.

Only non-programmable calculators that have been approved from the college list of non-programmable calculators are permitted in this examination. Please state on your answer book the name and type of machine used.

Complete all rough work in the answer book and cross through any work that is not to be assessed.

Possession of unauthorised material at any time when under examination conditions is an assessment offence and can lead to expulsion from QMUL. Check now to ensure you do not have any unauthorised notes, mobile phones, smartwatches or unauthorised electronic devices on your person. If you do, raise your hand and give them to an invigilator immediately.

It is also an offence to have any writing of any kind on your person, including on your body. If you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms, it will be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. A mobile phone that causes a disruption in the exam is also an assessment offence.

Exam papers must not be removed from the examination room.

Examiners: H. Maruri-Aguilar, M. Benning

This page is for marking purposes only.

Do not write on it.

Question	Mark	Comments
1	/ 10	
2	/ 29	
3	/ 20	
4	/ 23	
5	/ 18	
Total		

Question 1 [10 marks].

(a) Describe the problem of dimensionality reduction in unsupervised learning. **[6]**

(b) List two techniques for this problem. **[4]**

Write your solutions here

Question 2 [29 marks]. As part of **Karhunen-Loeve** expansion of the covariance matrix of a centered data set $\mathbf{X}$ with $n = 100$ observations in $p = 5$ variables, the following matrix was computed.

$$\boldsymbol{\Lambda} = \begin{pmatrix} 75.93 & 0 & 0 & 0 & 0 \\ 0 & 68.544 & 0 & 0 & 0 \\ 0 & 0 & 38.767 & 0 & 0 \\ 0 & 0 & 0 & 26.228 & 0 \\ 0 & 0 & 0 & 0 & 4.746 \end{pmatrix}$$

(a) Complete the following table and determine a number of components using an 80% threshold. **[12]**

Write your solutions here

	Standard deviation	Proportion of variance	Cumulative proportion
PC1			
PC2			
PC3			
PC4			
PC5			

(b) Using the matrix $\mathbf{\Lambda}$ above, determine if the data was scaled to compute the covariance matrix and briefly explain why. **[4]**

(c) Write (do not derive) the formula that links $\mathbf{\Lambda}$ with $\mathbf{D}$. Recall that $\mathbf{\Lambda}$ is the eigenvalue matrix of the Karhunen-Loeve decomposition of the covariance matrix $\mathbf{\Sigma}$; and that $\mathbf{D}$ is the matrix of eigenvalues of the singular value decomposition of matrix $\mathbf{X}$. **[6]**

Write your solutions here

(d) Use the formula you wrote to determine numerically the eigenvalues d_i of the singular value decomposition of the data matrix $\mathbf{X}$. [7]

Write your solutions here

Question 3 [20 marks].

(a) Explain what is meant by **single** linkage in agglomerative clustering. **[3]**

(b) Consider the following distance matrix

$$\begin{array}{c c} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \left(\begin{array}{rrrrr} 0 & 7 & 4 & 11 & 10 \\ 7 & 0 & 11 & 10 & 11 \\ 4 & 11 & 0 & 15 & 14 \\ 11 & 10 & 15 & 0 & 1 \\ 10 & 11 & 14 & 1 & 0 \end{array}\right) \end{array},$$

where row and columns are indexed as usual by individuals.

(i) If agglomerative **single** linkage clustering were to be performed, which individuals would be merged first and why? **[4]**

Write your solutions here

(ii) Explain why in the first step the result is the same regardless of the linkage used. **[3]**

(iii) Assume you are at a step in agglomerative clustering in which individuals 1,2,3 belong to one cluster and individuals 4,5 belong to another cluster. Using **single** linkage, find the distance between these two clusters. **[5]**

Write your solutions here

(iv) Using **average** linkage, give the distance between clusters in Question (biii). **[5]**

Write your solutions here

Question 4 [23 marks]. The following data are the results of a classification analysis. The output includes the validation output `Ytrue` and the classifications obtained with three trained classification algorithms termed `Y1`, `Y2` and `Y3`.

```
##       Ytrue Y1 Y2 Y3
##  [1,]     1  1  0  1
##  [2,]     0  0  1  0
##  [3,]     1  1  0  0
##  [4,]     0  0  1  1
##  [5,]     0  1  1  0
##  [6,]     0  0  1  1
##  [7,]     0  0  0  0
##  [8,]     0  0  1  0
##  [9,]     0  0  0  0
## [10,]     1  1  0  0
## [11,]     1  1  0  1
## [12,]     1  1  0  0
```

(a) Complete the following confusion matrices. **[9]**

Write your solutions here

		Predicted (`Y1`)	
		0	1
True (`Ytrue`)	0		
	1		

		Predicted (`Y2`)	
		0	1
True (`Ytrue`)	0		
	1		

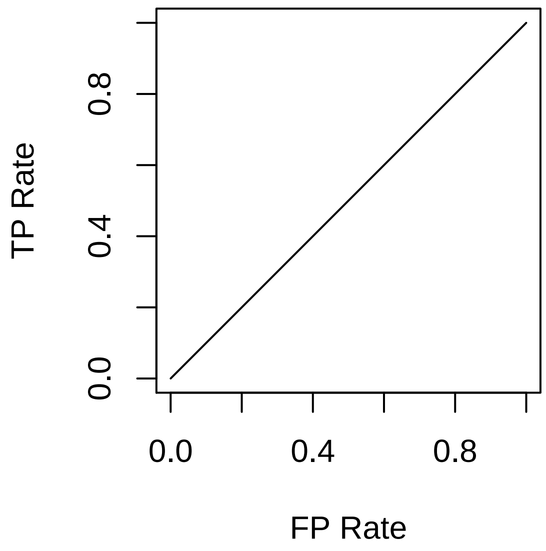
		Predicted (`Y3`)	
		0	1
True (`Ytrue`)	0		
	1		

(b) Compute the False Positive Rate (FPR) and True Positive Rate (TPR) for each confusion matrix, completing in the table below. **[6]**

(c) Plot your results in the ROC graph below and briefly comment on the performance of classifiers. Which is the best classifier? **[8]**

Write your solutions here

Confusion matrix	FPR	TPR
`Y1`		
`Y2`		
`Y3`		



Question 5 [18 marks].

(a) The Lasso criterion is $L = \frac{1}{2}||\mathbf{Y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_1$. Explain what the components of the Lasso criterion are. **[3]**

(b) Explain what are the solutions to lasso as $\lambda \to 0$. Also as $\lambda \to \infty$. **[2]**

Write your solutions here

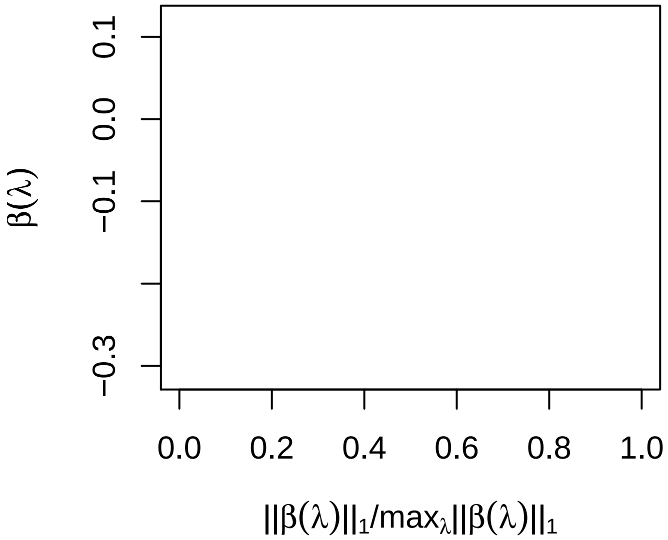
(c) The following table contains output from a lasso fit to model with $d = 3$ variables and $n = 20$ observations. For each row in the table, compute s, the proportion of shrinkage defined as $s = s(\lambda) = ||\beta(\lambda)||_1 / \max_\lambda ||\beta(\lambda)||_1$ and write its value in the correct position to complete the table. **[6]**

Write your solutions here

λ	β_1	β_2	β_3	s
0	0.12057	-0.31144	-0.07388	
1.16364	0.01818	-0.23636	0	
1.41935	0	-0.21774	0	
5.6	0	0	0	

(d) Using your completed information, add the lasso paths to the following plot. In your plot, label each path according to its corresponding variable. **[7]**

Write your solutions here



Extra space for calculations

End of Paper – An appendix of 2 pages follows.

Matrices and their decompositions for data handling

Data set: $\mathbf{X}$
Sample covariance matrix of a centered data set: $\mathbf{\Sigma} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$.
Karhunen-Loeve decomposition of the covariance matrix: $\mathbf{\Sigma} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T$.
The total (sample) variance is $\mathrm{tr}(\mathbf{\Sigma}) = \sum_{i=1}^{p} \lambda_i$
Singular value decomposition of $\mathbf{X}$: $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$.

Object	Size	Notes
n	-	Number of observations
p	-	Number of variables
$\mathbf{X}$	$n \times p$	$n > p$
$\mathbf{\Sigma}$	$p \times p$	(i, j) element is the sample covariance between columns i, j of $\mathbf{X}$
$\mathbf{A}$	$p \times p$	Eigenvectors $\mathbf{a}_i$ are columns; $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}$
$\mathbf{\Lambda}$	$p \times p$	Diagonal with eigenvalues λ_i of $\mathbf{\Sigma}$
$\mathbf{U}$	$n \times p$	Eigenvectors $\mathbf{u}_i$ are columns; $\mathbf{U}^T\mathbf{U} = \mathbf{I}$
$\mathbf{D}$	$p \times p$	Diagonal with singular eigenvalues d_i of $\mathbf{X}$
$\mathbf{V}$	$p \times p$	Eigenvectors $\mathbf{v}_i$ are columns; $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$

As in lectures, it is assumed that the rank of $\mathbf{X}$ is p so that $\mathbf{X}^T\mathbf{X}$ is invertible; that the diagonal entries in either $\mathbf{\Lambda}$ or $\mathbf{D}$ are all distinct and positive numbers.
For a square full rank diagonal matrix $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_m)$, we have $\mathbf{W} = \mathbf{W}^T$; that $\mathbf{W}^T\mathbf{W} = \mathbf{W}\mathbf{W}^T = \mathbf{W}^2 = \mathrm{diag}(w_1^2, \ldots, w_m^2)$ and $\mathbf{W}^{-1} = \mathrm{diag}(1/w_1, \ldots, 1/w_m)$.

Common distances d_{ij} between points $\mathbf{x}_i$ and $\mathbf{x}_j$

Name	Distance d_{ij}	Equivalent notation
Manhattan	$\sum_{l=1}^{p} \lvert x_{il} - x_{jl} \rvert$	$\lvert\lvert\mathbf{x}_i - \mathbf{x}_j\rvert\rvert_1$
Euclidean	$\sqrt{\sum_{l=1}^{p} \left(x_{il} - x_{jl}\right)^2}$	$\lvert\lvert\mathbf{x}_i - \mathbf{x}_j\rvert\rvert_2 = \sqrt{\lvert\lvert\mathbf{x}_i - \mathbf{x}_j\rvert\rvert_2^2}$
Minkowski	$\left(\sum_{l=1}^{p} \lvert x_{il} - x_{jl} \rvert^m\right)^{1/m}$	$\lvert\lvert\mathbf{x}_i - \mathbf{x}_j\rvert\rvert_m$
Mahalanobis	$\sqrt{\left(\mathbf{x}_{i\cdot} - \mathbf{x}_{j\cdot}\right)^T \mathbf{\Sigma}^{-1} \left(\mathbf{x}_{i\cdot} - \mathbf{x}_{j\cdot}\right)}$	
Absolute correlation	$\sqrt{1 - \lvert\rho_{ij}\rvert}$	

The Manhattan norm of a vector $\mathbf{x}$ is $||\mathbf{x}||_1 = \sum_{i=1}^{p} |x_i|$ (Manhattan distance between $\mathbf{x}$ and the origin); and the Euclidean norm of $\mathbf{x}$ is $||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{p} x_i^2}$ (Euclidean distance between $\mathbf{x}$ and the origin).

Classification

Name	Comment
TP	True positive
FP	False positive
FN	False negative
TN	True negative
$P = TP + FN$	Positives
$N = FP + TN$	Negatives
$TPR = \frac{TP}{P}$	True positive rate (Sensitivity or Recall)
$FPR = \frac{FP}{N}$	False positive rate
$FNR = \frac{FN}{P}$	False negative rate
$TNR = \frac{TN}{N}$	True negative rate (Specificity)
$\frac{TP}{TP+FP}$	Precision
$2\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$	F1 score
$\frac{TP+TN}{P+N}$	Accuracy
$\frac{FP+FN}{P+N}$	Error rate

Penalised regression

Lasso criterion:

$$L = \frac{1}{2}||\mathbf{Y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_1$$

Ridge regression criterion:

$$R = ||\mathbf{Y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_2^2.$$

End of Appendix.

Main Examination period 2020 – May – Semester B

MTH6101: Introduction to Machine Learning

Duration: 2 hours

Student number | | | | | | | | | **Desk number** | | |

Make and model of calculator used ______________________________

Apart from this page, you are not permitted to read the contents of this question paper until instructed to do so by an invigilator.

Write your solutions in the spaces provided in this exam paper. If you need more paper, ask an invigilator for an additional booklet and attach it to this paper at the end of the exam.

You should attempt ALL questions. Marks available are shown next to the questions.

Only non-programmable calculators that have been approved from the college list of non-programmable calculators are permitted in this examination. Please state on your answer book the name and type of machine used.

Complete all rough work in the answer book and cross through any work that is not to be assessed.

Possession of unauthorised material at any time when under examination conditions is an assessment offence and can lead to expulsion from QMUL. Check now to ensure you do not have any unauthorised notes, mobile phones, smartwatches or unauthorised electronic devices on your person. If you do, raise your hand and give them to an invigilator immediately.

It is also an offence to have any writing of any kind on your person, including on your body. If you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms, it will be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. A mobile phone that causes a disruption in the exam is also an assessment offence.

Exam papers must not be removed from the examination room.

Examiners: H. Maruri-Aguilar, M. Benning

This page is for marking purposes only.

Do not write on it.

Question	Mark	Comments
1	/ 10	
2	/ 29	
3	/ 20	
4	/ 23	
5	/ 18	
Total		

Question 1 [10 marks].

(a) Describe the problem of dimensionality reduction in unsupervised learning. **[6]**

(b) List two techniques for this problem. **[4]**

Write your solutions here

(a) In the dimensionality reduction problem the data has a high number of dimensions [2] and data needs to be mapped into low dimensions [2] while still preserving relevant information [2].

Total (a) [6]

(b) Examples of techniques for this problem are Principal Component Analysis [2], Factor Analysis [2] and Multidimensional Scaling [2].

Only two examples are needed. Total (b) [4]

Total [10]

Seen in lectures, bookwork.

Question 2 [29 marks]. As part of **Karhunen-Loeve** expansion of the covariance matrix of a centered data set $\mathbf{X}$ with $n = 100$ observations in $p = 5$ variables, the following matrix was computed.

$$\mathbf{\Lambda} = \begin{pmatrix} 75.93 & 0 & 0 & 0 & 0 \\ 0 & 68.544 & 0 & 0 & 0 \\ 0 & 0 & 38.767 & 0 & 0 \\ 0 & 0 & 0 & 26.228 & 0 \\ 0 & 0 & 0 & 0 & 4.746 \end{pmatrix}$$

(a) Complete the following table and determine a number of components using an 80% threshold. **[12]**

Write your solutions here

	Standard deviation	Proportion of variance	Cumulative proportion
PC1	8.7138	0.3545	0.3545
PC2	8.2791	0.32	0.6744
PC3	6.2263	0.181	0.8554
PC4	5.1213	0.1224	0.9778
PC5	2.1785	0.0222	1

Looking at the column of 'Cumulative proportion' in the table above, we suggest 3 components which in this case explain 85.54% of the total variability.

Computations use eigenvalues, take square root; convert them to proportions then to cumulative proportions. Each column [3] to total for Table [9].
Number of components [3].

Total (a) [12]

Seen in lectures, coursework.

(b) Using the matrix $\mathbf{\Lambda}$ above, determine if the data was scaled to compute the covariance matrix and briefly explain why. **[4]**

(c) Write (do not derive) the formula that links $\mathbf{\Lambda}$ with $\mathbf{D}$. Recall that $\mathbf{\Lambda}$ is the eigenvalue matrix of the Karhunen-Loeve decomposition of the covariance matrix $\mathbf{\Sigma}$; and that $\mathbf{D}$ is the matrix of eigenvalues of the singular value decomposition of matrix $\mathbf{X}$. **[6]**

Write your solutions here

(b) We simply compute the trace of this matrix, which is $\sum_{i=1}^{5} \lambda_i = 214.215$. As the sum is not equal to p, then the data was not scaled.

Comment [2], brief explanation [2]

Total (b) [4]

(c) The formula that links eigenvalue matrices is

$$\mathbf{\Lambda} = \frac{1}{n-1}\mathbf{D}^2.$$

Formula [6]

Total(c) [6]

Both items seen in lectures.

(d) Use the formula you wrote to determine numerically the eigenvalues d_i of the singular value decomposition of the data matrix $\mathbf{X}$. **[7]**

Write your solutions here

The previous formula $\mathbf{\Lambda} = \frac{1}{n-1}\mathbf{D}^2$ links individual eigenvalues λ_i from K-L to those d_i of svd as $\lambda_i = d_i^2/(n-1)$ so that $d_i = \sqrt{(n-1)\lambda_i}$ [2]. We apply the formula directly so that

$d_1 = \sqrt{(100-1)75.93} = \sqrt{7517.07} = 86.701$;
$d_2 = \sqrt{(100-1)68.544} = \sqrt{6785.856} = 82.376$;
$d_3 = \sqrt{(100-1)38.767} = \sqrt{3837.933} = 61.951$;
$d_4 = \sqrt{(100-1)26.228} = \sqrt{2596.572} = 50.957$ and
$d_5 = \sqrt{(100-1)4.746} = \sqrt{469.854} = 21.676$.

Allow full marks if formula was -implicitly- used correctly then each eigenvalue [1] to total [5].

Total (d) [7]

Seen in lectures and coursework.

Question 3 [20 marks].

(a) Explain what is meant by **single** linkage in agglomerative clustering. **[3]**

(b) Consider the following distance matrix

$$\begin{array}{c c} & \begin{array}{ccccc} 1 & 2 & 3 & 4 & 5 \end{array} \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{array} & \begin{pmatrix} 0 & 7 & 4 & 11 & 10 \\ 7 & 0 & 11 & 10 & 11 \\ 4 & 11 & 0 & 15 & 14 \\ 11 & 10 & 15 & 0 & 1 \\ 10 & 11 & 14 & 1 & 0 \end{pmatrix} \end{array},$$

where row and columns are indexed as usual by individuals.

(i) If agglomerative **single** linkage clustering were to be performed, which individuals would be merged first and why? **[4]**

Write your solutions here

(a) In single linkage, the distance between two clusters [1] is defined as the distance between the two closest [2] elements of each cluster. This is the "nearest neighbor" distance.

Total (a) [3]

(bi) At this point, the smallest distance between clusters in the table is 1 which is that between clusters '4' and '5' and thus these two clusters are joined [4].

Total (bi) [4]

Seen in lectures and coursework.

(ii) Explain why in the first step the result is the same regardless of the linkage used. **[3]**

(iii) Assume you are at a step in agglomerative clustering in which individuals 1,2,3 belong to one cluster and individuals 4,5 belong to another cluster. Using **single** linkage, find the distance between these two clusters. **[5]**

Write your solutions here

(bii) At the start, each individual is a cluster so the distance between clusters is a single entry in the distance matrix [1]. The minimum, maximum and average coincide in this case so there is no difference between linkages for this first step [2].

Total (bii) [3]

The distances involved are a subset of the distance matrix, indeed those in the intersection between rows 1,2,3 and columns 4,5. These distances are 11, 10, 15, 10, 11, 14 [2]. The single linkage distance is the minimum of those distances so the distance asked is 10 [3].

Allow full marks if correct figure.

Total (biii) [5]

Seen in lectures, coursework.

(iv) Using **average** linkage, give the distance between clusters in Question (biii). **[5]**

Write your solutions here

(biv) Here the set of relevant distances is the same as in the previous question: 11, 10, 15, 10, 11, 14 [2]. The average linkage uses the average of these values so the distance asked is 11.8333 [3].

Allow full marks if correct figure.

Total (biv) [5]

Seen in lectures, coursework.

Question 4 [23 marks]. The following data are the results of a classification analysis. The output includes the validation output `Ytrue` and the classifications obtained with three trained classification algorithms termed `Y1`, `Y2` and `Y3`.

```
##       Ytrue Y1 Y2 Y3
##  [1,]     1  1  0  1
##  [2,]     0  0  1  0
##  [3,]     1  1  0  0
##  [4,]     0  0  1  1
##  [5,]     0  1  1  0
##  [6,]     0  0  1  1
##  [7,]     0  0  0  0
##  [8,]     0  0  1  0
##  [9,]     0  0  0  0
## [10,]     1  1  0  0
## [11,]     1  1  0  1
## [12,]     1  1  0  0
```

(a) Complete the following confusion matrices. **[9]**

Write your solutions here

		Predicted (`Y1`) 0	1
True (`Ytrue`)	0	6	1
	1	0	5

		Predicted (`Y2`) 0	1
True (`Ytrue`)	0	2	5
	1	5	0

		Predicted (`Y3`) 0	1
True (`Ytrue`)	0	5	2
	1	3	2

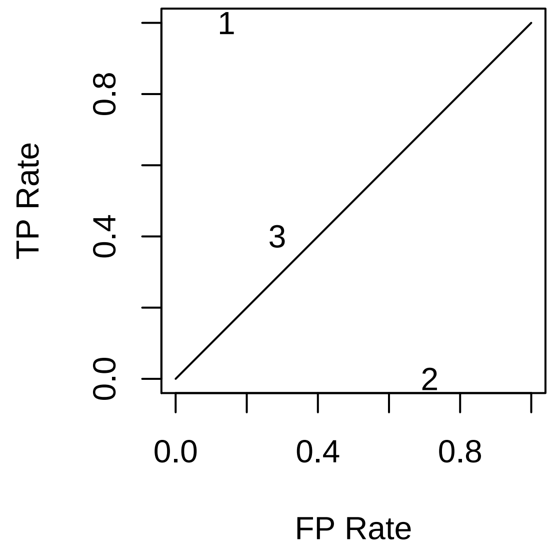
Each matrix [3] for Total (a) [9]
Seen in lectures, coursework.

(b) Compute the False Positive Rate (FPR) and True Positive Rate (TPR) for each confusion matrix, completing in the table below. **[6]**

(c) Plot your results in the ROC graph below and briefly comment on the performance of classifiers. Which is the best classifier? **[8]**

Write your solutions here

Confusion matrix	FPR	TPR
`Y1`	0.14286	1
`Y2`	0.71429	0
`Y3`	0.28571	0.4



The best classifier is `Y1`, followed by `Y3` whose performance is of a random classifier. The classifier `Y2` performs the worst.

In table, each entry [1] so Total (b) [6]

Each point in the plot [2] so for the plot [6]. Best classifier or meaningful comment [2] thus Total (c) [8]

Seen in lectures and coursework.

Question 5 [18 marks].

(a) The Lasso criterion is $L = \frac{1}{2}||\mathbf{Y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_1$. Explain what the components of the Lasso criterion are. **[3]**

(b) Explain what are the solutions to lasso as $\lambda \to 0$. Also as $\lambda \to \infty$. **[2]**

Write your solutions here

(a) The criterion has the term $||\mathbf{Y} - \mathbf{X}\beta||_2^2$ which is the usual sum of squares of the error from standard regression [1].
The term $||\beta||_1$ is the absolute size of coefficient vector [1] and the parameter λ is a non-negative quantity that controls the amount of penalization [1].

Each item [1] to Total (a) [3]

(b) When $\lambda \to 0$, the solution of Lasso is the ordinary least squares estimator [1].
As $\lambda \to \infty$, the solution of Lasso is $\beta = \mathbf{0}$, that is shrinkage of all coefficients to zero [1].

Each item [1] to Total (b) [2]

Both items seen in lectures

(c) The following table contains output from a lasso fit to model with $d = 3$ variables and $n = 20$ observations. For each row in the table, compute s, the proportion of shrinkage defined as $s = s(\lambda) = ||\beta(\lambda)||_1 / \max_\lambda ||\beta(\lambda)||_1$ and write its value in the correct position to complete the table. **[6]**

Write your solutions here

λ	β_1	β_2	β_3	s
0	0.12057	-0.31144	-0.07388	1
1.16364	0.01818	-0.23636	0	0.50315
1.41935	0	-0.21774	0	0.4304
5.6	0	0	0	0

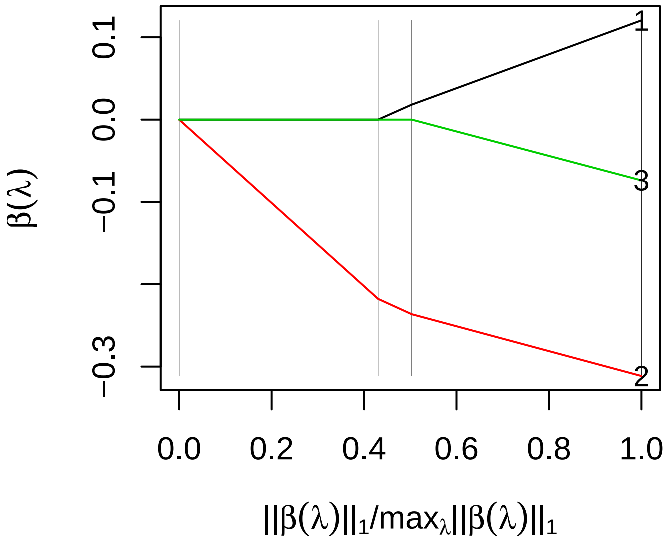
Computation of s involve the sums of absolute coefficients 0.5059, 0.25455, 0.21774, 0 which are then divided by the maximum 0.5059 to give the values in the table.

Each correct value of s [1] and if all correct add [2] to Total (c) [6].

Seen in lectures and coursework.

(d) Using your completed information, add the lasso paths to the following plot. In your plot, label each path according to its corresponding variable. [7]

Write your solutions here



Each correct path [2] and if all correct add [1] to Total (d) [7].

Seen in lectures and coursework.

Extra space for calculations

End of Paper – An appendix of 2 pages follows.

Matrices and their decompositions for data handling

Data set: $\mathbf{X}$
Sample covariance matrix of a centered data set: $\mathbf{\Sigma} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$.
Karhunen-Loeve decomposition of the covariance matrix: $\mathbf{\Sigma} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T$.
The total (sample) variance is $\mathrm{tr}(\mathbf{\Sigma}) = \sum_{i=1}^{p} \lambda_i$
Singular value decomposition of $\mathbf{X}$: $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$.

Object	Size	Notes
n	-	Number of observations
p	-	Number of variables
$\mathbf{X}$	$n \times p$	$n > p$
$\mathbf{\Sigma}$	$p \times p$	(i,j) element is the sample covariance between columns i, j of $\mathbf{X}$
$\mathbf{A}$	$p \times p$	Eigenvectors $\mathbf{a}_i$ are columns; $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}$
$\mathbf{\Lambda}$	$p \times p$	Diagonal with eigenvalues λ_i of $\mathbf{\Sigma}$
$\mathbf{U}$	$n \times p$	Eigenvectors $\mathbf{u}_i$ are columns; $\mathbf{U}^T\mathbf{U} = \mathbf{I}$
$\mathbf{D}$	$p \times p$	Diagonal with singular eigenvalues d_i of $\mathbf{X}$
$\mathbf{V}$	$p \times p$	Eigenvectors $\mathbf{v}_i$ are columns; $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$

As in lectures, it is assumed that the rank of $\mathbf{X}$ is p so that $\mathbf{X}^T\mathbf{X}$ is invertible; that the diagonal entries in either $\mathbf{\Lambda}$ or $\mathbf{D}$ are all distinct and positive numbers.
For a square full rank diagonal matrix $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_m)$, we have $\mathbf{W} = \mathbf{W}^T$; that $\mathbf{W}^T\mathbf{W} = \mathbf{W}\mathbf{W}^T = \mathbf{W}^2 = \mathrm{diag}(w_1^2, \ldots, w_m^2)$ and $\mathbf{W}^{-1} = \mathrm{diag}(1/w_1, \ldots, 1/w_m)$.

Common distances d_{ij} between points $\mathbf{x}_i$ and $\mathbf{x}_j$

Name	Distance d_{ij}	Equivalent notation
Manhattan	$\sum_{l=1}^{p} \lvert x_{il} - x_{jl} \rvert$	$\lvert\lvert \mathbf{x}_i - \mathbf{x}_j \rvert\rvert_1$
Euclidean	$\sqrt{\sum_{l=1}^{p} \left(x_{il} - x_{jl}\right)^2}$	$\lvert\lvert \mathbf{x}_i - \mathbf{x}_j \rvert\rvert_2 = \sqrt{\lvert\lvert \mathbf{x}_i - \mathbf{x}_j \rvert\rvert_2^2}$
Minkowski	$\left(\sum_{l=1}^{p} \lvert x_{il} - x_{jl} \rvert^m\right)^{1/m}$	$\lvert\lvert \mathbf{x}_i - \mathbf{x}_j \rvert\rvert_m$
Mahalanobis	$\sqrt{\left(\mathbf{x}_{i\cdot} - \mathbf{x}_{j\cdot}\right)^T \mathbf{\Sigma}^{-1} \left(\mathbf{x}_{i\cdot} - \mathbf{x}_{j\cdot}\right)}$	
Absolute correlation	$\sqrt{1 - \lvert \rho_{ij} \rvert}$	

The Manhattan norm of a vector $\mathbf{x}$ is $||\mathbf{x}||_1 = \sum_{i=1}^{p} |x_i|$ (Manhattan distance between $\mathbf{x}$ and the origin); and the Euclidean norm of $\mathbf{x}$ is $||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{p} x_i^2}$ (Euclidean distance between $\mathbf{x}$ and the origin).

Classification

Name	Comment
TP	True positive
FP	False positive
FN	False negative
TN	True negative
$P = TP + FN$	Positives
$N = FP + TN$	Negatives
$TPR = \frac{TP}{P}$	True positive rate (Sensitivity or Recall)
$FPR = \frac{FP}{N}$	False positive rate
$FNR = \frac{FN}{P}$	False negative rate
$TNR = \frac{TN}{N}$	True negative rate (Specificity)
$\frac{TP}{TP+FP}$	Precision
$2\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$	F1 score
$\frac{TP+TN}{P+N}$	Accuracy
$\frac{FP+FN}{P+N}$	Error rate

Penalised regression

Lasso criterion:

$$L = \frac{1}{2}||\mathbf{Y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_1$$

Ridge regression criterion:

$$R = ||\mathbf{Y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_2^2.$$

End of Appendix.

Queen Mary University of London

Main Examination period 2020 – May – Semester B

MTH6101: Introduction to Machine Learning - SAMPLE

Duration: 2 hours

Student number | | | | | | | | | **Desk number** | | |

Make and model of calculator used ______________________________

Apart from this page, you are not permitted to read the contents of this question paper until instructed to do so by an invigilator.

Write your solutions in the spaces provided in this exam paper. If you need more paper, ask an invigilator for an additional booklet and attach it to this paper at the end of the exam.

You should attempt ALL questions. Marks available are shown next to the questions.

Only non-programmable calculators that have been approved from the college list of non-programmable calculators are permitted in this examination. Please state on your answer book the name and type of machine used.

Complete all rough work in the answer book and cross through any work that is not to be assessed.

Possession of unauthorised material at any time when under examination conditions is an assessment offence and can lead to expulsion from QMUL. Check now to ensure you do not have any unauthorised notes, mobile phones, smartwatches or unauthorised electronic devices on your person. If you do, raise your hand and give them to an invigilator immediately.

It is also an offence to have any writing of any kind on your person, including on your body. If you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms, it will be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. A mobile phone that causes a disruption in the exam is also an assessment offence.

Exam papers must not be removed from the examination room.

Examiners: H. Maruri-Aguilar, M. Benning

This page is for marking purposes only.

Do not write on it.

Question	Mark	Comments
1	/ 29	
2	/ 24	
3	/ 28	
4	/ 19	
Total		

Question 1 [29 marks].
A data set consists of measurements of amounts of chemical compounds found in material samples of a mineral processing facility. Seven measurements were available for each sample of a total of $n = 200$ samples. The units of the measurements are milligrammes.
From a Principal Component Analysis of these data, the following are the PC loadings.

```
##          PC1      PC2      PC3      PC4      PC5      PC6      PC7
## Na  0.48335 -0.00822  0.11289 -0.22665  0.36347 -0.09448 -0.74910
## Au  0.37907  0.07811 -0.86478  0.31936 -0.01328  0.01060  0.00901
## NO3 0.46602 -0.00584  0.04666 -0.45909 -0.75041  0.00548  0.08190
## Al  0.00564 -0.70429 -0.06960 -0.00191  0.01109 -0.69995  0.09513
## Cu  0.48040 -0.00519  0.12123 -0.24015  0.52372  0.09342  0.64328
## Mg  0.41794 -0.03411  0.46375  0.76035 -0.17366 -0.00960  0.02684
## Zn  0.00311 -0.70469 -0.05134  0.00872 -0.00604  0.70156 -0.09205
```

The corresponding eigenvalues of the Karhunen-Loeve decomposition of the covariance matrix of these data are 4.11178, 1.98865, 0.49036, 0.29693, 0.07676, 0.02234, 0.01318.

(a) Each of the data columns was centered around its mean and standardised to have unit variance. From the output you have been given, briefly explain how we can confirm that the data was indeed standardized by column. **[4]**

Write your solutions here

(b) Complete the following table. **[15]**

Write your solutions here

	Standard deviation	Proportion of variance	Cumulative proportion
PC1			
PC2			
PC3			
PC4			
PC5			
PC6			
PC7			

(c) Using the table you just computed, suggest a number of components so that at least 80% of the total variability is explained. **[4]**

(d) Interpret the components you suggested. **[6]**

Write your solutions here

Question 2 [24 marks].

(a) Explain what is meant by **single** linkage in agglomerative clustering. **[3]**

(b) Consider the following data matrix

$$\mathbf{X} = \begin{pmatrix} 2 & -1 & -3 & 3 \\ 1 & 1 & 1 & 3 \\ 3 & -4 & -3 & -1 \\ -4 & 3 & -2 & -3 \\ -4 & 3 & -3 & 2 \end{pmatrix},$$

(i) Using Manhattan distance, complete the missing entries in the distance matrix below. **[4]**

Write your solutions here

$$\text{b(i):} \begin{pmatrix} 0 & \boxed{} & 8 & 17 & 11 \\ \boxed{} & 0 & 15 & \boxed{} & 12 \\ 8 & 15 & 0 & 17 & 17 \\ 17 & \boxed{} & 17 & 0 & \boxed{} \\ 11 & 12 & 17 & \boxed{} & 0 \end{pmatrix},$$

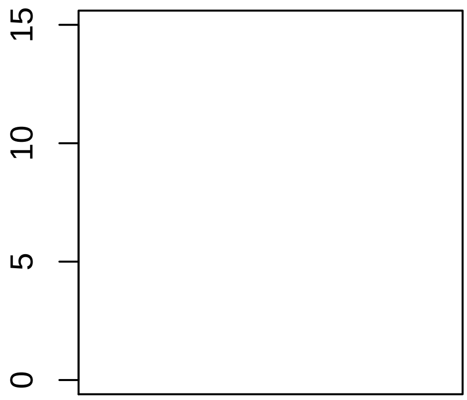
(ii) Perform agglomerative cluster for these data using the distance matrix above and single linkage. As part of your procedure, compute and show updated distance matrices at every step of the clustering procedure. **[12]**

Write your solutions here

Write your solutions here

(iii) Complete the dendrogram in the plot given below. **[5]**

Write your solutions here



Question 3 [28 marks]. The following data are the validation results of a classification analysis. The output below includes the validation output `Ytrue` and the classifications obtained with four classification algorithms termed `Y1`, `Y2`, `Y3` and `Y4`.

```
##       Ytrue Y1 Y2 Y3 Y4
##  [1,]     1  1  0  1  0
##  [2,]     0  0  0  0  0
##  [3,]     0  0  1  0  1
##  [4,]     1  1  0  1  0
##  [5,]     0  0  1  0  1
##  [6,]     0  0  1  1  1
##  [7,]     0  0  1  0  1
##  [8,]     1  1  0  1  1
##  [9,]     0  1  1  0  0
## [10,]     0  0  1  1  0
```

(a) Complete the following confusion matrices. **[14]**

Write your solutions here

		Predicted (`Y1`)	
		0	1
True (`Ytrue`)	0		
	1		

		Predicted (`Y2`)	
		0	1
True (`Ytrue`)	0		
	1		

		Predicted (`Y3`)	
		0	1
True (`Ytrue`)	0		
	1		

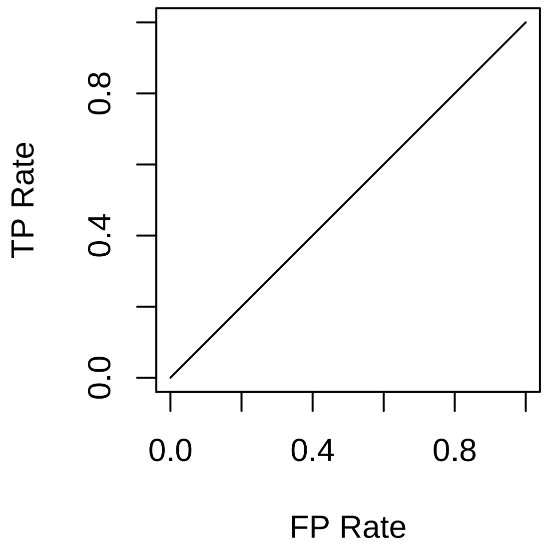
		Predicted (`Y3`)	
		0	1
True (`Ytrue`)	0		
	1		

(b) Compute the False Positive Rate (FPR) and True Positive Rate (TPR) for each confusion matrix, completing in the table below. **[6]**

(c) Plot your results in the ROC graph below and briefly comment on the performance of classifiers. Which is the best classifier? **[8]**

Write your solutions here

Confusion matrix	FPR	TPR
Y1		
Y2		
Y3		
Y4		



Question 4 [19 marks].

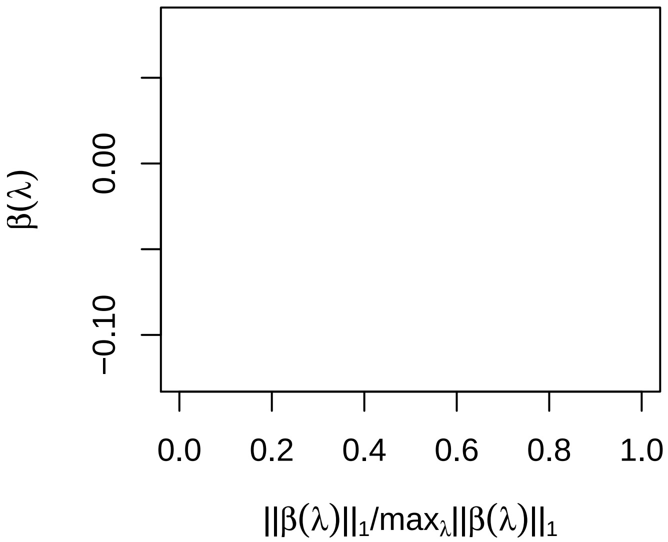
(a) The following table contains output from a lasso fit to model with $d = 5$ variables and $n = 40$ observations. For each row in the table, compute s, the proportion of shrinkage defined as $s = s(\lambda) = ||\beta(\lambda)||_1 / \max_\lambda ||\beta(\lambda)||_1$ and write its value in the correct position to complete the table. **[8]**

Write your solutions here

λ	β_1	β_2	β_3	β_4	β_5	s
0	0.08267	-0.0257	-0.12478	-0.0834	-0.07588	
1.10315	0.04204	0	-0.10618	-0.05665	-0.05547	
2.29358	0	0	-0.08636	-0.0304	-0.02125	
3.14574	0	0	-0.07455	-0.00839	0	
3.54098	0	0	-0.06557	0	0	
6	0	0	0	0	0	

(b) Using your completed information, add the lasso paths to the following plot. In your plot, label each path according to its corresponding variable. **[11]**

Write your solutions here



Extra space for calculations

End of Paper – An appendix of 2 pages follows.

Matrices and their decompositions for data handling

Data set: $\mathbf{X}$
Sample covariance matrix of a centered data set: $\boldsymbol{\Sigma} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$.
Karhunen-Loeve decomposition of the covariance matrix: $\boldsymbol{\Sigma} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^T$.
The total (sample) variance is $\text{tr}(\boldsymbol{\Sigma}) = \sum_{i=1}^{p} \lambda_i$
Singular value decomposition of $\mathbf{X}$: $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$.

Object	Size	Notes
n	-	Number of observations
p	-	Number of variables
$\mathbf{X}$	$n \times p$	$n > p$
$\boldsymbol{\Sigma}$	$p \times p$	(i, j) element is the sample covariance between columns i, j of $\mathbf{X}$
$\mathbf{A}$	$p \times p$	Eigenvectors $\mathbf{a}_i$ are columns; $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}$
$\boldsymbol{\Lambda}$	$p \times p$	Diagonal with eigenvalues λ_i of $\boldsymbol{\Sigma}$
$\mathbf{U}$	$n \times p$	Eigenvectors $\mathbf{u}_i$ are columns; $\mathbf{U}^T\mathbf{U} = \mathbf{I}$
$\mathbf{D}$	$p \times p$	Diagonal with singular eigenvalues d_i of $\mathbf{X}$
$\mathbf{V}$	$p \times p$	Eigenvectors $\mathbf{v}_i$ are columns; $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$

As in lectures, it is assumed that the rank of $\mathbf{X}$ is p so that $\mathbf{X}^T\mathbf{X}$ is invertible; that the diagonal entries in either $\boldsymbol{\Lambda}$ or $\mathbf{D}$ are all distinct and positive numbers.
For a square full rank diagonal matrix $\mathbf{W} = \text{diag}(w_1, \ldots, w_m)$, we have $\mathbf{W} = \mathbf{W}^T$; that $\mathbf{W}^T\mathbf{W} = \mathbf{W}\mathbf{W}^T = \mathbf{W}^2 = \text{diag}(w_1^2, \ldots, w_m^2)$ and $\mathbf{W}^{-1} = \text{diag}(1/w_1, \ldots, 1/w_m)$.

Common distances d_{ij} between points $\mathbf{x}_i$ and $\mathbf{x}_j$

Name	Distance d_{ij}	Equivalent notation
Manhattan	$\sum_{l=1}^{p} \lvert x_{il} - x_{jl} \rvert$	$\lvert\lvert \mathbf{x}_i - \mathbf{x}_j \rvert\rvert_1$
Euclidean	$\sqrt{\sum_{l=1}^{p} (x_{il} - x_{jl})^2}$	$\lvert\lvert \mathbf{x}_i - \mathbf{x}_j \rvert\rvert_2 = \sqrt{\lvert\lvert \mathbf{x}_i - \mathbf{x}_j \rvert\rvert_2^2}$
Minkowski	$\left(\sum_{l=1}^{p} \lvert x_{il} - x_{jl} \rvert^m\right)^{1/m}$	$\lvert\lvert \mathbf{x}_i - \mathbf{x}_j \rvert\rvert_m$
Mahalanobis	$\sqrt{(\mathbf{x}_{i\cdot} - \mathbf{x}_{j\cdot})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_{i\cdot} - \mathbf{x}_{j\cdot})}$	
Absolute correlation	$\sqrt{1 - \lvert \rho_{ij} \rvert}$	

The Manhattan norm of a vector $\mathbf{x}$ is $||\mathbf{x}||_1 = \sum_{i=1}^{p} |x_i|$ (Manhattan distance between $\mathbf{x}$ and the origin); and the Euclidean norm of $\mathbf{x}$ is $||\mathbf{x}||_2 = \sqrt{\sum_{i=1}^{p} x_i^2}$ (Euclidean distance between $\mathbf{x}$ and the origin).

Classification

Name	Comment
TP	True positive
FP	False positive
FN	False negative
TN	True negative
$P = TP + FN$	Positives
$N = FP + TN$	Negatives
$TPR = \frac{TP}{P}$	True positive rate (Sensitivity or Recall)
$FPR = \frac{FP}{N}$	False positive rate
$FNR = \frac{FN}{P}$	False negative rate
$TNR = \frac{TN}{N}$	True negative rate (Specificity)
$\frac{TP}{TP+FP}$	Precision
$2\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$	F1 score
$\frac{TP+TN}{P+N}$	Accuracy
$\frac{FP+FN}{P+N}$	Error rate

Penalised regression

Lasso criterion:

$$L = \frac{1}{2}||\mathbf{Y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_1$$

Ridge regression criterion:

$$R = ||\mathbf{Y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_2^2.$$

End of Appendix.

Queen Mary
University of London

Main Examination period 2020 – May – Semester B

MTH6101: Introduction to Machine Learning - SAMPLE

Duration: 2 hours

Student number | | | | | | | | | **Desk number** | | |

Make and model of calculator used ______________________________

Apart from this page, you are not permitted to read the contents of this question paper until instructed to do so by an invigilator.

Write your solutions in the spaces provided in this exam paper. If you need more paper, ask an invigilator for an additional booklet and attach it to this paper at the end of the exam.

You should attempt ALL questions. Marks available are shown next to the questions.

Only non-programmable calculators that have been approved from the college list of non-programmable calculators are permitted in this examination. Please state on your answer book the name and type of machine used.

Complete all rough work in the answer book and cross through any work that is not to be assessed.

Possession of unauthorised material at any time when under examination conditions is an assessment offence and can lead to expulsion from QMUL. Check now to ensure you do not have any unauthorised notes, mobile phones, smartwatches or unauthorised electronic devices on your person. If you do, raise your hand and give them to an invigilator immediately.

It is also an offence to have any writing of any kind on your person, including on your body. If you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms, it will be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. A mobile phone that causes a disruption in the exam is also an assessment offence.

Exam papers must not be removed from the examination room.

Examiners: H. Maruri-Aguilar, M. Benning

This page is for marking purposes only.

Do not write on it.

Question	Mark	Comments
1	/ 29	
2	/ 24	
3	/ 28	
4	/ 19	
Total		

Question 1 [29 marks].
A data set consists of measurements of amounts of chemical compounds found in material samples of a mineral processing facility. Seven measurements were available for each sample of a total of $n = 200$ samples. The units of the measurements are milligrammes.
From a Principal Component Analysis of these data, the following are the PC loadings.

```
##         PC1      PC2      PC3      PC4      PC5      PC6      PC7
## Na  0.48335 -0.00822  0.11289 -0.22665  0.36347 -0.09448 -0.74910
## Au  0.37907  0.07811 -0.86478  0.31936 -0.01328  0.01060  0.00901
## NO3 0.46602 -0.00584  0.04666 -0.45909 -0.75041  0.00548  0.08190
## Al  0.00564 -0.70429 -0.06960 -0.00191  0.01109 -0.69995  0.09513
## Cu  0.48040 -0.00519  0.12123 -0.24015  0.52372  0.09342  0.64328
## Mg  0.41794 -0.03411  0.46375  0.76035 -0.17366 -0.00960  0.02684
## Zn  0.00311 -0.70469 -0.05134  0.00872 -0.00604  0.70156 -0.09205
```

The corresponding eigenvalues of the Karhunen-Loeve decomposition of the covariance matrix of these data are 4.11178, 1.98865, 0.49036, 0.29693, 0.07676, 0.02234, 0.01318.

(a) Each of the data columns was centered around its mean and standardised to have unit variance. From the output you have been given, briefly explain how we can confirm that the data was indeed standardized by column. **[4]**

Write your solutions here

To determine if analysis was carried out standardising columns, we need to look at the total sum of eigenvalues [2]. This sum is

$$4.11178+1.98865+0.49036+0.29693+0.07676+0.02234+0.01318 = 7$$

As this total equals p, then the analysis was indeed performed in standardised columns [2].

Total (a) [4]

Seen in lectures and lab.

(b) Complete the following table. **[15]**

Write your solutions here

	Standard deviation	Proportion of variance	Cumulative proportion
PC1	2.0278	0.5874	0.5874
PC2	1.4102	0.2841	0.8715
PC3	0.7003	0.0701	0.9415
PC4	0.5449	0.0424	0.984
PC5	0.277	0.011	0.9949
PC6	0.1495	0.0032	0.9981
PC7	0.1148	0.0019	1

Square roots of eigenvalues give standard deviations; the proportion scales eigenvalues to percentages, from which the cumulative column is computed. Each column [5] to Total (b) [15]

Seen in lectures and lab.

(c) Using the table you just computed, suggest a number of components so that at least 80% of the total variability is explained. **[4]**

(d) Interpret the components you suggested. **[6]**

Write your solutions here

(c) Looking at the column of 'Cumulative proportion' in the said Table, we suggest 2 components which in this case explain 87.15% of the total variability [4].

Total (c) [4]

(d) The following interpretation is generated by seeing the coefficients of PC loadings as weigthed averages or contrasts: PC1 is a weighted average of variables Na,Au,NO3,Cu,Mg; PC2 is a contrast between variables Au and Al,Zn; PC3 is a contrast between variables Na,Cu,Mg and Au,Al,Zn..

Each sensible interpretation [3] to Total (d) [6]

Seen in lectures and coursework.

Question 2 [24 marks].

(a) Explain what is meant by **single** linkage in agglomerative clustering. **[3]**

(b) Consider the following data matrix

$$\mathbf{X} = \begin{pmatrix} 2 & -1 & -3 & 3 \\ 1 & 1 & 1 & 3 \\ 3 & -4 & -3 & -1 \\ -4 & 3 & -2 & -3 \\ -4 & 3 & -3 & 2 \end{pmatrix},$$

(i) Using Manhattan distance, complete the missing entries in the distance matrix below. **[4]**

Write your solutions here

(a) In single linkage, the distance between two clusters is defined as the distance between the two closest elements of each cluster [3]. This is the "nearest neighbor" distance.

Total (a) [3]

$$\text{b(i):} \quad \begin{pmatrix} 0 & \boxed{7} & 8 & 17 & 11 \\ \boxed{7} & 0 & 15 & \boxed{16} & 12 \\ 8 & 15 & 0 & 17 & 17 \\ 17 & \boxed{16} & 17 & 0 & \boxed{6} \\ 11 & 12 & 17 & \boxed{6} & 0 \end{pmatrix},$$

Distances are computed with the usual Manhattan distance, e.g.
$d_{12} = 1 + 2 + 4 + 0 = 7;$
$d_{24} = 5 + 2 + 3 + 6 = 16$
$d_{45} = 0 + 0 + 1 + 5 = 6$

Each distance [1] all fine add [1] to Total (bi) [4]

Both items seen in lectures.

(ii) Perform agglomerative cluster for these data using the distance matrix above and single linkage. As part of your procedure, compute and show updated distance matrices at every step of the clustering procedure. **[12]**

Write your solutions here

(bii) The initial step clusters individuals '4' and '5' at the distance 6. The updated distance matrix is

	1	2	3	45
1	0	7	8	11
2	7	0	15	12
3	8	15	0	17
45	11	12	17	0

The next step merges clusters '1' and '2' at the distance 7. The updated table of distances at this stage is the following.

	12	3	45
12	0	8	11
3	8	0	17
45	11	17	0

The procedure continues by finding the minimum distance between clusters in the newly updated distance table. This minimum value is 8 at which point the clusters '12' and '3' are merged. This is the updated table of distances at this stage.

	123	45
123	0	11
45	11	0

The last step of this agglomerative clustering process is the trivial merging of the clusters '123' and '45' at the distance 11.

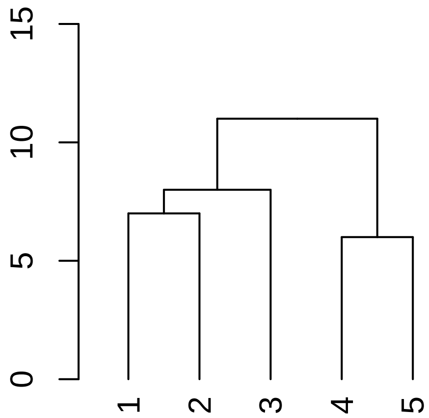
Each updated step with table worth [4] marks to Total (bii) [12]

Seen in lectures and coursework.

Write your solutions here

(iii) Complete the dendrogram in the plot given below. **[5]**

Write your solutions here



Correct diagram one mark per branch to Total (biii) [5]

Seen in lectures and coursework.

Question 3 [28 marks]. The following data are the validation results of a classification analysis. The output below includes the validation output `Ytrue` and the classifications obtained with four classification algorithms termed `Y1`, `Y2`, `Y3` and `Y4`.

```
##       Ytrue Y1 Y2 Y3 Y4
##  [1,]     1  1  0  1  0
##  [2,]     0  0  0  0  0
##  [3,]     0  0  1  0  1
##  [4,]     1  1  0  1  0
##  [5,]     0  0  1  0  1
##  [6,]     0  0  1  1  1
##  [7,]     0  0  1  0  1
##  [8,]     1  1  0  1  1
##  [9,]     0  1  1  0  0
## [10,]     0  0  1  1  0
```

(a) Complete the following confusion matrices. **[14]**

Write your solutions here

		Predicted (`Y1`) 0	1
True (`Ytrue`)	0	6	1
	1	0	3

		Predicted (`Y2`) 0	1
True (`Ytrue`)	0	1	6
	1	3	0

		Predicted (`Y3`) 0	1
True (`Ytrue`)	0	5	2
	1	0	3

		Predicted (`Y3`) 0	1
True (`Ytrue`)	0	3	4
	1	2	1

Each matrix [3] if all fine extra [2] to Total (a) [14].

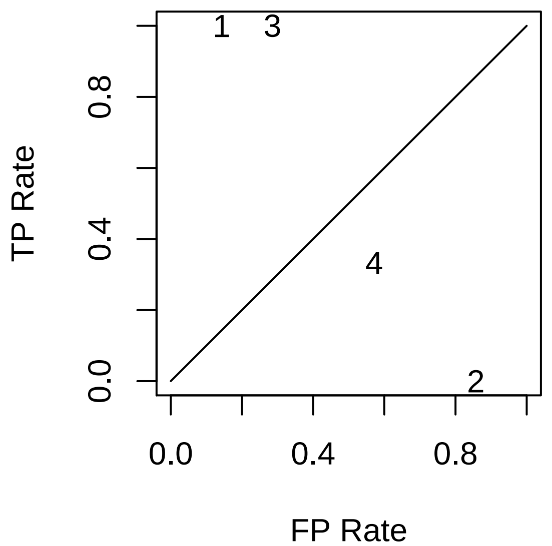
Seen in lectures and coursework.

(b) Compute the False Positive Rate (FPR) and True Positive Rate (TPR) for each confusion matrix, completing in the table below. **[6]**

(c) Plot your results in the ROC graph below and briefly comment on the performance of classifiers. Which is the best classifier? **[8]**

Write your solutions here

Confusion matrix	FPR	TPR
`Y1`	0.14286	1
`Y2`	0.85714	0
`Y3`	0.28571	1
`Y4`	0.57143	0.33333



The best classifier is `Y1`, followed by `Y3` then `Y4` whose performance is similar to a random classifier. The classifier `Y2` performs the worst.

Each column of the Table [3] so Total (b) [6].

Plot correctly done [6]. Best classifier and/or meaningful comment [2] thus Total (c) [8].

Seen in lectures and coursework.

Question 4 [19 marks].

(a) The following table contains output from a lasso fit to model with $d = 5$ variables and $n = 40$ observations. For each row in the table, compute s, the proportion of shrinkage defined as $s = s(\lambda) = ||\beta(\lambda)||_1 / \max_\lambda ||\beta(\lambda)||_1$ and write its value in the correct position to complete the table. **[8]**

Write your solutions here

λ	β_1	β_2	β_3	β_4	β_5	s
0	0.08267	-0.0257	-0.12478	-0.0834	-0.07588	1
1.10315	0.04204	0	-0.10618	-0.05665	-0.05547	0.66341
2.29358	0	0	-0.08636	-0.0304	-0.02125	0.35169
3.14574	0	0	-0.07455	-0.00839	0	0.21135
3.54098	0	0	-0.06557	0	0	0.1671
6	0	0	0	0	0	0

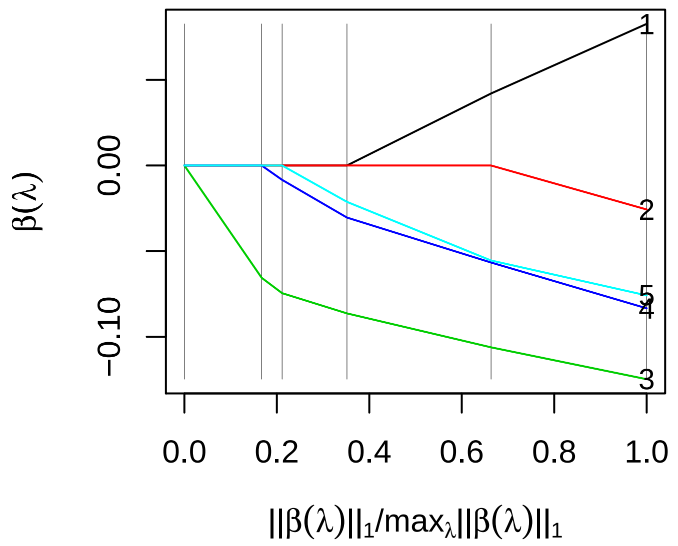
Computation of s involve the sums of absolute coefficients 0.39243, 0.26034, 0.13801, 0.08294, 0.06557, 0 which are then divided by the maximum 0.39243 to give the values in the table.

Each correct value of s [1] and if all correct add [2] to Total (c) [8]

Seen in lectures and coursework.

(b) Using your completed information, add the lasso paths to the following plot. In your plot, label each path according to its corresponding variable. **[11]**

Write your solutions here



Each correct path [2] and if all correct add [1] to Total (d) [11]

Seen in lectures and coursework.

Extra space for calculations

End of Paper – An appendix of 2 pages follows.

Matrices and their decompositions for data handling

Data set: $\mathbf{X}$
Sample covariance matrix of a centered data set: $\mathbf{\Sigma} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$.
Karhunen-Loeve decomposition of the covariance matrix: $\mathbf{\Sigma} = \mathbf{A}\mathbf{\Lambda}\mathbf{A}^T$.
The total (sample) variance is $\mathrm{tr}(\mathbf{\Sigma}) = \sum_{i=1}^{p} \lambda_i$
Singular value decomposition of $\mathbf{X}$: $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$.

Object	Size	Notes
n	-	Number of observations
p	-	Number of variables
$\mathbf{X}$	$n \times p$	$n > p$
$\mathbf{\Sigma}$	$p \times p$	(i, j) element is the sample covariance between columns i, j of $\mathbf{X}$
$\mathbf{A}$	$p \times p$	Eigenvectors $\mathbf{a}_i$ are columns; $\mathbf{A}^T\mathbf{A} = \mathbf{A}\mathbf{A}^T = \mathbf{I}$
$\mathbf{\Lambda}$	$p \times p$	Diagonal with eigenvalues λ_i of $\mathbf{\Sigma}$
$\mathbf{U}$	$n \times p$	Eigenvectors $\mathbf{u}_i$ are columns; $\mathbf{U}^T\mathbf{U} = \mathbf{I}$
$\mathbf{D}$	$p \times p$	Diagonal with singular eigenvalues d_i of $\mathbf{X}$
$\mathbf{V}$	$p \times p$	Eigenvectors $\mathbf{v}_i$ are columns; $\mathbf{V}^T\mathbf{V} = \mathbf{V}\mathbf{V}^T = \mathbf{I}$

As in lectures, it is assumed that the rank of $\mathbf{X}$ is p so that $\mathbf{X}^T\mathbf{X}$ is invertible; that the diagonal entries in either $\mathbf{\Lambda}$ or $\mathbf{D}$ are all distinct and positive numbers.
For a square full rank diagonal matrix $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_m)$, we have $\mathbf{W} = \mathbf{W}^T$; that $\mathbf{W}^T\mathbf{W} = \mathbf{W}\mathbf{W}^T = \mathbf{W}^2 = \mathrm{diag}(w_1^2, \ldots, w_m^2)$ and $\mathbf{W}^{-1} = \mathrm{diag}(1/w_1, \ldots, 1/w_m)$.

Common distances d_{ij} between points $\mathbf{x}_i$ and $\mathbf{x}_j$

Name	Distance d_{ij}	Equivalent notation
Manhattan	$\sum_{l=1}^{p} \lvert x_{il} - x_{jl} \rvert$	$\lvert\lvert \mathbf{x}_i - \mathbf{x}_j \rvert\rvert_1$
Euclidean	$\sqrt{\sum_{l=1}^{p} \left(x_{il} - x_{jl}\right)^2}$	$\lvert\lvert \mathbf{x}_i - \mathbf{x}_j \rvert\rvert_2 = \sqrt{\lvert\lvert \mathbf{x}_i - \mathbf{x}_j \rvert\rvert_2^2}$
Minkowski	$\left(\sum_{l=1}^{p} \lvert x_{il} - x_{jl} \rvert^m\right)^{1/m}$	$\lvert\lvert \mathbf{x}_i - \mathbf{x}_j \rvert\rvert_m$
Mahalanobis	$\sqrt{\left(\mathbf{x}_{i\cdot} - \mathbf{x}_{j\cdot}\right)^T \mathbf{\Sigma}^{-1} \left(\mathbf{x}_{i\cdot} - \mathbf{x}_{j\cdot}\right)}$	
Absolute correlation	$\sqrt{1 - \lvert \rho_{ij} \rvert}$	

The Manhattan norm of a vector $\mathbf{x}$ is $\lvert\lvert \mathbf{x} \rvert\rvert_1 = \sum_{i=1}^{p} \lvert x_i \rvert$ (Manhattan distance between $\mathbf{x}$ and the origin); and the Euclidean norm of $\mathbf{x}$ is $\lvert\lvert \mathbf{x} \rvert\rvert_2 = \sqrt{\sum_{i=1}^{p} x_i^2}$ (Euclidean distance between $\mathbf{x}$ and the origin).

Classification

Name	Comment
TP	True positive
FP	False positive
FN	False negative
TN	True negative
$P = TP + FN$	Positives
$N = FP + TN$	Negatives
$TPR = \frac{TP}{P}$	True positive rate (Sensitivity or Recall)
$FPR = \frac{FP}{N}$	False positive rate
$FNR = \frac{FN}{P}$	False negative rate
$TNR = \frac{TN}{N}$	True negative rate (Specificity)
$\frac{TP}{TP+FP}$	Precision
$2\frac{\text{Precision}\cdot\text{Recall}}{\text{Precision}+\text{Recall}}$	F1 score
$\frac{TP+TN}{P+N}$	Accuracy
$\frac{FP+FN}{P+N}$	Error rate

Penalised regression

Lasso criterion:

$$L = \frac{1}{2}||\mathbf{Y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_1$$

Ridge regression criterion:

$$R = ||\mathbf{Y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_2^2.$$

End of Appendix.