QP Pattern:

- **Instance based learning: 4-8 marks**
- **SVM: 4 - 8 marks**
- **Bayesian Learning: 4 - 8 marks**
- **Ensemble Learning: 4 - 8 marks**
- **Unsupervised Learning: 4 - 8 marks**
- **Model Evaluation and Comparison: 4 - 8 marks**
- **Pre-midterm topics: 5-10 marks**

Course No.	: DSECLZG565/ AIMLCLZ565
Course Title	: Machine Learning
Nature of Exam	: Open Book
Weightage	: 40%
Duration	: 2 Hours
Date of Exam	:

Note:

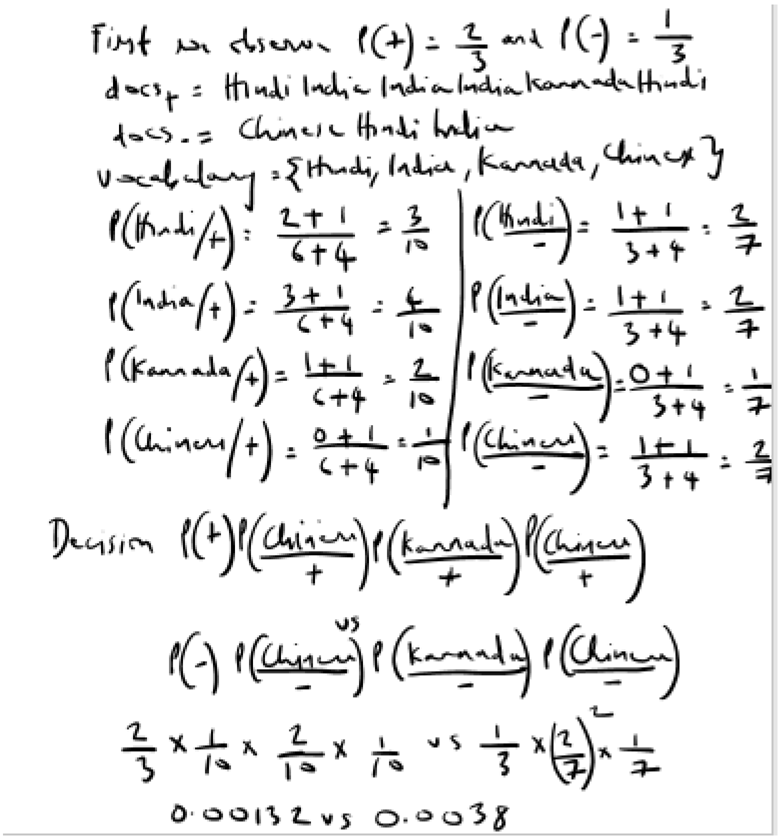
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Question 1:

Consider the following dataset for text classification where three training instances are given with corresponding classifications into the '+' or –'- category: [5]

Hindi India India	+
India Kannada Hindi	+
Chinese Hindi India	-

Showing all intermediate calculations, find the appropriate classification for the test instance: "Chinese Kannada Chinese" using the Multinomial NB text classification approach.



First we observe $P(+) = \frac{2}{3}$ and $P(-) = \frac{1}{3}$

$docs_+$ = Hindi India India India Kannada Hindi

$docs_-$ = Chinese Hindi India

vocabulary = {Hindi, India, Kannada, Chinese}

$P(Hindi/+) = \frac{2+1}{6+4} = \frac{3}{10}$ | $P\left(\frac{Hindi}{-}\right) = \frac{1+1}{3+4} = \frac{2}{7}$

$P(India/+) = \frac{3+1}{6+4} = \frac{4}{10}$ | $P\left(\frac{India}{-}\right) = \frac{1+1}{3+4} = \frac{2}{7}$

$P(Kannada/+) = \frac{1+1}{6+4} = \frac{2}{10}$ | $P\left(\frac{Kannada}{-}\right) = \frac{0+1}{3+4} = \frac{1}{7}$

$P(Chinese/+) = \frac{0+1}{6+4} = \frac{1}{10}$ | $P\left(\frac{Chinese}{-}\right) = \frac{1+1}{3+4} = \frac{2}{7}$

Decision $P(+)P\left(\frac{Chinese}{+}\right)P\left(\frac{Kannada}{+}\right)P\left(\frac{Chinese}{+}\right)$

vs

$P(-)\,P\left(\frac{Chinese}{-}\right)P\left(\frac{Kannada}{-}\right)P\left(\frac{Chinese}{-}\right)$

$\frac{2}{3} \times \frac{1}{10} \times \frac{2}{10} \times \frac{1}{10}$ vs $\frac{1}{3} \times \left(\frac{2}{7}\right)^2 \times \frac{1}{7}$

0.00132 vs 0.0038

Marking Scheme:

Positive Conditional Probabilities → 1.5 mark

Negative Conditional Probs → 1.5 mark

Decision - 2 marks

Question 2:

You have been tasked to create a discriminative model using a linear Support Vector Machine method. Consider the below training dataset for training the model, where X1, and X2 are independent features and Y is the target variable. [4]

X1	**X2**	**Y**
3	2	Positive
5	3	Positive
-2	-2	Negative
4	4	Positive
3	-1	Positive
1	0	Negative
-1	-1	Negative
0	2	Negative

1	4	Negative
-1	2	Negative
4	5	Positive

Answer the following questions:

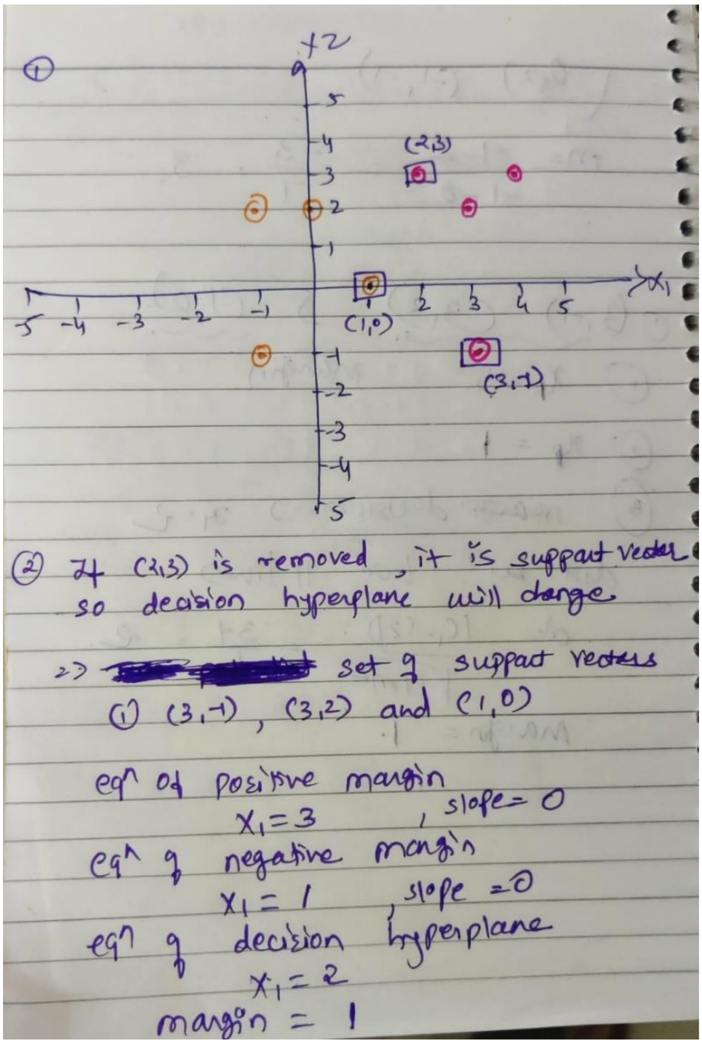
A. Find the support vectors [1]

B. Determine the equation of hyperplane [3]

Solution:

A. Support vectors are (3,2), (1,0) and (3,-1) [1M- if one of the SVs is wrong then 0 M]

B.



② If (2,3) is removed, it is support vector so decision hyperplane will change

=> [redacted] set of support vectors

① (3,-1), (3,2) and (1,0)

eqⁿ of positive margin

$x_1 = 3$, slope $= 0$

eqⁿ of negative margin

$x_1 = 1$, slope $= 0$

eqⁿ of decision hyperplane

$x_1 = 2$

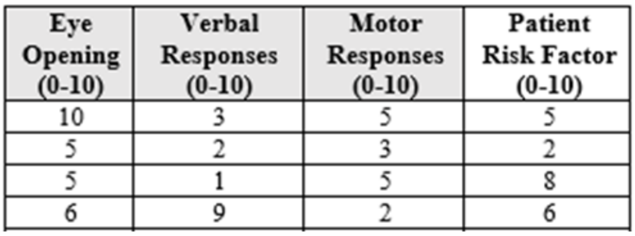
margin = 1

The solution obtained using the Lagrange method is equally acceptable.

Question 3:

Use case: The Glasgow Coma Scale assesses patients according to three aspects of responsiveness: eye-opening, motor, and verbal responses. Reporting each of these separately provides a clear, communicable picture of a patient's state.

[3 + 5 = 8 Marks]



Eye Opening (0-10)	Verbal Responses (0-10)	Motor Responses (0-10)	Patient Risk Factor (0-10)
10	3	5	5
5	2	3	2
5	1	5	8
6	9	2	6

Note: *Wherever applicable use only Manhattan distance & No scaling is required. Round all the calculation to 4 decimal places if any. Use average as the aggregation function for the final estimation wherever required unless specific function is recommended. Show all steps. Calculation error will also be penalized.*

Predict the risk factor of new patient with below observation using both the following independent experiments for the above training data.

Query Instance: <Eye Opening = 5, Verbal Responses = 5, Motor Responses = 5>

i) Predict the risk factor using 3-NN model.

ii) If the initial estimation is proposed as locally weighted regression model instead, use: ***Patient Risk = 10 – 0.1 $X_{VERBALRESPONSES}$ – 0.1 $X_{MOTORRESPONSES}$,*** and apply 2-NN with kernel: $K(d(x_q,x_i)) = (-1)/ d(x_q,x_i)$ and apply the gradient descent only for one iteration with learning rate = 0.1. Apply gradient descent only for one iteration and predict the risk factor for the query instance.

Answer Key:

a) Prediction with average of below 3-NN = 5

Manhattan Distance
7
5
4
8
10

b) Only for the top 2-NN Below are calculated:
Y-Pred = {9.5, 9.4}
Delta gradient for Wverbalresponse = {3, 0.35}
Delta gradient for Wmotorresponse = {4.5, 1.75}
Delta gradient for Wo = {1.5, 0.35}
New (Wverbalresponse, Wmotorresponse, Wo) = (0.435, 0.725, 10.185)
Prediction with new regression equation = 4.385

Marking Scheme:

a)

2 mark: Distance calculation between test and all other instances. Order the results of the distance.

1 mark: Results of 3-NN average

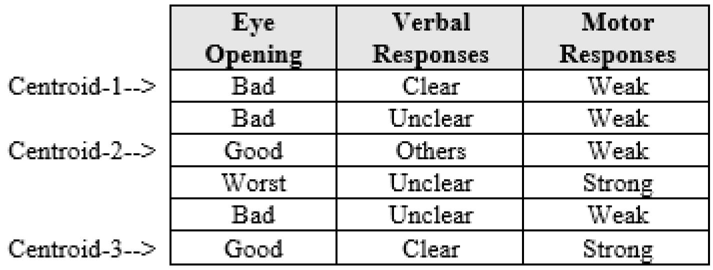
b)

1 mark : Y-Pred calculation
0.5 mark : Delta gradient for Wverbalresponse calculation
0.5 mark : Delta gradient for Wmotorresponse calculation
1 mark : Delta gradient for Wo calculation
1 mark : New (Wverbalresponse, Wmotorresponse, Wo)
1 mark : Prediction with new regression equation

Question 4:

Use case: The Glasgow Coma Scale assesses patients according to three aspects of responsiveness: eye-opening, motor, and verbal responses. Reporting each of these separately provides a clear, communicable picture of a patient's state. Quantified values of attributes are discretized in below data.

[4 + 1 + 2 = 7 Marks]



	Eye Opening	Verbal Responses	Motor Responses
Centroid-1-->	Bad	Clear	Weak
	Bad	Unclear	Weak
Centroid-2-->	Good	Others	Weak
	Worst	Unclear	Strong
	Bad	Unclear	Weak
Centroid-3-->	Good	Clear	Strong

a) Use following distance measure to cluster the given patients into three clusters using k-modes clustering algorithm for only one iteration. Show the step by step working of the Expectation and Maximization step. The centroids are marked in the given table. Assume all the features are categorical in nature and use only the following distance metric for your calculation. Round-off all the proximity values to two decimal places.

$$distance\ (data1\ , data2) = 10 * \left(1 - \frac{Number.of.matching.categorical.attributes}{Total.number.of.categorical.attributes}\right)$$

$$Median\ of\ categorical\ attributes\ within\ a\ cluster = \begin{cases} Mode\ of\ attribute\ value & if\ cluster\ size\ is\ odd \\ t & otherwise \end{cases}$$

$$where\ \ t = \ Least\ frequent\ attrbute\ value\ observed\ in\ the\ entire\ training\ data$$

b) Calculate the new centroids

c) State if the below given statement is true or false w.r.t to given data whose centroids are sampled with replacements. Justify your statement with plagiarism free explanations.

"If the number of clusters expected is equivalent to the number of data points/instances given for training, then the algorithm is guaranteed to converge in atmost one iteration of Expectation & Maximization"

Answer Key:

a) Only the three non-centroid points are required to be compared against the centroids. Highlighted are the members of clusters after Expectation step

	Eye Opening	Verbal Responses	Motor Responses	Distance with Centroid-1	Distance with Centroid-2	Distance with Centroid-3
Centroid-1-->	Bad	Clear	Weak			
	Bad	Unclear	Weak	3.33	6.67	10
Centroid-2-->	Good	Others	Weak			
	Worst	Unclear	Strong	10	10	6.67
	Bad	Unclear	Weak	3.33	6.67	10

Centroid-3-->	Good	Clear	Strong			

Maximization:
New Centroid 1 = (Bad, Unclear, Weak)
Centroid 2 : No Change
New Centroid 3 = (Worst, Others, Strong)

b) FALSE. Note due to sampling method chosen, same data might be chosen to be arbitrary centroids of multiple clusters. Hence convergence may needs more than one EM iteration in worst case.

Marking Scheme:

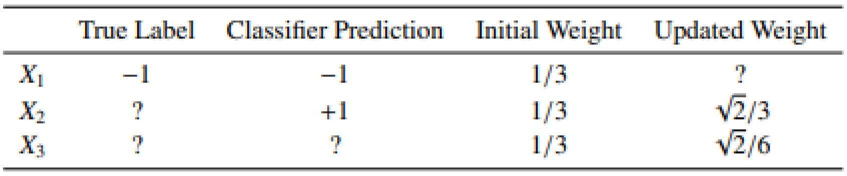
a)
1 mark: Distance Calculation w.r.t C1
1 mark: Distance Calculation w.r.t C2
0.5 mark: Distance Calculation w.r.t C3
0.5 mark: New Centroid using mode of member's value for centroid 1
1 mark: New Centroid using given formula for centroid 3

b)
1 mark: No new calculation are required, All the values used in part a) has to correctly referred to find the difference in value

c)
1 mark: Answer "False"
1 mark: Right Justification

Question 5:

In a single iteration of AdaBoost on three sample points, we initiate the process with uniform weights assigned to the sample points. The ground truth labels and predictions are binary, taking values of either +1 or −1. The table provided below contains some missing values. [3]

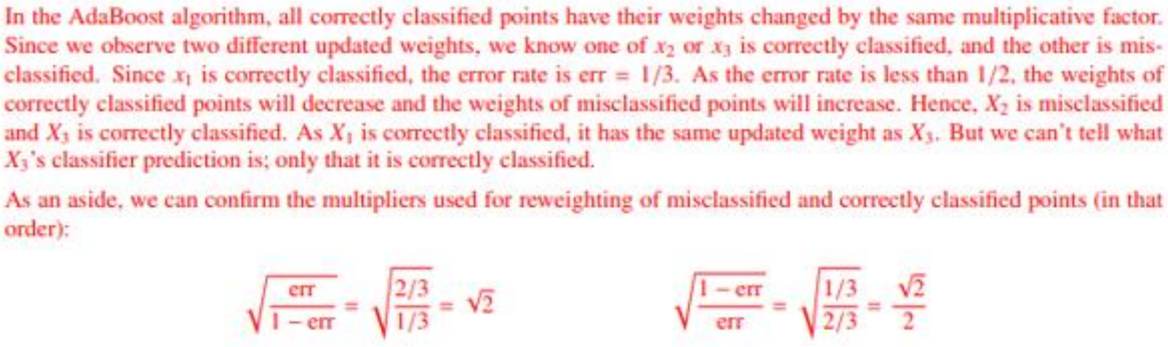


	True Label	Classifier Prediction	Initial Weight	Updated Weight
X_1	-1	-1	$1/3$	?
X_2	?	$+1$	$1/3$	$\sqrt{2}/3$
X_3	?	?	$1/3$	$\sqrt{2}/6$

Answer the following:

a) Find the updated weight (before normalization) of X1 instance. Note, no need to normalize the values. [1.5]

b) Identify which instances/data points were misclassified in the first iteration. Justify your answer. [1.5]



In the AdaBoost algorithm, all correctly classified points have their weights changed by the same multiplicative factor. Since we observe two different updated weights, we know one of x_2 or x_3 is correctly classified, and the other is misclassified. Since x_1 is correctly classified, the error rate is err = 1/3. As the error rate is less than 1/2, the weights of correctly classified points will decrease and the weights of misclassified points will increase. Hence, X_2 is misclassified and X_3 is correctly classified. As X_1 is correctly classified, it has the same updated weight as X_3. But we can't tell what X_3's classifier prediction is; only that it is correctly classified.

As an aside, we can confirm the multipliers used for reweighting of misclassified and correctly classified points (in that order):

$$\sqrt{\frac{\text{err}}{1-\text{err}}} = \sqrt{\frac{2/3}{1/3}} = \sqrt{2} \qquad \sqrt{\frac{1-\text{err}}{\text{err}}} = \sqrt{\frac{1/3}{2/3}} = \frac{\sqrt{2}}{2}$$

Part a) [1.5 marks for the correct updated value, otherwise 0]

Part b) 0.5 marks for the correct answer, 1 mark for justification]

a) Random Forest is a bagging model that incorporates feature randomness. Clarify the rationale behind introducing feature randomness in Random Forest. [2]

2 marks for justification

Decision trees in a Random Forest may be highly correlated, especially when there are a few dominating features that provide most of the information for splitting. Hence, feature randomness is also introduced in Random forest

Question 6:

Illustrate a situation in which a lack of interpretability in a model could result in ethical issues. Explain how interpretability could mitigate these concerns.

Solution:

A scenario where lack of model interpretability could raise ethical concerns is in automated hiring systems. If a machine learning model is used to screen job applicants and is not interpretable, it might unintentionally favor or discriminate against candidates based on sensitive attributes like gender, ethnicity, or age without providing any rationale.

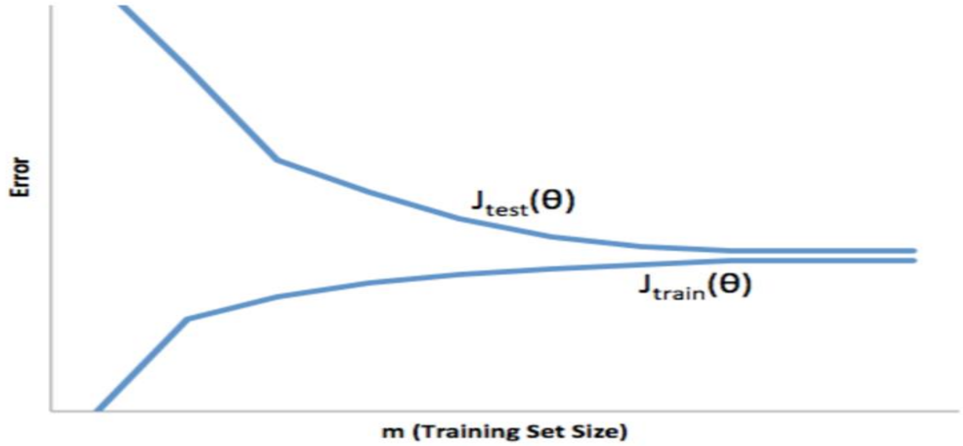
Mitigation through Interpretability:

Bias Detection: Interpretability allows stakeholders to examine the model's decision-making process, helping identify if and how certain features (e.g., gender or ethnicity) are disproportionately affecting the model's decisions.

Accountability: It ensures that decisions made by ML models are transparent and accountable. If a model unfairly discriminates against a group of applicants, interpretability helps trace back the decision to its underlying cause, enabling corrective measures.

Regulatory Compliance: By making models interpretable, organizations can ensure compliance with anti-discrimination laws and regulations, thus avoiding legal and reputational damage.

a) You have trained a ML model and discovered that it yields unacceptably high error on the test data. You also plotted a learning curve for both test data and training data as shown below. Comment on the performance of this ML model. Additionally, discuss strategies to address such cases, including the approaches and measures you would take in such scenarios. [2.5]



High Bias [1 mark if explanation is also provided otherwise 0.5 marks]
https://www.dataquest.io/blog/learning-curves-machine-learning/
[1.5 mark to suggest the strategies to avoid high bias]

b)

Question 7:

a) You are fitting a logistic regression model to predict whether an email is spam (class 1) or not (class 0) based on the length of the email's subject line (Feature). The model's coefficients are:

Coefficient:	0.03
Intercept:	-1.2

Calculate the predicted probability of an email being spam for an instance with a subject line length of 50 characters and classify the instance, assuming 70% is the threshold. Additionally, justify the assignment of a specific class to the given instance. [5 marks]

To calculate the predicted probability of an email being spam for an instance with a subject line length of 50 characters, you can use the logistic regression equation:

$\text{logit}(p)=\beta0+\beta1\times\text{Feature}$

Where:

logit(p) is the logarithm of the odds of the positive class (spam) probability,
- $\beta0$ is the intercept (bias) coefficient,
- $\beta1$ is the coefficient for the feature (subject line length).

In this case:

- $\beta0=-1.2$ (Intercept)
- $\beta1=0.03$ (Coefficient)

Feature=50 (Subject line length)

Substitute these values into the equation:

$\text{logit}(p)=-1.2+0.03\times50$

$\text{logit}(p)=-1.2+1.5$

$\text{logit}(p)=0.3$

Now, to convert the logit back into a probability p, you can use the sigmoid (logistic) function:

$p=1/\ (1+e\text{^}(-\text{logit}(p))$ **[1 marks for the equation]**

Substitute the logit value:

$p=1/(1+e\text{^}-0.31$ **[1 marks for calculation]**

$p=1/1+0.7408181$

$p\approx0.57444$

So, the predicted probability of an email with a subject line length of 50 characters being classified as spam is approximately 0.57444 or 57.44%.

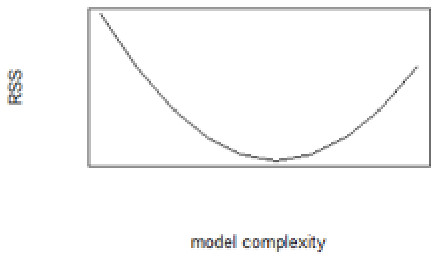
Since the predicted probability is 57.4% which is less than the threshold, i.e. 70%, hence the predicted class would be “Not Spam” **[0.5 marks for the correct classification, 0.5 marks for the justification]**

b) What are the shortcomings of using Entropy (Information Gain) as a heuristic measure while building decision trees?
[1.5 marks]

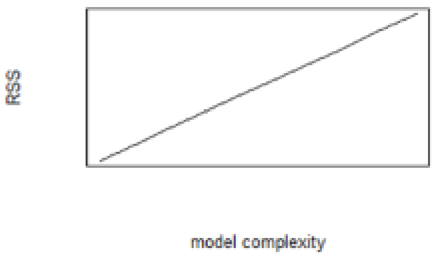
Information gain measure is biased towards attributes with a large number of distinct values. Eg. Product_ID (unique for every tuple), resulting in large number of partitions as Info$_{\text{product\_ID}}$ (D) = 0, Such partitioning is useless.

c) Which of the following charts of Residual Sum of Squares (RSS) and model complexity represent training phase for a fixed dataset?

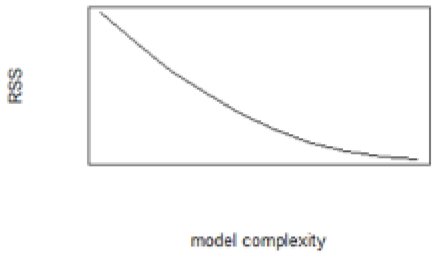
a



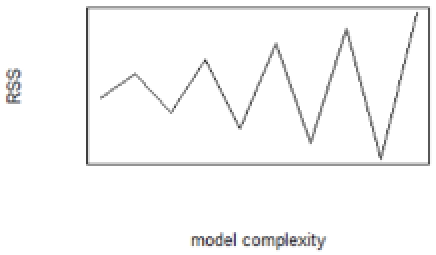
b



c



d



Answer c) [0.5 marks]
[1marks for the justification]