AHMEDABAD UNIVERSITY
School of Engineering and Applied Science
2025-2026 Monsoon

CSE500:in Statistical Learning
Makeup Final Examination
Friday January 2nd, 2026, 12:00-14:00

Name:	Roll no.	A U 2	Marks /80

Write your answers in the answer book. Attach this sheet to the answer book and then submit it. Answer all questions. Students are allowed 5 A-4 sized cheat sheets. Access to net is not allowed.

1. [15 points] Given n observations $\{(x1, y1), (x2, y2), \ldots, (xn, yn)\}$ where each input xi is a p-dimensional vector and the corresponding output yi is a positive real number. Now, suppose you train a model using the n input observations using k-nearest neighboours, and you are also given a test point $x0$ for which you need to predict its response, please answer the following questions:

(a) [5 points] Given a test point $x0$, how will you predict its response? How many parameters will this model contain?

(b) [5 points] How does the choice of k affect the training error of k-nearest neighbours? Explain in terms of impact of k on bias and variance.

(c) [5 points] From model interpretability and prediction accuracy perspective, what would be a good choice for k? Justify.

2. [10 points] Given the following situations where we either fit the points in 2D using polynomials, which of these models overfit, under-fit and are right fit? Justify your answer and reason in terms of bias and variance *Note the sine curve is the true function and the other curve is the polynomial determined by the model.*

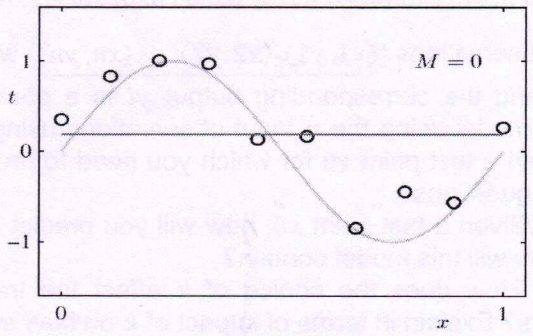


Figure 1.

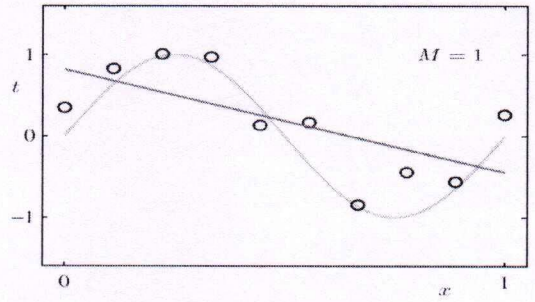


Figure 2.

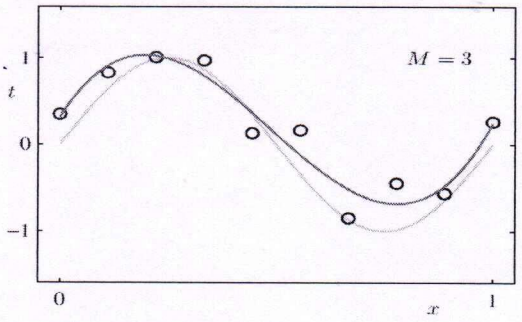


Figure 3.

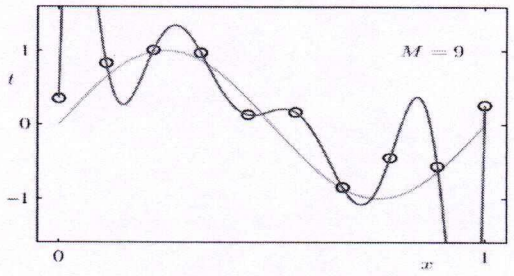
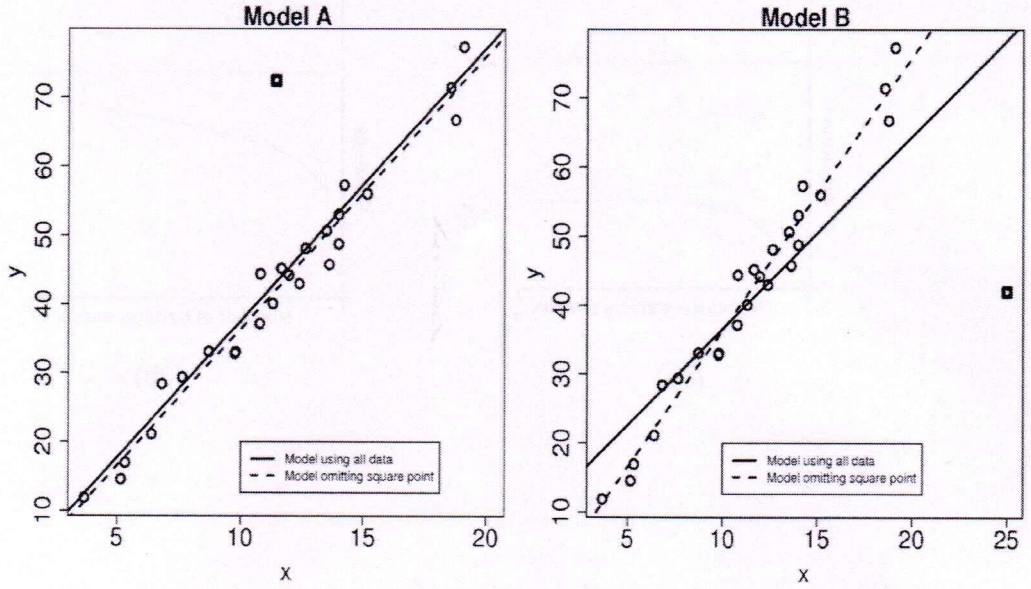
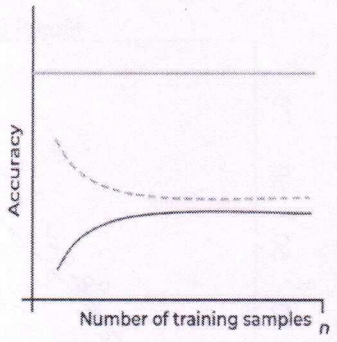


Figure 4.

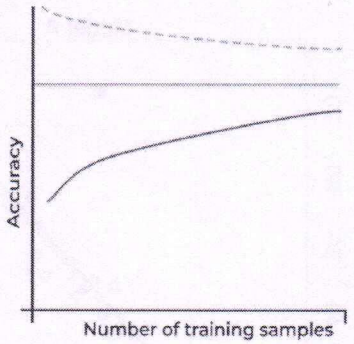
3. [5 points] You are given two models: Model A and Model B. For Model A, the square point that is an outlier, and for Model B, the square point is a high leverage point. How do these two points affect the model accuracy and the model variability? Justify using what measures you use to quantify their effects.



4. [10 points] You are given the curves of training (green-dotted) and test (blue-bottom line) accuracy for different input sizes. Red line (horizontal line) indicates the desired accuracy. In these situations, determine whether there is a problem and if so, what is the problem how will you fix the model to improve prediction accuracy?

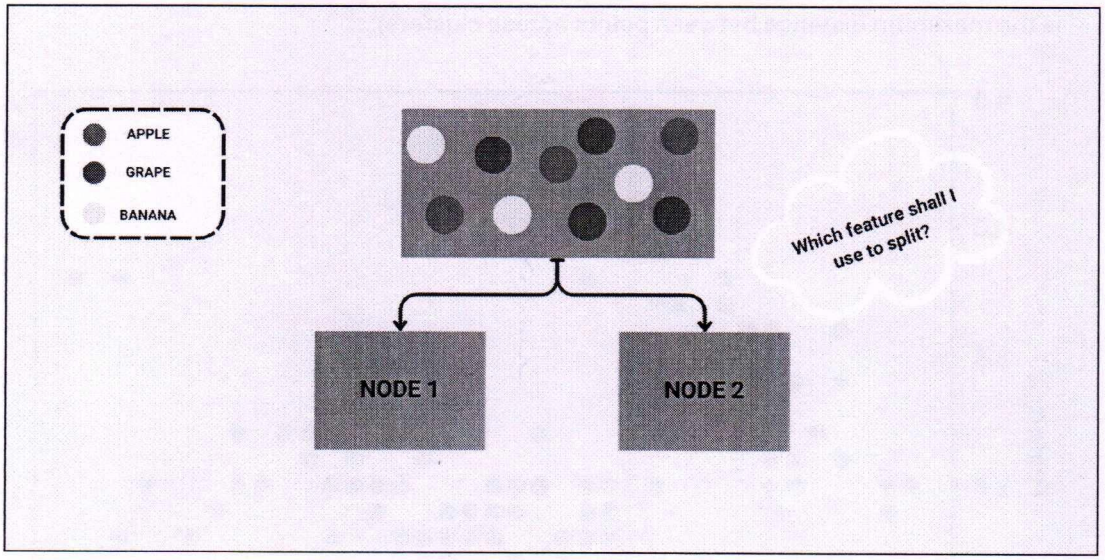


(3a)

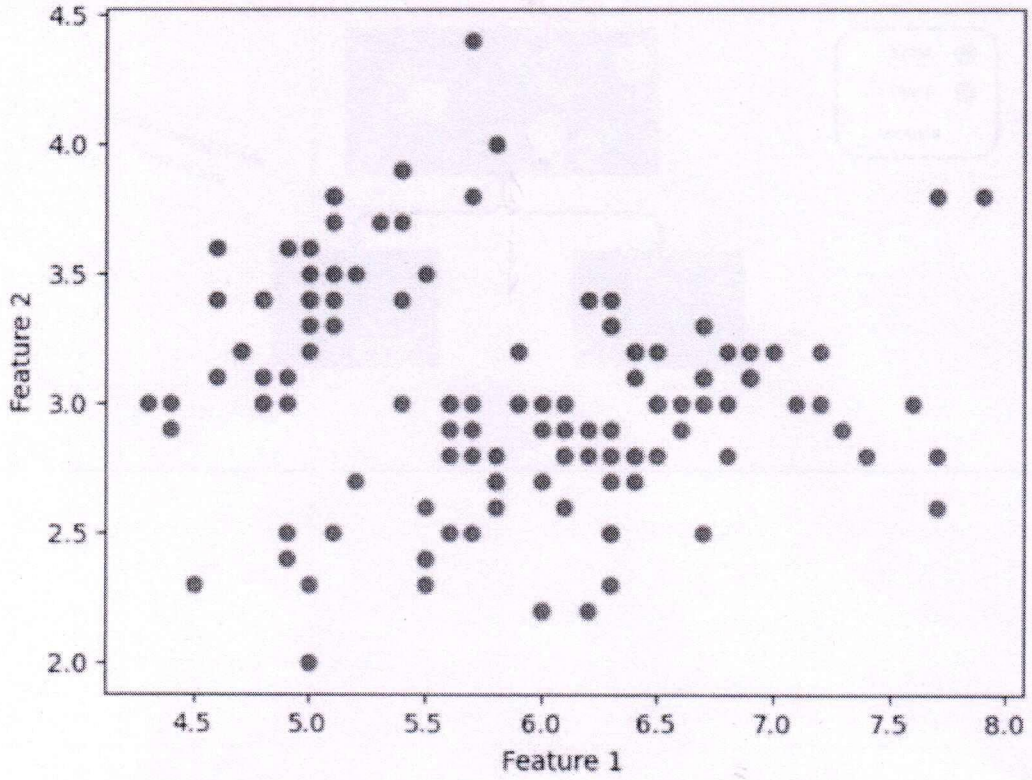


(3b)

5. [10 points] You are given a decision tree node containing 10 objects (3 Apples, 4 grapes and 3 bananas) that needs to be partitioned into two leaf-nodes so that (i) Gini-index is maximized (ii) entropy is minimized. Show how you will do the partitioning.



6. [10 points] Given the following cluster of points where the distance between these points is measured in terms of their 2D Euclidean distance. For these points show the approximate hierarchical clustering using (i) single linkage (i.e. distance between clusters is the minimum distance between points across clusters) and (ii) max linkage (i.e. distance between clusters is the maximum distance between points across clusters).



7. [15 points]: Simulate the tossing of three types of coins (a) a fair coin; (b) a coin with p(H) = 0.4 and (c) a coin with p(H) = 0.6. An experiment involving 10-coin tosses is done with each coin. Let X denoting the number of coin tosses that results in Heads. Please answer the following questions:

(i) [4 points] What is the mean(X), and Standard deviation of X in each situation?
(ii) [4 points] How would the distribution of the random variable X for these 3 coins be different?
(iii) [4 points] What is P|X-mean(X)| < 2standard deviation(X) in these three distributions?
(iv) [3 point] If you want to minimize the standard deviation of X what would you do?

8. [5 points] I am interested in estimating a population parameter from a population involving k categories of people by using random sampling as well as stratified sampling, please explain how stratified sampling results in better prediction accuracy when compared to randomized sampling? What is the trade-off involved in choosing between randomized sampling and stratified sampling?