School of Engineering and Applied Science (SEAS), Ahmedabad University

Name: _____ Roll no.: _____

MAT502 Advanced Statistics
END Term Exam (Winter Sem 2025);
Marks 40, Weightage: 40 %
02 May 2025, Friday, 14.00 – 16.00

Instructions:

1 All questions are compulsory.
2 Figures to right indicate full marks.
3 All acronyms and symbols carry their usual meaning.
4 Assume suitable data wherever necessary and specify it clearly in your answers.

Q:1 **Solve following problems.** **[08]**

(a) Urn A contains 2 white, 1 black and 3 red balls; Urn B contains 3 white, 2 black and 4 red balls and Urn C contains 4 white, 3 black and 2 red balls. One urn is chosen at random and 2 balls are drawn at random from the urn. If one chosen balls happens to be red and black what is the probability that both balls have come from urn B? **(04)**

(b) The distribution of number of road accidents per day in a given city is estimated by Poisson distribution with mean equal to 4. Find the number of days out of 100 days that **(04)**

(i) there will be no accident

(ii) at least two accidents

(iii) at most 3 accidents

Q:2 **Do as directed: (Any two)** **[08]**

(a) Enlist the reasons of usage of Non-probability sampling methods in different surveys. Enlist popular types of Non-probability sampling methods. Explain any one of Non-probability sampling methods in brief. **(04)**

(b) Explain the concept of Null hypothesis. Statistically, when two hypothesis are being evaluated, can any one of them be given the name 'null hypothesis? Give reasons for your answer. Discuss Type I and Type II errors in context of evaluation of hypothesis. **(04)**

(c) A personnel manager has received complaints that the stenographers in the company have become slower and do not have the requisite speeds in stenography. The Company expects the stenographers to have a minimum speed of 90 words per minute. The personnel manager decides to conduct a stenography test on a random sample of 15 stenographers. However, he is clear in his mind that unless the sample evidence is strongly against it, he would accept that the mean speed is at least 90 w.p.m. After the test, it is found that the mean speed of the 15 stenographers tested is 86.2 w.p.m. What should the personnel manager conclude at a significance level of 5%, if it is known that the standard deviation of the speed of all stenographers is 10 w.p.m. No separate table is provide. You may assume suitable data /values needed if any and clearly mention them. **(04)**

Q:3 Answer the following questions as directed. [24]

(a) Consider following pattern points with their (X,Y) coordinates in in a given 2-D space. **(06)**

Point No	X	Y	Point No	X	Y
1	6	6	**6**	5	4
2	2	3	**7**	7	7
3	3	3	**8**	7	5
4	6	7	**9**	2	4
5	4	5	**10**	4	2

Using Vector quantization technique assume (case 1) where that threshold distance equal to 2.5; determine the clusters with following parameters :

(i) Number of clusters (ii) Cluster centres (iii) Cluster members for all clusters

In above example, consider another case (i.e. Case 2) when threshold distance is changed to 4 compute above parameters again. Critically comment on the charrachristics of both cases in order to classify points.

(b) For a certain application 10 different instances attributes and classification are given as under, USE Decision tree (ID3) algorithm to compute entropies and information gain and create decision tree. **(06)**

instance	Attribute –a1	Attribute –a2	Attribute –a3	Classification
1	True	Hot	High	No
2	True	Hot	High	No
3	False	Hot	High	Yes
4	False	Cool	Normal	Yes
5	False	Cool	Normal	Yes
6	True	Cool	High	No
7	True	Hot	High	No
8	True	Hot	Normal	Yes
9	False	Cool	Normal	Yes
10	False	Cool	High	Yes

(c) Explain 'Hypothesis' and 'Hypothesis space' in brief. For the evaluation of the given hypothesis, explain trade-off between 'fitting' and 'generalisation' **(02)**

(d) Explain need of 'Feature reduction'. Enlist and explain in brief, types of feature reduction techniques. **(02)**

(e) Consider the dataset with six data points $\{(x_1,y_1), (x_2,y_2)\ldots (x_6,y_6)\}$ Where, $x_1= (1,0)$, $x_2= (0,1)$, $x_3= (0,-1)$, $x_4= (-1,0)$, $x_5= (2,2)$, $x_6= (-2,-2)$ and labels are given by $y_1=y_2=y_5=1$ and $y_3=y_4=y_6=-1$. A hard margin support vector machine is trained on the above dataset. Which one of the following datasets is possible set of support vectors? **(02)**

Note: Justify your answer with reasoning else no marks will be awarded.

(i) $\{x_1,x_2,x_5\}$

(ii) $\{x_3,x_4,x_5\}$

(iii) $\{x_4,x_5\}$

(iv) $\{x_1,x_2,x3,x_4\}$

(f) A researcher plans to have improved performance of a neural architecture to be built for an application. Previous studies for similar neural architectural problems suggest that in general, network will be complex. Critically evaluate role of following heuristics to enable the researcher for considerations for his/her plans: **(06)**

(i) Ensemble learning

(ii) Early stopping

(iii) Drop out

All the best
