Total No. of Pages 02 Roll No. ..... **Set B**

FIFTH SEMESTER **B.Tech. (CSE)**

MID SEMESTER EXAMINATION ***Sept, 2024***

CO327 MACHINE LEARNING

Time: 1:30 Hours *Max. Marks: 20*

Note: Answer ***ALL*** questions.
Assume suitable missing data, if any.
CO# is course outcome(s) related to the question.

1. For each of the following scenarios, determine whether machine learning is a suitable approach. If it is suitable, recommend the specific type of machine learning model (e.g., supervised, unsupervised, reinforcement learning, etc.) and justify your choice based on the problem's structure, data characteristics, and real-world constraints (such as scalability, interpretability, or data limitations). If machine learning is not suitable, propose a more appropriate non-ML method and explain your reasoning.

Supervised Unsupervised NO ML

[a] Predicting crop diseases in a large-scale agricultural setting using images from drone-mounted cameras and environmental data where disease occurrences are sporadic and data is imbalanced.

Unsupervised

[b] Classifying types of soil based on chemical composition and texture, but the available data is limited and lacks labeled examples.

Reinforcement

[c] Optimizing irrigation schedules in a smart farm where sensor data is available in real-time, computational resources are constrained, and decisions need to be interpretable by farmers.

Supervised OR. NO model unsupervised

[d] Modeling the growth of a plant species under controlled greenhouse conditions with predefined growth stages and no variability in environmental factors. **[1+1+1+1] [CO1]**

2. An energy management company is developing a system to detect abnormal energy consumption in households (Positive class: Anomaly, Negative class: Normal Consumption). The system monitors thousands of households daily and flags certain households as potentially having abnormal energy usage. After running the system for a day on 15,000 households, the following confusion matrix is obtained:

Table 1: Confusion Matrix

	Predicted: Anomaly	**Predicted: Normal**
Actual: Anomaly	120	30
Actual: Normal	200	14,650

[a] Calculate the precision and recall of the abnormal energy consumption detection system. **[2] [CO3]**

[b] If the company's goal is to identify as many households with abnormal energy consumption as possible, even at the risk of flagging some normal households, which metric, precision or recall, should be prioritized? Justify your answer. **[2] [CO4]**

3. A financial services company is building a logistic regression model to predict whether a customer is likely to default on a loan (outcome: y=1, default, y=0, no default) based on the following features: Income (x_1, in thousands of ₹), Credit Score (x_2), and Number of Dependents (x_3). The coefficients of a trained model are given as $w_0 = 1.0, \boldsymbol{w} = [-0.02, 0.03, -0.1]^T$. The company uses a threshold of $P(y = 1) = 0.5$ to o classify customers as default or not.

[a] Consider a customer with Income = ₹50,000, Credit Score = 700, and Number of Dependents = 2. Will this customer be classified as likely to default or not? **[2] [CO4]**

[b] How might adjusting the threshold to $P(y = 1) = 0.3$ affect the false positive and false negative rates? Explain the trade-off. **[2] [CO4]**

[c] The company is concerned about overfitting since it only has a small dataset. Explain how adding L2 regularization (Ridge Regression) would affect the model, particularly the coefficients. How would this technique help prevent overfitting? **[2] [CO4]**

4. A technology blog wants to classify emails into two categories: Promotions and Updates. The blog uses a Naive Bayes classifier to predict the category of a email based on certain keywords. The dataset consists of the frequency of these keywords in emails that are labeled as either Promotions or Updates. The probability of an email being in the Promotions class is 60%.

Table II: Keyword Frequencies in Email Categories

Category	~~AI~~ Discount	~~Game~~ Sale	~~Robot~~ News	~~Player~~ Offer	~~Score~~ Update
~~Technology~~ Promotion	50	10	45	5	3
~~Sports~~ Update	5	40	2	60	55

[a] Given a new email with the following keywords: Discount, Sale, and New, classify the email category using the Naive Bayes classifier. **[3]**

[b] If the classifier prioritizes reducing false positives (incorrectly marking an Update email as Promotions), how should the prior probability for Promotions be adjusted to reflect this goal? Explain how changing the prior probability will affect the classifier's behavior and its impact on false positives. **[2+2] [CO3, CO4]**

---Best of Luck---

Q2:

		Predicted +	Predicted −
Actual	+	120 (TP)	30 (FN)
Actual	−	200 (FP)	14650 (TN)

a) Precision $= \frac{TP}{TP+FP} = \frac{120}{120+200} = \frac{120}{320} = 0.375$ (½)

Recall $= \frac{TP}{TP+FN} = \frac{120}{120+30} = \frac{120}{150} = 0.8$ (½)

2 marks

b) **Recall**

Among all the actual +ve, we want to find which are predicted +ve, even at the risk of including -ves. 2

Q3:

$y=0$ no default — $y=1$ defaulter

a).

$$Z = -0.02x_1 + 0.03x_2 - 0.1x_3 + 1.0$$
$$= -0.02 * 50{,}000 + 0.03 * 700 - 0.1 * 2 + 1.0$$
$$= -1000 + 21 - 0.2 + 1.0 = -978.2$$

1 mark

assuming 50

$Z = 20.8$

$g(z) = \frac{1}{1+e^{-20.8}}$

$= \frac{1}{1+0} = 1$

defaulter

$$g(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{978.2}} = \frac{1}{1+(2.72)^{978.2}} \sim 0$$

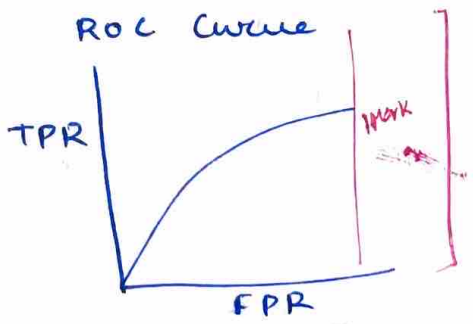
1 Mark

not defaulter 1 Mark

b) $TPR = \frac{TP}{TP+FN}$

$FPR = \frac{FP}{FP+TN}$

1 mark



2 mark

FP↑ FN↓

c). Ridge Regression

$$L = \frac{1}{n}\sum_{i=1}^{m}(y_i - \hat{y}_i)^2 + \lambda \sum_{i=1}^{n} w_i^2$$

Add squared magnitude of the coefficient as penalty term. — 2 mark

Q4 a)

only formula: ½

$P(\text{Promotion} \mid \text{Discount, Sale, New}) =$

$$P(\text{Dis} \mid \text{Promotion}) \cdot P(\text{Sale} \mid \text{Promotion}) \cdot P(\text{New} \mid \text{Promotion}) \cdot P(\text{promotion})$$

1 mark

$$= \frac{50}{113} \cdot \frac{10}{113} \cdot \frac{45}{113} \cdot \frac{60}{100} = \frac{13500}{1442897} = 0.009356$$

$P(\text{Update} \mid \text{Discount, Sale, New}) =$

$$P(\text{Discount} \mid \text{update}) \cdot P(\text{Sale} \mid \text{update}) \cdot P(\text{New} \mid \text{update}) \cdot P(\text{update})$$

1 mark

$$= \frac{5}{162} \cdot \frac{40}{162} \cdot \frac{2}{162} \cdot \frac{40}{100} = \frac{160}{4251528}$$

$$= 3.76 \times 10^{-5}$$

Denom wrong: ½
Probability multiply mis: 1
Coding mistake in Denom: 2½

Promotion — 1 mark

b) Prior probability for promotion should be reduced if mails are being incorrectly marked as promotions. — 2 mark

if PP of promotion is high: incorrectly marked as promotio
PP of update is high: incorrectly marked as updat
So 0.5 should be a good no. to ensure balance.