SOUMIK ROY 2260074

CSBS 3rd Year 6th Sem.

Assignment: MACHINE LEARNING
CSBS 3232

Date: 07/05/25

① Briefly explain the contexts where logistic regression is used.

Ans: Logistic regression is used for:

→ Binary classification tasks, such as spam detection (spam or not), disease prediction (disease or not) or fraud detection.

→ Medical diagnosis (eg. if a tumor is malignant or benign)

→ Customer behaviour modeling, like churn prediction (will the customer leave or stay).

→ Credit scoring to determine if a customer is likely to default.

② Explain the significance of decomposing a multi-class classification problem into binary classification tasks.

Ans:

Decomposing multi-class problems into binary-tasks helps

- Simplify complex classification into manageable binary decisions.
- Improving traning efficiency and performance using methods like One-vs-Rest or One-vs-One.
- Allow the use of simpler binary classifiers that may generalize better on limited data.

④ Apply the Least Square Method on the given data set to predict the output value for the test value of $x = 4.5$.

X	1	2	3	4	5
Y	3	7	8	12	15

Ans:-

Given:-

x: 1, 2, 3, 4, 5

y: 3, 7, 8, 12, 15

∴ slope (m) and (c) intercept:

$$m = \frac{n\Sigma XY - \Sigma X \Sigma Y}{n\Sigma X^2 - (\Sigma X)^2}, \quad c = \frac{\Sigma Y - m\Sigma X}{N}$$

$\Sigma X = 15, \quad \Sigma Y = 45, \quad \Sigma XY = 1\times3 + 2\times7 + 3\times8 + 4\times12 + 5\times15$

$= 164$

$\Sigma X^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 55, \quad n = 5$

$\therefore m = \frac{5\times164 - 15\times45}{5\times55 - 15^2} = \frac{820 - 675}{275 - 225}$

$= \frac{145}{50} = 2.9$

$\therefore c = \frac{45 - 2.9\times15}{5} = \frac{45 - 43.5}{5} = \frac{1.5}{5} = 0.3$

$\therefore \quad y = 2.9x + 0.3$

$\therefore$ Prediction for $x = 4.5$

$\Rightarrow \quad 2.9\times 4.5 = 13.05 + 0.3 = 13.35$

(5) Explain how a higher VC dimension affects a model's ability to generalize. What happens when the VC dimension is too high or too low relative to the data set size?

Ans:- VC (Vapnik-Chervonenkis) Dimension measures the capacity of a model class to shatter data points i.e. correctly classify all label combinations.

High VC Dimension can lead to of overfitting i.e model too complex for the data whereas Low VC Dimension may lead to underfit i.e model too simple. Ideal VC Dimension should grow with data size, but not too quickly, to maintain generalization.

(6) For a logistic regression model with two input features x_1, x_2, the weight, the weight vector is given as $w = [2, -1]$ and bias $= 0.5$. If a new data point $(x_1, x_2) = (3, 2)$ is given, compute the output of the logistic function determine whether the model classifies it as class 0 or class 1.

Ans:

[P.T.O]

Ans:

Given:

$w = [2, -1]$, $b = 0.5$

$$z = w_1 x_1 + w_2 x_2 + b = 2 \times 3 + (-1) \times 2 + 0.5$$
$$= 6 - 2 + 0.5$$
$$= 4.5$$

∴ Logistic Function,

$$\sigma(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{-4.5}} \approx \frac{1}{1+0.0111}$$
$$= 0.989$$

Since, $0.989 > 0.5$, the model classifies the point as class 1.

(7) Define the VC Dimension of a hypothesis's class. How does it help in understanding the learning capacity of a model?

Ans:

VC Dimension is the maximum number of data points that a hypothesis can shatter. It reflects the capacity or complexity of a model class.

It helps in understanding:

- Learning bounds: • VC dimension helps determine how many training examples are needed for a model to learn accurately
  • A higher VC dimension implies that more training data is needed to ensure low generalization error.
- Generalization Ability:
  • It helps predict how well a model trained on limited data will perform on new, unseen data.
  • A balanced VC dimension improves the ability to generalize from training to testing data.
- Overfitting / Underfitting Risk:
  • A very high VC dimension may mean the model is too flexible → overfitting.
  • A very low VC dimension may mean the model is too simple → Underfitting.
  • Choosing a model with an appropriate VC dimension helps find the right trade-off.