Lab - 4.

Consider a binary classification problem where we want to predict whether a student will pass or fail based on their study hours. The logistic regression model has been trained, and the learned parameters are $a_0 = -5$ (intercept) and $a_1 = 0.8$ (coefficient of study hours).

(a) Write the logistic regression equation for this problem.

$$p(y=1/x) = \frac{1}{1+e^{-(-5+0.8x)}}$$

(b) Calculate the probability that a student who studies for 7 hours will pass.

Substitute $x = 7$.

$$z = -5 + 0.8 \times 7 = 0.6$$

$$p(\text{pass}) = \frac{1}{1+e^{-0.6}}$$

$$= 0.6479.$$

(c) Determine the predicted class for this student based on threshold of 0.5,

if $p(\text{pass}) \geq 0.5$ student will pass else he will fail

2] Consider $z = [2, 1, 0]$ for three classes, Apply softmax function to find probability of values of three classes.

$$\text{Softmax}(z_k) = \frac{e^{z_k}}{\sum_{i=1}^{K} e^{z_i}}$$

$$P(1) = \frac{e^2}{e^2+e^1+e^0} = 0.665$$

$$P(2) = \frac{e^1}{e^2+e^1+e^0} = 0.295$$

$$P(3) = \frac{e^0}{e^2+e^1+e^0} = 0.09$$

Answer:

After building the logistic regression models, write the answer for the following questions in your observation book.

1. For dataset file "HR-comma-sep.csv".

(i) which variable did you identify as having a direct and clear impact on employee retention? why?

- Satisfaction level
- Time spent in company.
- Number of projects
- Salary

These variables are choosen based on trends in data visualization

(ii) what was the accuracy of your logistic regression model? Do you think this is a good accuracy? why or why not?

The accuracy of logistic regression was 78.1%. This accuracy is fairly good, It suggest that the model ~~capturing~~ captures most of the properties of hiring employee intension.

2) For Zoo dataset.

(i). Did you perform any data preprocessing steps If yes, what were they, and why are they necessary. → Yes.

* Dropped 'animal_name' column
* checked for missing values.
* Converted categorical variable if needed.

(ii). were there are any missing or inconsistent values in dataset? How did you handle them?

No missing values found in the dataset.

→ If incase of inconsistency, we could have used mean/mode imputation or removed them.

(iii). what does the confusion matrix tell you about the performance of your model

→ It shows how well the model predicted different class types.

→ A high number of correct predictions, along the diagonal of the matrix indicates good performance

(iv) Which class types were most frequently misclassified? why do you think this happened?

→ amphibians, birds/reptiles

Reasons :
- since they have similar features.
- logistic regression assumes linear decision boundaries, which may not work well for complex class separation.

KNN (K - nearest neighbours) :

Consider the following dataset, for k=3 and test data (x, 35, 100) as a [Person, Age, Salary] and predict the target.

Person	Age	Salary	Distance	Rank	Target
A	18	50	52.8		
B	23	55	46.6		
C	29	70	31.9	2	N
D	41	60	40.4	3	Y
E	43	70	31.1	1	Y
F	58	90	60.1		

Step 1 : Distance (d) = $\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

$(x_2, x_1) = (35, 100)$

$d_1 = \sqrt{(35-18)^2+(100-50)^2} = 52.8$

$d_2 = \sqrt{(35-23)^2+(100-55)^2} = 46.6.$

Step 2 : Identify 3 nearest neighbours

1) E (31.1, Y)
2) C (31.9, N)
3) D (40.4, Y)

Step 3 : majority voting

Since 2 out of 3 belong to class 'Y'

∴ the predicted class for x(35, 10) is 'Y'

Random Forest.

16/04/2025

1] Start with the data.

Input

Training dataset D with n example and m features
Number of trees T to create

2] For each tree t from 1 to T

Step 1: Randomly select a bootstrap sample of size n from Dataset D.
This subset is used to train tree t.

Step 2: For each node in the tree.

* Randomly select k features from total m features
* choose the best feature among k features to split the node based on a criterion
* Repeat until tree reaches the stop condition.

3] For each test sample x:

Pass x through each of the T decision tree T to get a prediction.

→ If it's a classification program each tree votes for a class
→ If its a regression problem, each tree provides a predicted value

④ Output:

For classification: The final prediction is the majority vote of all T trees

For regression: the final prediction is the average of predictions for all T trees.

5] End.

16.04