LAB-4

Logistic Regression

1) 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 $a_0 = -5$ and $a_1 = 0.8$

a) Write logistic regression equation of the 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 the probability of value of three classes

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

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

$$P(2) = \frac{e^1}{e^2+e^1+e^0} = 0.245$$

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

1) For dataset file "HR_comma_sep.csv"

i) Which variables 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 were chosen 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%. This accuracy is fairly good. It suggests that the model captures most of properties affecting employee retention

2) For zoo dataset

i) Did you perform any data preprocessing steps? if yes, what are they? and why were they necessary?

• Dropped the animal_name column
• checked for missing values
• converted categorical variables if needed.

ii) Were there any missing or inconsistent values in dataset? How did you handle them

No missing values were found in the dataset. If there were inconsistencies we could have used mean/mode imputation

iii) what does the cofusion matrix tell you about the performance of model

→ Confusion matrix showed 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?

→ The most frequently misclassified classes were likely amphibians, birds or reptiles as they share similar features

→ possible reasons are feature overlap and simplified model