02/04/25

LAB-4
Logistic Regression

Exercise-1

① Consider a binary classification problem where want to predict whether a student will pass or fail based on their study hours. The model is trained and learned parameters are $a_0 = -5$ & $a_1 = 0.8$

a. Write logistic regression equation

b. Calculate probability that student who studies for 7 hours will pass.

c. Determine the predicted class (pass or fail) for this student based on a threshold of 0.5.

(A) ⓐ Logistic Regression Equation is:

$$P(\text{Pass}) = \frac{1}{1+e^{-(a_0+a_1 \cdot x)}}$$

ⓑ
$$P(\text{Pass}) = \frac{1}{1+e^{-(-5+0.8\times7)}}$$

$$= \frac{1}{1+e^{-0.6}}$$

$$= \frac{1}{1+0.548} \approx 0.645$$

∴ It is approximately 64.5%.

ⓒ Since $P(\text{Pass}) = 0.645$ which is greater than threshold, therefore predicted class is "Pass".

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

(A) The softMax function for vector z is given by :-

$$P_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

Computing probabilities for each class:

$$P_1 = \frac{e^2}{11.107} = \frac{7.389}{11.107} \approx 0.665$$

$$P_2 = \frac{e^1}{11.107} = \frac{2.718}{11.107} \approx 0.245$$

$$P_3 = \frac{e^0}{11.107} = \frac{1}{11.107} \approx 0.090$$

Exercise - 2

(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?

(A) key variables for Employee Retention

- Satisfaction Level :- Lower satisfaction increases attrition
- Number of Projects & Average Monthly Hours :- Overworking leads to burnout.
- Promotion and Salary :- Lack of growth opportuni impact retention.

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

(A) Achieved accuracy is 78.43%.
Accordingly, it is an accuracy which is good and key patterns is been captured successfully.

(2) For Zoo dataset

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

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

(iii) What does the confusion matrix tell you about the performance of your model?

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

(A) (i) Data Preprocessing

- Removed animal_name.
- standardized numerical features.
- split dataset (80% train, 20%. test).

(ii) Handling Missing Data

- No missing or inconsistent values found

(iii) Confusion Matrix Insights

- Achieved 100% accuracy, no misclassifications.

(iv) Misclassified class Types

- None due to well-separated data features.