02/04

# LAB - 03

# LOGISTIC REGRESSION

Q1. Consider a binary classification problem where we want to predict whether a student will pass or fail ...

Given $a_0 = -5$ (intercept)

$a_1 = 0.8$ (coefficient)

a) Write the logistic regression equation for this problem.

$$\text{sigmoid}(z) = P(x) = \frac{1}{1+e^{-z}} \qquad z = a_0 + a_1 x.$$

$$z = -5 + 0.8x \qquad P(x) = \frac{1}{1+e^{-(-5+0.8x)}}$$

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

$$\text{Probability}(x \mid \text{pass}) = \frac{1}{1+e^{5-0.8(7)}} = \frac{1}{1+e^{-0.6}}$$

$$= 0.6457$$

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

If threshold $= 0.5$

$P(x = 7) > 0.5$

Student will $\boxed{\text{pass}}$

Q2. Consider z = [2,1,0] for three classes. Apply softmax function

$$\boxed{\text{softmax}(z_k) = \frac{e^{z_k}}{\sum_{j=1}^{k} e^{z_j}}}$$

$$\text{softmax}(z_1) = \frac{e^2}{e^2+e^1+e^0} = \frac{7.39}{7.39+2.72+1} \approx 0.665$$

$$\text{softmax}(z_2) = \frac{e^1}{e^2+e^1+e^0} = \frac{2.72}{7.39+2.72+1} \approx 0.244$$

$$\text{softmax}(z_3) = \frac{e^0}{e^2+e^1+e^0} = \frac{1}{7.39+2.72+1} \approx 0.095$$

Probabilities for three classes = 66.5%, 24.4%, 9.1%.

Q. HR_comma_sep. csv

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

variables such as satisfaction_level, average_monthly_hours, number_project, time_spend_company, salary.

- satisfaction level : strong negative correlation
- average_monthly_hours : strong positive correlation
- number_project : positive correlation
- time_spend_company : strong positive correlation
- salary : strong negative correlation

(ii) What was the accuracy of your logistic regression model?

The accuracy was around (0.79) (79%

Q. Zoo.csv

(1) Did you perform any data preprocessing steps? If yes, what were they?

① Handling categorial data

class_type → class_mapping

② Splitting dataset

80% training, 20% testing

③ Drop irrelevant columns

Drop animal_name, ~~class~~ as it is irrelevant.

④ Feature Scaling .

standardScaler() to standardize numerical features        mean = 0

variance = 1

(ii) Were there any missing or inconsistent values in the dataset ?

Ans        No missing values.

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

The confusion matrix represents how well the model classifies differ classes.

Diagonal values : ~~True~~ Correct classification

Other values : Incorrect classification

(iv) Which class types were most frequently misclassified ?

Ans        Class 3 , 5 ,and 6 were misclassified .

Possible reasons

① Feature similarity

② Small sample size