Lab-4

14/12/25

Logistic Regression

Bafna Gold
Date Page

1) Logistic Regression for Binary Classification

Q1) 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 & the learned parameters are $a_0 = -5$ (intercept) & $a_1 = 0.8$ (coefficient for study hours).

a) Write a logistic regression equation for this problem.

→ $$P(pass) = \frac{1}{1+e^{-(a_0+a_1 x)}}$$

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

→ $$P(pass) = \frac{1}{1+e^{-(-5+0.8\times 7)}}$$

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

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

∴ It is approximately 64.5%.

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

→ Since $P(pass) = 0.645$ which is greater than threshold, therefore predicted class is "Pass"

2) Consider $z = [2, 1, 0]$ for 3 classes. Apply softmax function to find the probability values of 3 classes.

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

Computing probabilities for each class:

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

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

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

2) Multiclass Classification:

Qn. i) For dataset files "HR_comma_sep.csv"

i) Which variables did you identify as having a direct and clear impact on employee retention? Why?

→ * satisfaction_level → lower satisfaction increases attrition.

* number_project and avg_monthly_hours → Overworking leads to burnout.

* promotion_last_5years and salary → lack of growth opportunities impacts retention.

ii) What was the accuracy of your model? Do you think this is a good accuracy? Why or Why not?

→ Accuracy → 76.48%

→ It is good since it captures key patterns & relationships.

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?

→ 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.

9/4/25