Date 19/25

Lab 4

Logistic regression

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

2) write a logistic regression equation

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

where x = study hours

$$P(y=1|x) = \frac{1}{1+e^{-(-5+0.8 \times \text{study hrs})}}$$

probability =

$$P(y=1|x=7) = \frac{1}{1+e^{-(-5+0.8 \times 7)}}$$

$$= 0.6956$$

Given threshold = 0.5

P (Student) = 0.6956

P > threshold,

Student is in pass class

3) Given values $z = [2, 1, 0]$

Ans $e^z \Rightarrow [e^2, e^1, e^0]$

$\Rightarrow [7.389, 2.718, 1]$

$\Sigma e^z \Rightarrow 11.107$

Probabilities $\Rightarrow [0.665, 0.249, 0.09]$

=> In dataset file "HR_comma_sep.csv"

1) which variables did you identify as having a direct and clear impact on employee retention? why?

Ans Key variable are

=> Satisfaction level => Strong negative correlation

=> Time spent in company => Positive correlation

=> Work accident => Negative correlation

=> Salary => lower salary tends to more leaves

=> Department => Retention varies for departments

=> Model accuracy => Accuracy is 78.58% while learning and indicates if our imbalanced dataset, many leaves are missed.

Date:
Page No.:

6. Class ~~separable~~ ~~was~~ big was misclassified once. It could be because
a) Limited amount of data
b) Feature similarity in big and blue class.