

Lab-4

Logistic Regression

- (1) Binary classification: $a_0 = -5$ (intercept) and $a_1 = 0.8$ (coeff for study hours)

$$(a) \quad y = 0.8 + (-5)x \quad \cdot \quad \sigma = \frac{1}{1 + e^{-x}}$$
$$y = -5 + (0.8)(x)$$

$$(b) \quad x = 7, \quad y = -5 + (0.8)(7)$$
$$= -5 + 5.6$$

$$y = -5 + (0.8)(7) = -5 + 5.6 = 0.6$$

$$\sigma(x) = \frac{1}{1 + e^{-0.6}} = 0.64$$

$$\text{threshold} = 0.5 \quad \therefore 0.64 > 0.5$$

\therefore the student passes.

- (2) Multiclass classification:

$$x = [2, 1, 0]$$

$$p_2 = \frac{e^2}{e^2 + e^1 + e^0} = 0.64$$

$$p_1 = \frac{e^1}{e^2 + e^1 + e^0} = 0.244$$

$$p_0 = \frac{e^0}{e^2 + e^1 + e^0} = 0.099$$

→ HR dataset (binary)

(i) Satisfaction level, salary, time spent at the company, work accidents, department

(ii) Accuracy is 95.67%.

Reasonably good but not perfect.
Because the factor might be affected by factors outside the dataset.

→ EOR dataset (multiclass)

(i) animal column was dropped, as it was not useful for classification, just an identifier.
encoded class-type using LabelEncoder

(ii) Categorical variables converted to numeric

(iii) No missing values

(iv) All the numbers are in diagonal hence therefore no misidentification of classes.

(v) No class types were misclassified.

lab-4

KNN (K-Nearest Neighbors)

Person	Age	Salari	k	Target	distance	Rank
A	18	50	N		$39.82 \sqrt{914}$	1 5
B	23	55	N		$46.07 \sqrt{2169}$	5 4
C	24	70	N		$31.9 \sqrt{1021}$	2 2
D	41	60	Y		40.44	3 3
E	43	70	Y		31.04	1 1
F	38	40	Y		60.07	6 6
X	35	100	?			

$$\rightarrow \sqrt{(38-35)^2 + (40-100)^2} = \sqrt{3^2 + 60^2} = \sqrt{9 + 3600} = 60.07$$

$$\sqrt{(43-35)^2 + (70-100)^2} = \sqrt{8^2 + 30^2} = \sqrt{964} = 31.04$$

$$\sqrt{(41-35)^2 + (60-100)^2} = \sqrt{6^2 + 40^2} = \sqrt{1636} = 40.44$$

k=3 \rightarrow N Y ~~Y~~

\rightarrow ~~Y~~ Y

1 is dataset

\rightarrow How to choose k value?

\rightarrow testing multiple k values and computing their accuracy. The accuracy and error rate for the k's are compared and the most optimal k is selected.
here k=3.

diabetes dataset:

- 0) What is the purpose of feature scaling? How to perform it?

→ Feature scaling ensures all features contribute equally to the nearest neighbors. Scaling is done so that features like Glucose / age don't dominate the one with the smaller range.