classmate
Date 23/02/24
Page 1

Logistic Regression

It is a supervised ML algorithm which is used to perform binary classification. Eg. To predict whether an email is spam or non-spam or whether a person is healthy or unhealthy, etc.

Logistic regression is a **linear model**. As we all know that linear models have an assumption that their residuals should be **normally distributed** and predictions($\hat{y}$) should be **continuous** in nature and should be **unbounded** on either sides of the number line.

But to classify the output of this linear model as a category, we can use different approaches.

(a) ✗ **probability** → The probability of an event occuring is given as 'p' and not occuring is '1-p', but since the values will lie between 0 to 1 always it is continuous but bounded on both sides of the number line, hence we can't use it.

(b) ✗ **odds** → The odds of an event occuring is given by $\left(\frac{p}{1-p}\right)$, where p is probability.

Odds is also continuous, but its values can't be negative hence it is bounded on negative side, so we can't use it.

(c) ✓ **log-odds** → It is found by taking log of the odds. and its formula is given as—

$\log\left(\frac{p}{1-p}\right)$, It is continuous and unbounded on both sides and hence can be used to model linear relationships.

(d) We can even use other approaches like — probit, square root of arc sin and complimentary log-log. but log-odds has been found out to be giving best results.

In statistics, log of odds is also referred as **logit functions**.

Getting the probability back

We can use the log-odds as a linear function of the predictors.

$$\log\left(\frac{p}{1-p}\right) = \theta_0 + \theta_1 x$$

$$\Rightarrow \frac{p}{1-p} = e^{\theta_0+\theta_1 x}$$

$$\Rightarrow p = (1-p)e^{\theta_0+\theta_1 x}$$

$$\Rightarrow p = e^{\theta_0+\theta_1 x} - pe^{\theta_0+\theta_1 x}$$

$$\Rightarrow p + pe^{\theta_0+\theta_1 x} = e^{\theta_0+\theta_1 x}$$

$$\Rightarrow p\left(1+e^{\theta_0+\theta_1 x}\right) = e^{\theta_0+\theta_1 x}$$

$$\Rightarrow p = \frac{e^{\theta_0+\theta_1 x}}{1+e^{\theta_0+\theta_1 x}} \leftarrow \text{probability of positive class}$$

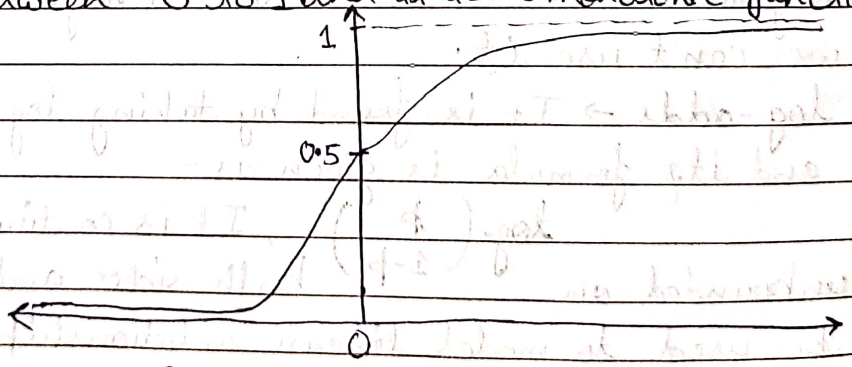
Let's assume $\theta_0+\theta_1 x = X$, then

$$\Rightarrow p = \frac{e^X}{1+e^X}$$, dividing numerator and denominator by e^X.

$$\Rightarrow p = \frac{1}{1+e^{-X}} \Rightarrow p = \frac{1}{1+e^{-(\theta_0+\theta_1 x)}}$$

Intrestingly, $\frac{1}{1+e^{-X}}$ is nothing but a **sigmoid/ logistic function**.

* It is the property of sigmoid function that given any real-value it squeezes its output between 0 to 1 and it is a monotonic function.



$$y = \theta_0 + \theta_1 x$$

$$\log\left(\frac{p}{1-p}\right) = \theta_0 + \theta_1 x$$

When, $y = 0$, then $p = 0.5$
$y = -\infty$, then $p = 0$
$y = \infty$, then $p = 1$

$$\boxed{p = \frac{1}{1+e^{-y}}}$$

→ In logistic regression, one beta is learnt per feature.

Imp:

Ques: Sigmoid function is a non-linear relationship but we still say logistic regression is a linear model. Why?

Ans: It is linear model because the **logit / log of odds** of the estimated probability response is a linear function of the predictors.

Loss function in case of logistic regression:

In linear regression, we used OLS or MSE as a loss function, but in case of logistic regression we can't use a MSE loss function because, the prediction ($\hat{y}$) comes out from non-linearity applied to logit / log-odds, hence the loss function becomes **non-convex**, hence we will not be able to find the global minima where loss will be minimum.

Therefore, for logistic regression we use a special kind of loss function known as **log-loss / cross-entropy**.

Given n data points, the log-loss / cross entropy is given as —

$$L_{log}(y, \hat{p}) = -(y \log(\hat{p}) + (1-y) \log(1-\hat{p}))$$

$$L_{log}(y, \hat{p}) = \begin{cases} \text{when } y=0, & \log(1-\hat{p}) \\ \text{when } y=1, & \log(\hat{p}) \end{cases}$$

The log-loss is a convex function, and its global minima can be found easily.

If we just remove the negative sign attached in front of the log-loss equation, it becomes **log-likelihood**. So, We can either minimize the logloss or we can maximize the log-likelihood.

→ No closed form solution exists for finding the optimal values of coefficients in case of logistic regression.

* Extension of logistic regression for multi-class classification—

In multi-class classification, we make use of the **softmax function** to get probabilities back from response variable 'y'.

Probability of class i, given k classes is given as—

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \text{for } i = 1, \dots, K \text{ and } z = (z_1, z_2, \dots z_K).$$

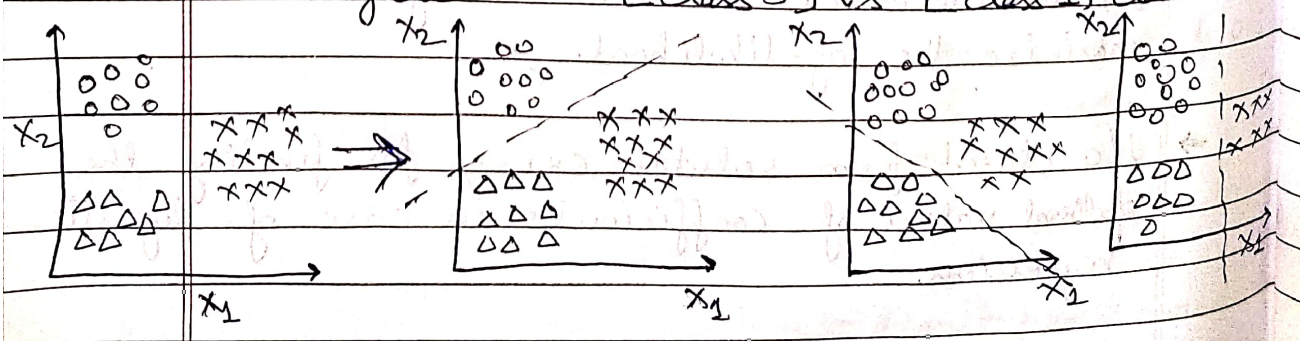
→ In multi-class classification, one theta(θ) is learnt per feature per class. Every class has it's own hyperplane.

* Other way of solving multi-class classification problem is converting it into binary classification problem—

(a) **One Vs Rest (OVR)** classification method—

In this, if we have 'n' classes, then we need to train 'n' binary classifiers i.e., the number of class labels present in the dataset and the number of generated binary classifier will be same. Suppose, we have three classes in a dataset and we want them to classify then, One Vs Rest / One Vs All will work as—

Classifier 1 → [class 1] vs [class 2, class 3]
classifier 2 → [class 2] vs [class 1, class 3]
classifier 3 → [class 3] vs [class 1, class 2]



Suppose, we have dataset in which we want to classify three different colours given as —

X_1	X_2	X_3	Y
a	b	c	Red
d	e	f	Blue
g	h	i	Green

three models will be created as —

①

X_1	X_2	X_3	Y
a	b	c	+1
d	e	f	-1
g	h	i	-1

②

X_1	X_2	X_3	Y
a	b	c	-1
d	e	f	+1
g	h	i	-1

Here, Blue treated as positive class.

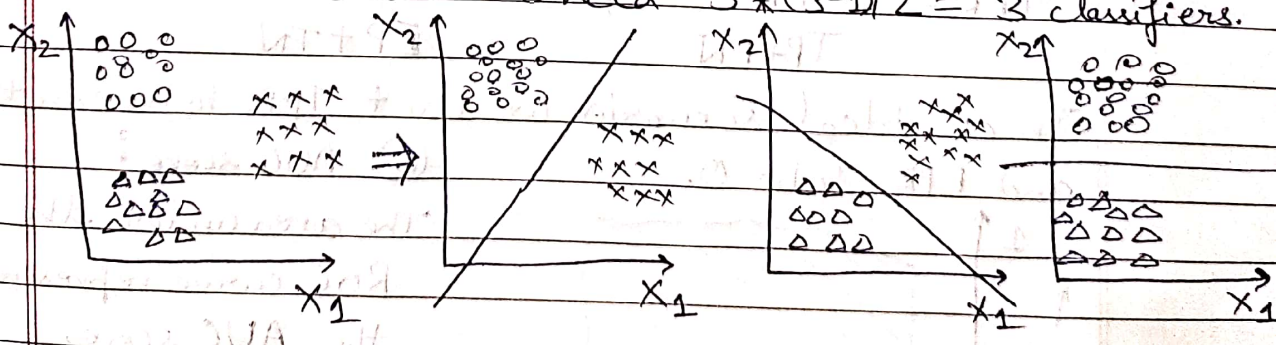
③

X_1	X_2	X_3	Y
a	b	c	-1
d	e	f	-1
g	h	i	+1

Here, green treated as positive class.

→ Red classifier yields positive with probability of 0.9
→ Blue classifier yields positive with probability of 0.4.
→ Green classifier yields negative with probability of 0.5.
Then, we take a majority voting that the final response belongs to positive class and since probability for red classifier is highest, it means the colour is red.

(b) One vs One (OvO) classification method —
In this method, if we have 'n' classes in our dataset, then we need $n*(n-1)/2$ binary classifiers, suppose we have three classes then we need $3*(3-1)/2 = 3$ classifiers.



Each binary classifier predicts one class label and the label with most votes is decided winner and is assigned with that label.

* Evaluation metrics for classification problem—

→ Confusion Matrix:

A 2×2 confusion matrix for binary classification. For multi-class classification it can go upto n×n.

Actual →	1	0
Predicted → 1	TP	FP → Type-I error
Predicted → 0	FN → Type-II error	TN

① Accuracy = $\frac{TP+TN}{TP+TN+FP+FN}$

② Precision = $\frac{TP}{TP+FP}$, Suppose, we have a spam classifier then we should reduce no. of false positives because if we predict non-spam mail as spam, then a person can miss important email.

③ Recall = $\frac{TP}{TP+FN}$, It is also known as **sensitivity**. In cancer detection, we should focus to reduce FN.

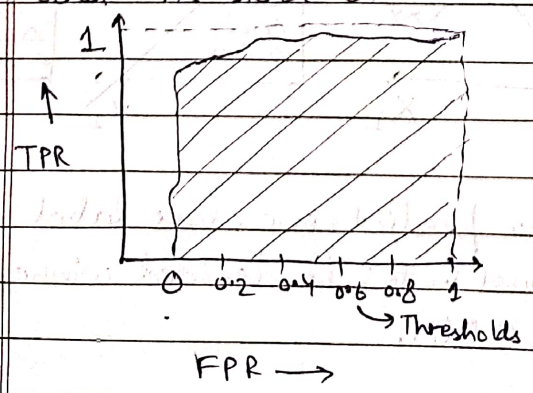
④ F-1 score: It is harmonic mean of precision and recall and its formula is given as—

$$F1\text{-score} = \frac{2 \times P \times R}{P+R}, \quad P = \text{Precision}, \quad R = \text{Recall}.$$

⑤ ROC-curve: ROC curve is an abbreviation for Receiver Operator Characteristic curve. It is a chart plotted between TPR and FPR at different threshold.

$$TPR = \frac{TP}{TP+FN}, \quad FPR = \frac{FP}{FP+TN}$$

In an ideal scenario, we want TPR to be exactly 1 and FPR to be 0.



⑥ AUC score:

The area under the ROC curve represents the AUC score.

** The AUC score is solely based on model and this metric does not change for varying thresholds.

Interview questions on Logistic Regression

Q.1 Why is logistic regression called regression if its predicting classifying labels and not continuous value?

Ans→ Refer page 2-3.

Q.2 What is the loss function used in logistic regression?

Ans→ We use cross-entropy / log-loss as a loss function in LR. Refer page 3.

Q.3 Why can't we use OLS/MSE as a loss function for logistic regression.

Ans→ The $\hat{y}$ (prediction/response) coming from the logistic regression model is from a non-linear sigmoid function and because of it the OLS/MSE function becomes non-convex and we can't find global minima for it.

Q.4 What does slope in logistic regression signify?

Ans→ In logistic regression, we model our data in form of -

$$\log\left(\frac{p}{1-p}\right) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$$

$$\Rightarrow \frac{p}{1-p} = e^{(\theta_0 + \theta_1 x_1 + \theta_2 x_2)}$$

So, we can say with every unit increase in x_1, the odds in favour of event changes by $e^{\theta_1 *}$.

Q.5 Can we use logistic regression for multi-class classification?

Ans→ Yes. Refer page 4-5.

Q.6 What are different evaluation metrics for classification?

Ans→ Refer page 6.