Unit 2.2
Logistic Regression

Logistic Regression

Disclaimer

The content is curated from online/offline resources and used for educational purpose only

Logistic Regression : Demo Case

I'm really rich!
I should be worthy!
CHOSEN ONE
Special Rock
SCALE

Learning Objectives

- Introduction to Logistic Regression
- Sigmoid Activation
- Training Process
- Parameter Estimation
- Evaluation Metrices
- Concept of Cross-Validation

Categorical Response Variables

Examples

Whether or not a person smokes

Binary Response

$$Y = \begin{cases} \text{Non} - \text{smoker} \\ \text{Smoker} \end{cases}$$

Success of a medical treatment

$$Y = \begin{cases} \text{Survives} \\ \text{Dies} \end{cases}$$

Opinion poll responses

Ordinal Response

$$Y = \begin{cases} \text{Agree} \\ \text{Neutral} \\ \text{Disagree} \end{cases}$$

Logistic Regression

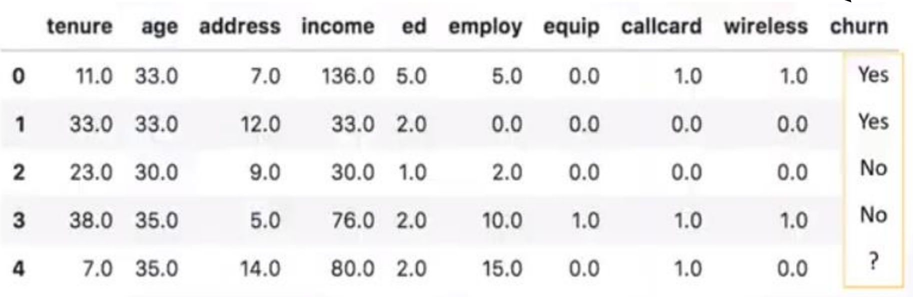
Introduction

Logistic regression is a statistical and machine learning technique for classifying records of a dataset, based on the values of the input fields.

Let’s say we have a telecommunication dataset which you’ll use to build a model based on logistic regression for predicting customer churn, using **the given features.**

INDEPENDENT VARIABLES

Dependent Variable



	tenure	age	address	income	ed	employ	equip	callcard	wireless	churn
0	11.0	33.0	7.0	136.0	5.0	5.0	0.0	1.0	1.0	Yes
1	33.0	33.0	12.0	33.0	2.0	0.0	0.0	0.0	0.0	Yes
2	23.0	30.0	9.0	30.0	1.0	2.0	0.0	0.0	0.0	No
3	38.0	35.0	5.0	76.0	2.0	10.0	1.0	1.0	1.0	No
4	7.0	35.0	14.0	80.0	2.0	15.0	0.0	1.0	0.0	?

Logistic Regression Applications

- **Predict the probability of a person having a heart attack** within a specified time period, based on our knowledge of the person's age, sex, and body mass index.
- **Predict the chance of mortality** in an injured patient
- **Predict whether a patient has a given diseas**e, such as diabetes, based on observed characteristics of that patient
- **Predict the likelihood of a customer purchasing a product** or halting a subscription.
- **Predict the probability of failure of a given process, system, or product**.

Note: All of these applications, we not only predict the class of each case, we also measure the probability of a case belonging to a specific class.

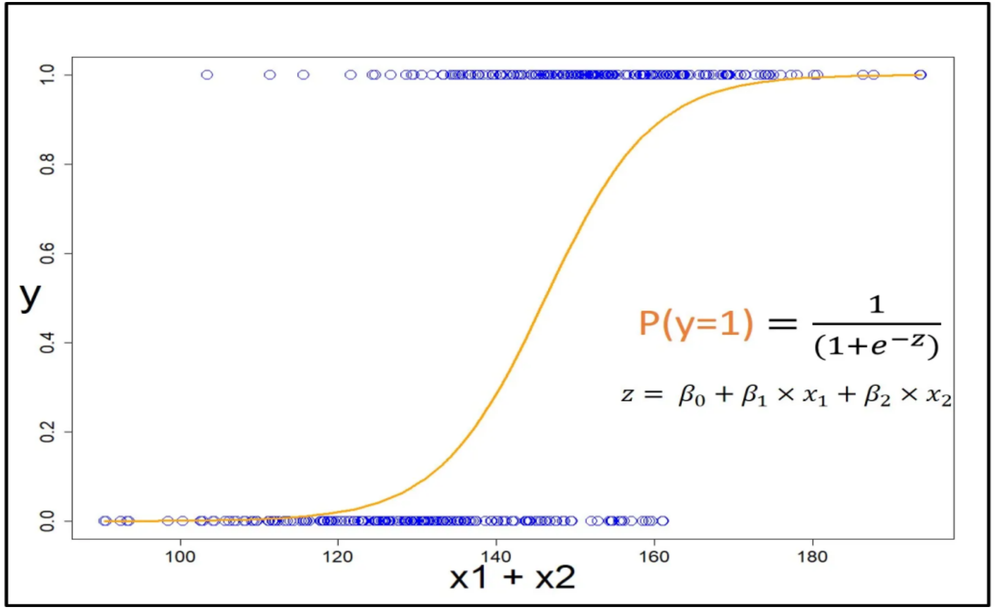
Logistic Regression Model

Ideally, a logistic regression model, so called ŷ, can **predict that the class of a customer is 1, given its features x (probability of customer falling in a particular class).**

$$\hat{y} = P(y = 1|x)$$

For class of customer =0

$$\hat{y} = P(y = 0|x) = 1 - P(y = 1|x)$$



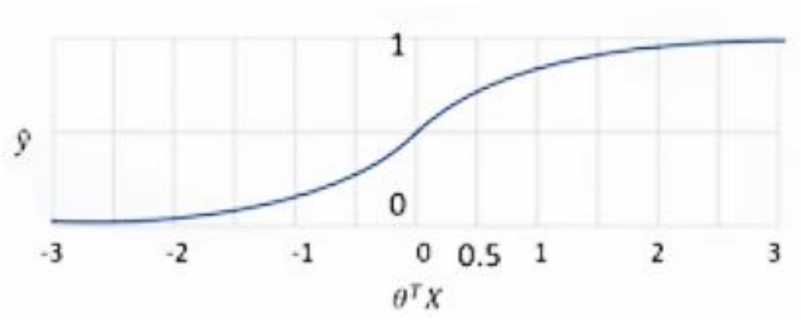
Reference

Logistic Regression

Logistic Regression Model

- To build such model, Instead of using $\theta^T X$ we use a specific function called sigmoid.

- $\sigma(\theta^T X)$ gives us the probability of a point belonging to a class, instead of the value of y directly.



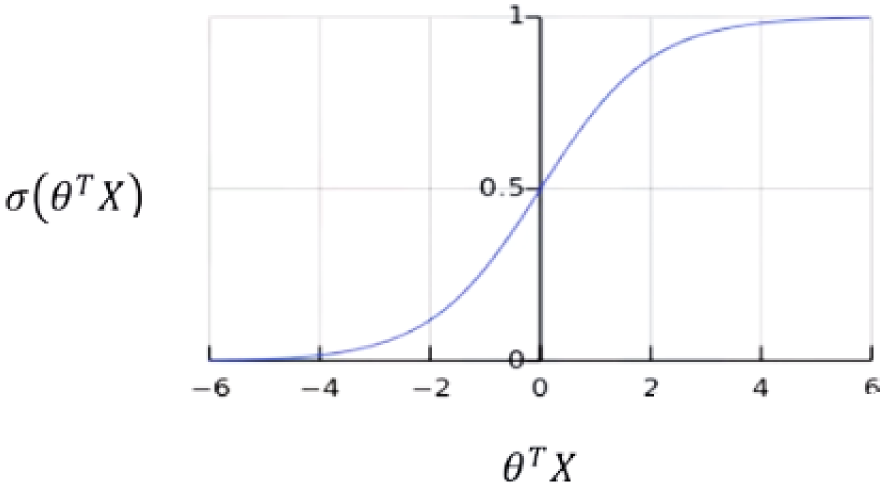
$$\hat{y} = \sigma(\theta^T X) = \sigma(\theta_0 + \theta_1 x_1 + \cdots)$$

Logistic Regression

Logistic Function

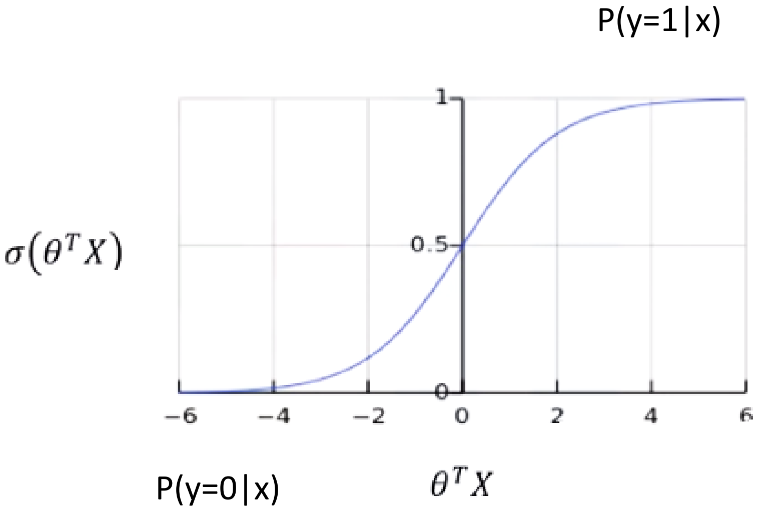
- **Sigmoid function, (the logistic function),** resembles the step function and is used by the following expression in the logistic regression.

$$\sigma(\theta^T X) = \frac{1}{1+e^{-\theta^T X}}$$



Sigmoid Function

- When $\theta^T X$ goes bigger, sigmoid function gets closer to 1, the P(y=1|x), goes up.
- When $\theta^T X$ goes very small, sigmoid function gets closer to 0, thus, the P(y=1|x), goes down.
- Sigmoid function’s output is always between 0 and 1, which makes it proper to interpret the **results as probabilities**.



Output of Model with Sigmoid Function

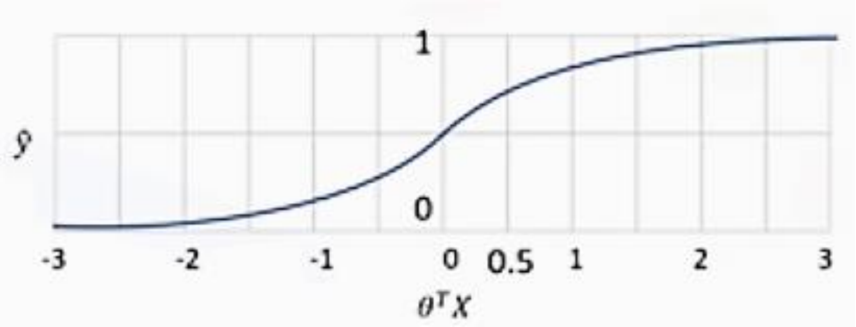
Ex:

- The probability of a customer staying with the company can be shown as probability of churn equals 1 given a customer’s income and age, assume, 0.7.

P(churn=1|income,age) = 0.7

- And the probability of churn is 0, for the same customer:

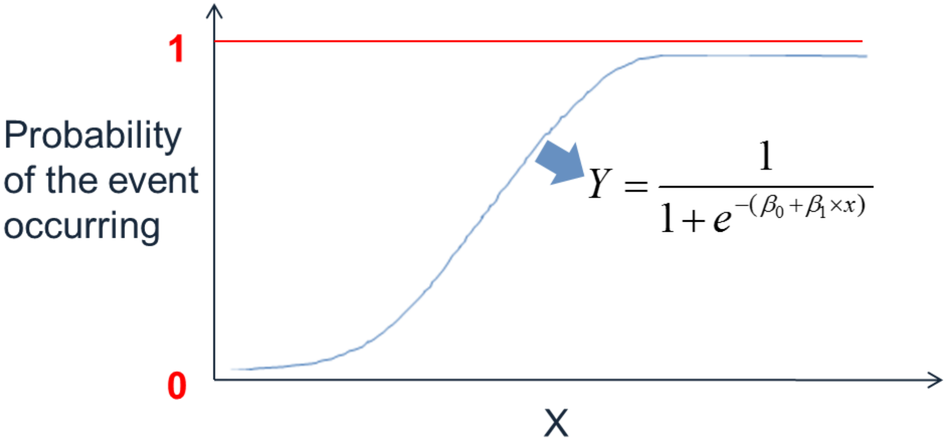
P(churn=0|income,age) = 1-0.7=0.3



Model Build?

How can we build such model ?

- First step towards building such model is to find $\boldsymbol{\theta}$, which can find through the training process.
- We can plan to learn from the data itself in training phase.



$$Y = \frac{1}{1+e^{-(\beta_0+\beta_1\times x)}}$$

Reference

Training Process

Step 1: Initialize θ vector with random values, assume [−1, 2].

Step 2: Calculate the model output, which is σ($\theta^T X$), for a sample customer.

X in $\theta^T X$ is the feature vector values. Ex: the age and income of the customer, assume, [2, 5].

θ is the confidence or weight that you've set in the previous step.

The probability that the customer belongs to class 1 is:

$$\hat{y} = \sigma(\theta^T X) = \sigma([-1,2] * [2,5]) = 0.7$$

Step 3: Compare the output of our model, ŷ, with the actual label of the customer, assume, 1 for churn.

Ex: Model's error = 1-0.7 = 0.3

This is the **error for only one customer out of all** the customers in the training set.

Training Process

Step 4: Calculate the error for all customers.

Add up to find total error, which is the cost of your model,

Cost function (error of the model) is the difference between the actual and the model’s predicted values. Therefore, the lower the cost, the better the model is at estimating the customer’s labels correctly.

We must try to minimize this cost !

Step 5: But, because the initial values for θ were chosen randomly, it’s very likely that the cost function is very high. So, we change the θ in **such a way** to hopefully reduce the total cost.

Step 6: After changing the values of θ, we go back to step 2 to start another iteration and calculate the cost of the model again **until the cost is low enough**.

Logistic Regression

Loss Function

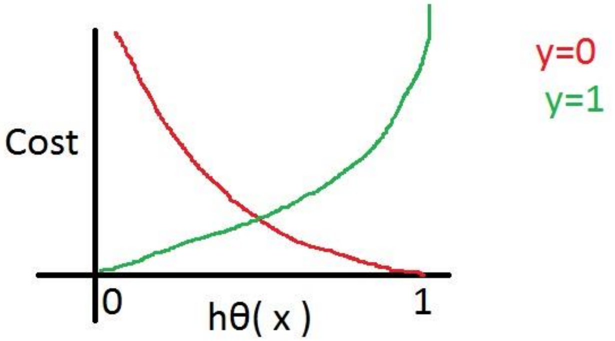
- Usually, the square of this equation is used because of the possibility of the negative result, and for the sake of simplicity, half of this value is considered as the cost function, through the derivative process.

$$\boldsymbol{Cost(\hat{y}, y) = \frac{1}{2}\left(\sigma\left(\theta^T X\right) - y\right)^2}$$

- Now, we can write the cost function for all the samples in our training set; for example, **for all customers, we can write it as the average sum of the cost functions of all cases.**

$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} Cost(\hat{y}, y)$$

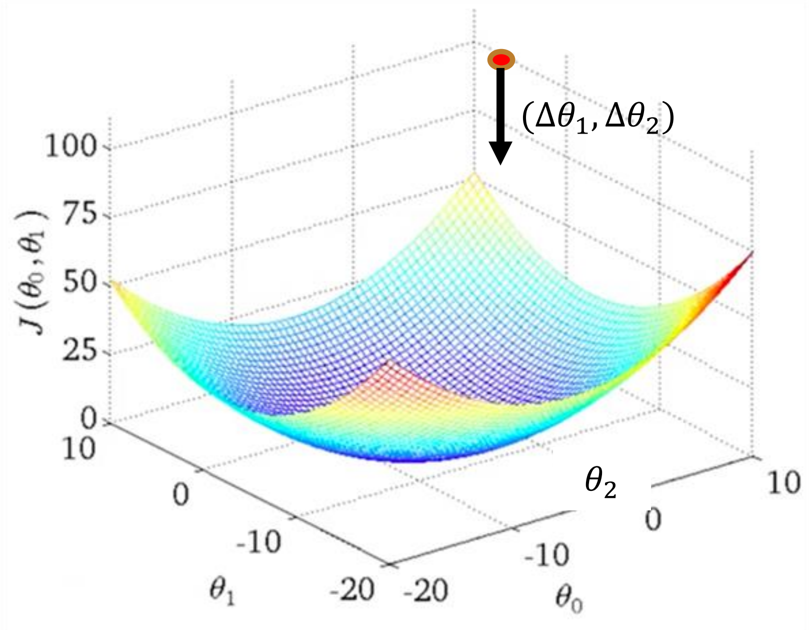
$$J(\theta) = \frac{1}{m}\sum_{i=1}^{m} y^i \log\left(\hat{y}^i\right) + \left(1 - y^i\right)\log(1 - \hat{y}^i)$$



Reference

Gradient Descent

- If we plot the cost function based on all possible values of $\theta 1$, $\theta 2$,
- **We call it “error curve” or “error bowl” of cost function.**
- It represents the error value for different values of parameters



Reference

Evaluation Metrics in Classification

- **Evaluation metrics explain the performance of a model.**
- Imagine that we have an historical dataset which shows the customer churn for a telecommunication company.
- We have trained the model, and now we want to calculate its accuracy using the test set.
- We pass the test set to our model, and we find the predicted labels.
- Now the question is, “**How accurate is this model?”**
- Basically, we compare the actual values in the test set with the values predicted by the model, to calculate the accuracy of the model.

Evaluation Metrics in Classification

- Evaluation metrics provide a key role in the development of a model, as they **provide insight to areas that might require improvement.**

There are different model evaluation metrics but we will talk about three popular metrics:

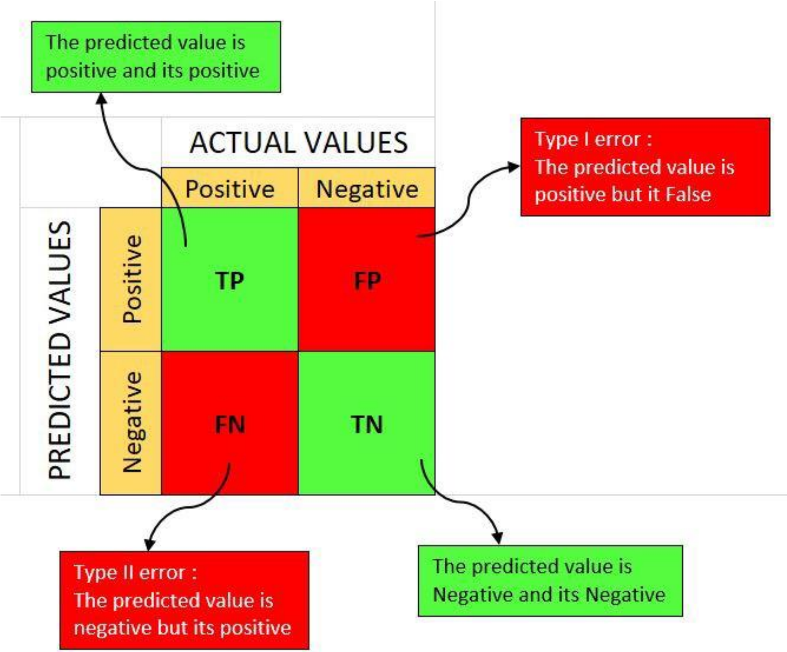
1. Jaccard index
2. F1-score
3. Log Loss

Jaccard Index

- Also known as the Jaccard Similarity Coefficient
- Let’s say ***y* = true labels** of the dataset.
- And **y_cap** be predicted values estimated by our classifier.
- Then we can define Jaccard Index as, the **size of intersection** divided by the size of union of true and predicted dataset

$$J(y,\hat{y}) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|} = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|}$$

Confusion Matrix



The predicted value is positive and its positive

		ACTUAL VALUES	
		Positive	Negative
PREDICTED VALUES	Positive	TP	FP
	Negative	FN	TN

Type I error :
The predicted value is positive but it False

Type II error :
The predicted value is negative but its positive

The predicted value is Negative and its Negative

Reference

Realization….

- The entries of the confusion matrix are the number of occurrences of each class for the dataset being analyzed.
- Let’s obtain the confusion matrix for our spam filtering algorithm, by using the function confusion_matrix:

```
from sklearn.metrics import confusion_matrix
print(confusion_matrix(y_test, y_pred))
```

- The output for spam detection system is:

```
[[724   7]
 [  1  136]]
```

Inference from Conf. Matrix

Let’s interpret results,

- Out of 731 actual instances of ‘not spam’ (first row), the classifier correctly predicted 724 of them.
- Out of 137 actual instances of ‘spam’ (second row), the classifier correctly predicted 136 of them.
- Out of all 868 emails, classifier correctly predicted 860 of them.
- This allow us to obtain the accuracy of classification from confusion matrix using below formulae

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

Logistic Regression

Precision, recall and f1-score

- **Precision:** When **positive result** is predicted, **how often is it correct**.
- Used when we need to limit the number of False Positives

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall:** When it is actually **positive result**, how often does it **predict correctly**.
- Used when we need to limit the number of False Negatives
- Recall also known as **“Sensitivity” or the True Positivite Rate (TPR)**

$$\text{Recall} = \frac{TP}{TP + FN}$$

Precision, recall and f1-score

- **f1-score:** this is just the harmonic mean of precision and recall

$$\text{f1-score} = 2 \times \frac{precision \times recall}{precision + recall}$$

- Useful when you need to take precision and recall into account
- If you try to optimize recall you algorithm will predict more examples to belong to positive class, but that may result in many false positive hence low precision
- On other hand, if you try to optimize precision, your model will predict very few examples as positive results (one with highest probability), but the recall will be very low
- **f1-score reaches is best value as 1 (represent perfect precision and recall) and the worst value as 0.**

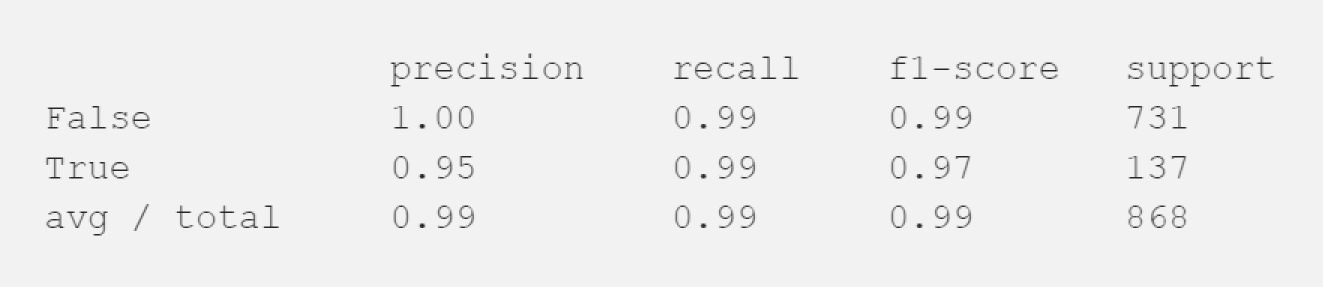
Logistic Regression

Simulation Output

from sklearn.metrics import classification_report

print(classification_report(y_test, y_pred))

The output is:



```
             precision    recall    f1-score    support
False        1.00         0.99      0.99        731
True         0.95         0.99      0.97        137
avg / total  0.99         0.99      0.99        868
```

Logarithmic (Log) Loss

- Now let's look at another accuracy metric for classifiers.
- **Sometimes, the output of a classifier is the “probability of a class label”, instead of the “label”.**
- For example, in logistic regression, the output can be the probability of customer churn, i.e., yes (or equals to 1).
- **This probability is a value between 0 and 1.**
- Logarithmic loss (also known as Log loss) measures the performance of a classifier where the predicted output is a probability value between 0 and 1

Log Loss

In order to calculate Log Loss, classifier needs to assign a probability to each class rather than yielding more likely class.

Log Loss can be defined as,

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij} \times \log(p_{ij})$$

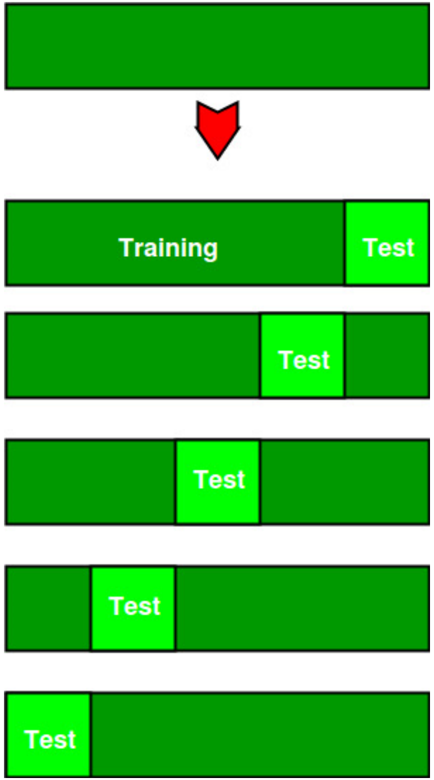
where,

N is the number of samples or instances.

M is the number of possible labels

Cross-Validation

- Machines may start with different random values.
- Performance may differ
- **Unrepeatable** results
- **Non-generalised** solution
- K-Fold cross-validation



Reference

Logistic Regression

Lab 1: Logistic Regression on Two Class dataset

Summary

In this session we have learned:

- What does it mean by term logistic regression
- How regression is a case of probability handling
- What should be needful properties of sigmoid function
- How to interpret logistic regression outcomes
- How to implement logistic regression for any given case
- Why model generalization is important?
- In this regression technique, why MSE is not used?

Quiz

Q1. What is the primary purpose of logistic regression?

a. Predicting continuous outcomes

b. Classifying data into two or more categories

c. Estimating the mean of a dependent variable

d. Analyzing the correlation between two variables

Answer: b. Classifying data into two or more categories

Logistic Regression

Quiz

Q2. In logistic regression, what is the output variable (dependent variable)?

a. Continuous variable

b. Categorical variable

c. Independent variable

d. None of the above

Answer: b. Categorical variable

Logistic Regression

Quiz

Q3. Which function is commonly used as the activation function in logistic regression?

a. ReLU (Rectified Linear Unit)

b. Sigmoid function

c. Tanh (Hyperbolic Tangent) function

d. Linear function

Answer: b. Sigmoid function

Logistic Regression

Quiz

Q4. When performing logistic regression, what evaluation metric is commonly used to assess the model's performance?

a. Mean squared error (MSE)

b. R-squared (R^2)

c. Accuracy, precision, recall, and F1-score

d. Correlation coefficient (r)

Answer: c. Accuracy, precision, recall, and F1-score

References

- https://en.wikipedia.org/wiki/Logistic_regression
- https://web.stanford.edu/~hastie/ElemStatLearn//printings/ESLII_print12.pdf
- https://www.coursera.org/learn/machine-learning
- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html
- https://towardsdatascience.com/a-practical-introduction-to-logistic-regression-eb13d5b075ac
- https://www.r-bloggers.com/2020/09/logistic-regression-in-r/
- "Introduction to Logistic Regression Analysis" by William J. Rudasill, Richard G. Lomax, and Hung-Chung Huang
- "Logistic Regression: A Self-Learning Text" by David G. Kleinbaum and Mitchel Klein

Logistic Regression

Thank You