Classification

Logistic regression

Tech Lead Data Science

Master en Data Science
2022-2023

Assembler
Institute of Technology

ÍNDICE

CLASSIFICATION

CLASSIFICATION

- Classification problems have an **independent categorical variable** Y.
- They are processes that consist on identifying to which category or class belongs a determined object, according to their dependent variables.
- Examples:
  - Fraud detection
  - Definition of a target in a marketing campaign
  - Medical diagnosis
  - Image classification

Assembler
Institute of Technology

STEPS

1. **Training:**

   We build a classifier (model) learning from a labeled train set.

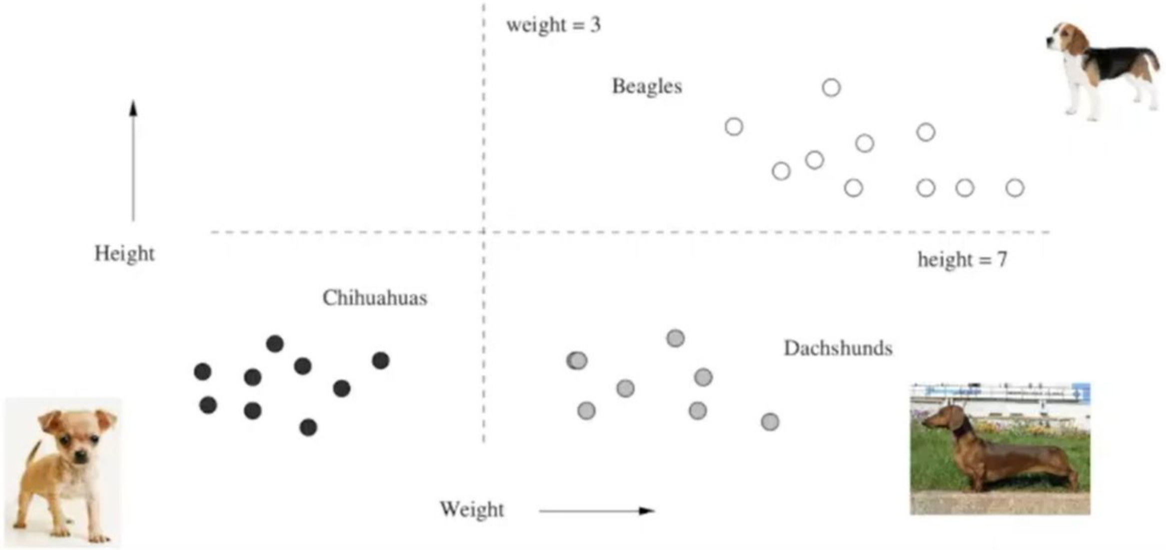
2. **Classification:**

   We use the model to classify.

3. **Evaluation:**

   We evaluate the model. This step can be included in the previous one.

CLASSIFICATION




```
REGLA
     if height > 7:
          print('Beagle')
     elif weight < 3:
          print('Chihuahua')
     else:
           print('Dachshund')
```

CLASSIFICATION

Life is not so easy: we cannot always develop clear rules☹️

To solve this kind of problem, we have machine learning models that **predict the probability of a given observation to belong to a particular class**:

- Bayesian models
- Logistic regression
- Decision Trees - Random forests
- Neural networks
- And more!!

CLASSIFICATION

Life is not so easy: we cannot always develop clear rules☹️

To solve this kind of problem, we have machine learning models that **predict the probability of a given observation to belong to a particular class**:

- Bayesian models
- **Logistic regression** ⭐
- Decision Trees - Random forests
- Neural networks
- And more!!

LOGISTIC REGRESSION

LOGISTIC REGRESSION

- Logistic Regression is a classification model, which is very easy to realize and achieves very good performance with linearly separable classes.
- Assumption about y:

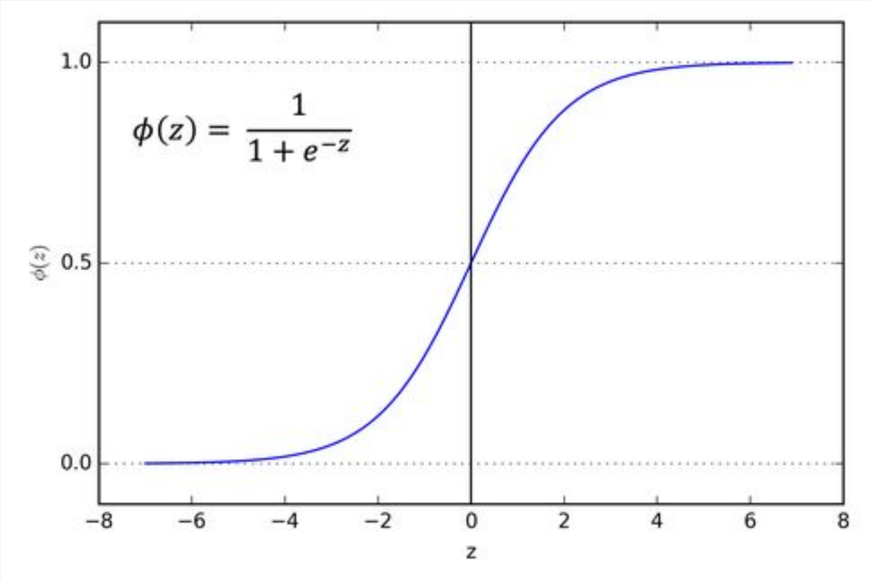
$$y \sim binomial(1, p) \longrightarrow y \begin{cases} 1 & p \\ 0 & 1 - p \end{cases}$$

- The predictor variables must be linearly independent.
- It is necessary to standardize the variables.
- It is a very sensitive model to atypical values or outliers .
- Logistic regression can be generalized to problems of more than two classes.

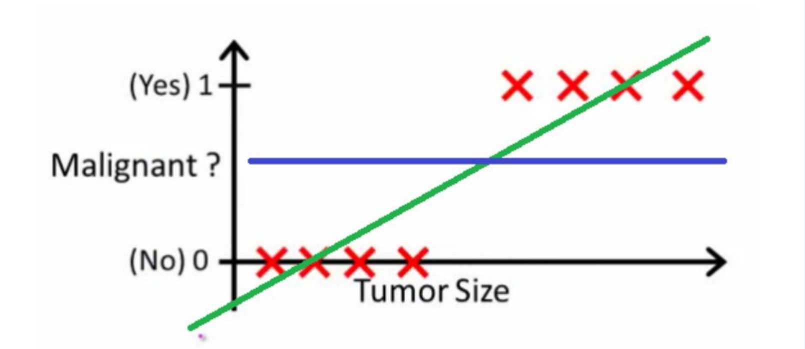
Assembler
Institute of Technology

LOGISTIC REGRESSION

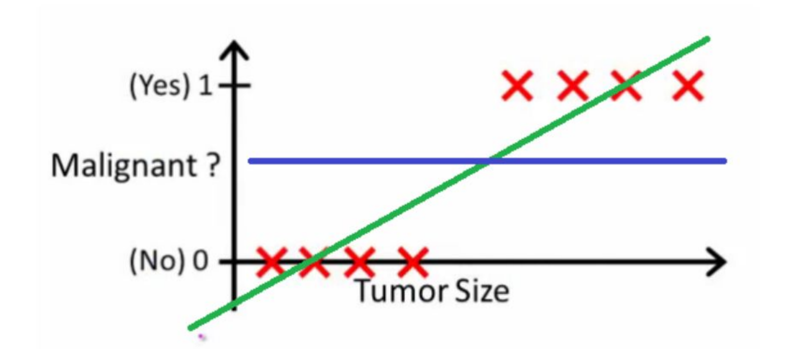
- It uses a logistic function

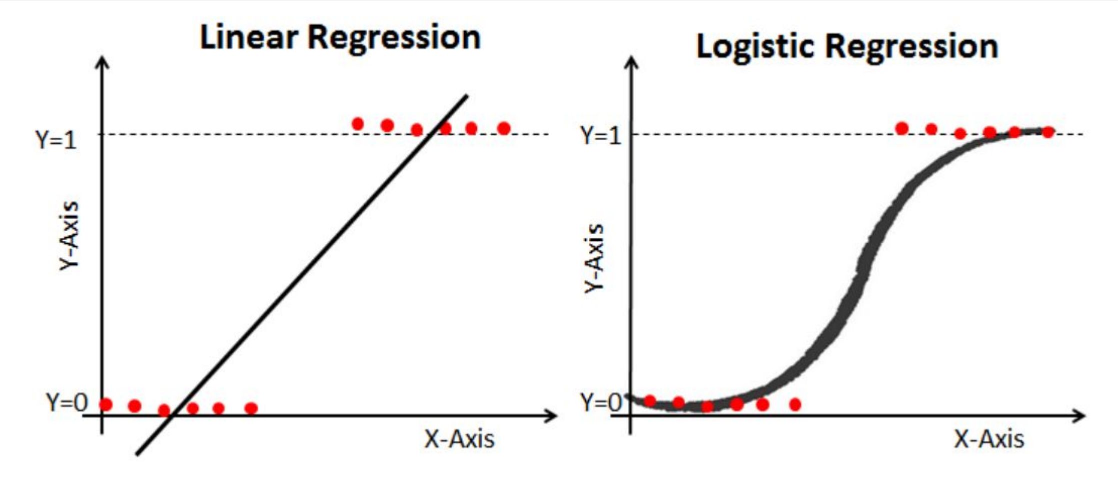


LOGISTIC REGRESSION



LOGISTIC REGRESSION





LOGISTIC REGRESSION

$$\hat{y}^{(i)} = \beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}$$

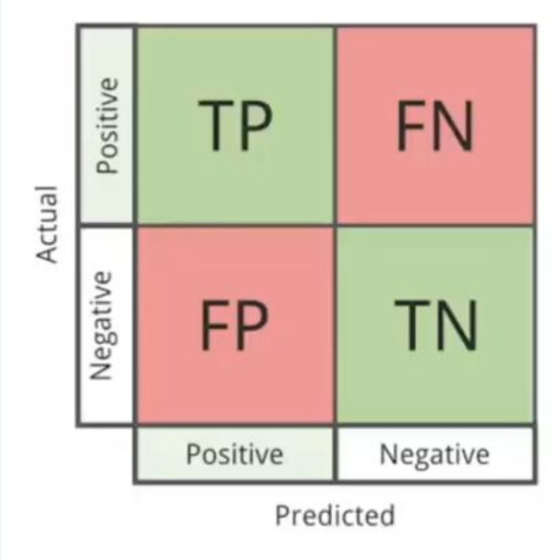
$$\phi(z) = \frac{1}{1 + e^{-z}} = \frac{1}{1 + exp(-(\beta_0 + \beta_1 x_1^{(i)} + \ldots + \beta_p x_p^{(i)}))}$$

EVALUATION METRICS

CONFUSION MATRIX

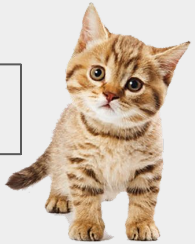
There are 4 possible values:

- True Positives (TP)
- True Negatives (TN)
- False Positives (FP)
- False Negatives (FN)



CONFUSION MATRIX

TYPE 2 ERROR

True positive CAT 	False negative NOT CAT 
False positive CAT 	True negative NOT CAT 

TYPE 1 ERROR

Assembler Institute of Technology

METRICS

- **Accuracy :** Percentage of cases in which our model was correct
- **Precision :** Percentage of values that have been classified as positive are actually positive
- **Recall :** Percentage of positive values that are identified
- **F1 Score :** Combines accuracy and comprehensiveness

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

ROC CURVE

- Represents the percentage of **true positives** (TPR or Recall) **against the false positives** ratio (FPR).
- Its values range from 0 to 1.