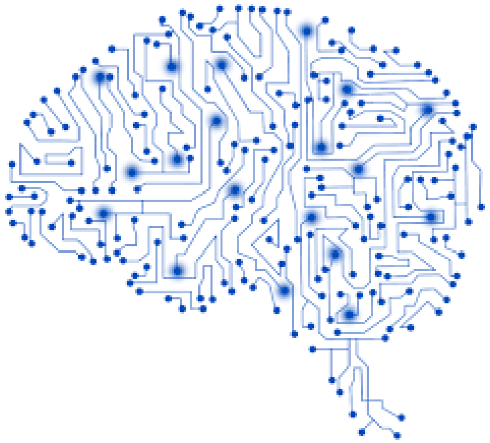
University of Minho
School of Engineering

ISLab



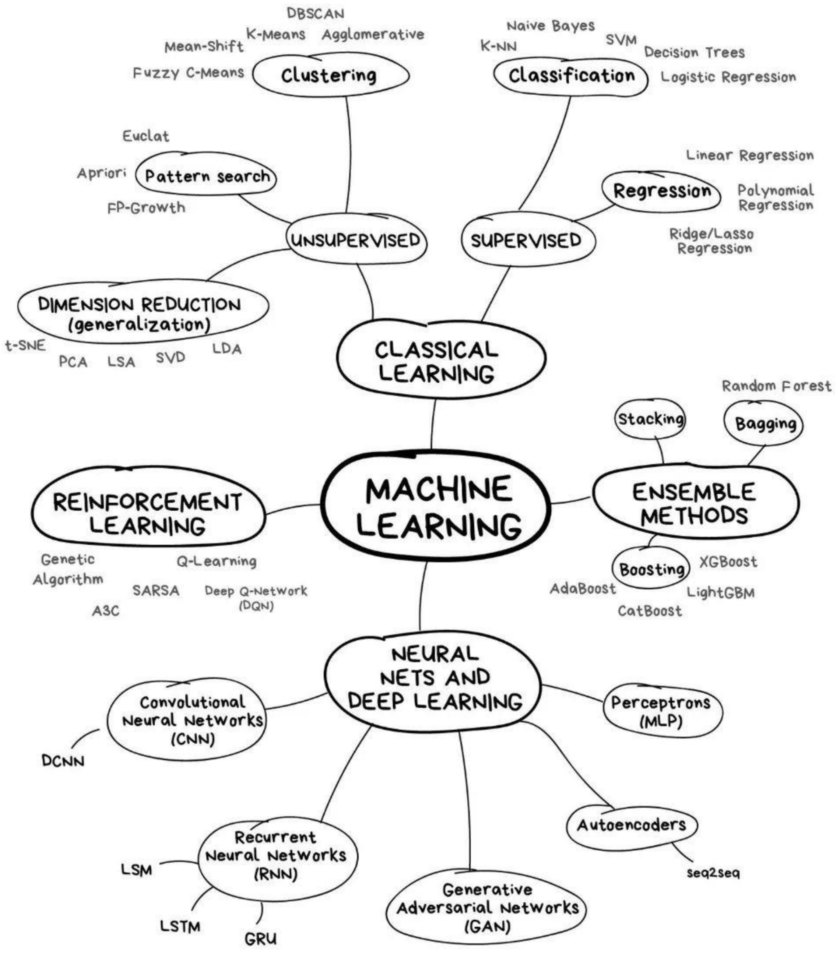
Dados e Aprendizagem Automática

Supervised Learning

Linear & Logistic Regression

DAA @ MEI/1º ano – 1º Semestre

Machine Learning



Source: The map of the machine learning world
Vasily Zubarev (vas3k.com)

Contents

- Supervised Learning:
  - The relationship between variables:
    - Covariance;
    - Correlation.
  - Linear Regression;
  - Logistic Regression.

The relationship between x and y
Covariance

Gives information on the degree to which two variables vary together.

$$\operatorname{cov}(x, y) = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{n-1}$$

- When X ↑ and Y ↑: cov (x,y) = pos.
- When X ↓ and Y ↑: cov (x,y) = neg.
- When no constant relationship: cov (x,y) = 0

The value obtained by covariance is dependent on the size of the data’s standard deviations: if large, the value will be greater than if small… even if the relationship between x and y is exactly the same in the large versus small standard deviation datasets.

The relationship between x and y
Correlation

- Correlation: is there a relationship between 2 variables?
- Correlation coefficients are indicators of the strength of the linear relationship between two different variables, x and y.
- The correlation coefficient (ρ) is a measure that determines the degree to which the movement of two different variables is associated.

Correlation coefficient

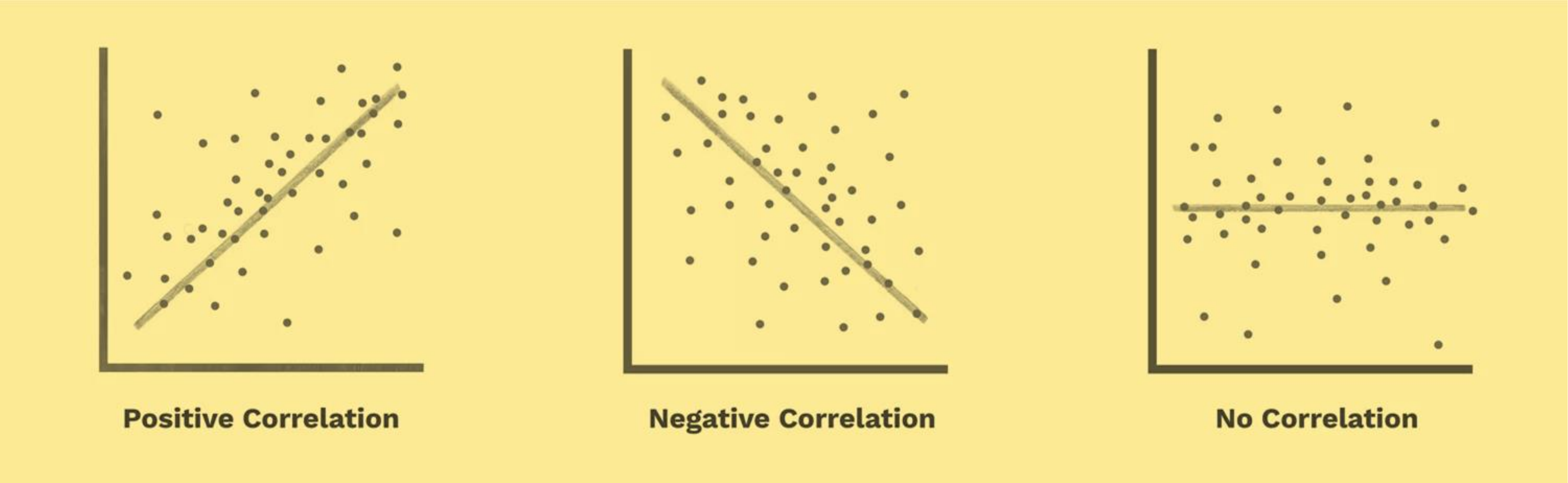
- **Pearson’s correlation coefficient** (linear correlation): shows linear correlation between two continuous variables

- **Spearman rank correlation coefficient:** non-parametric alternative to Pearson’s correlation coefficient

Correlation

- Measures the relative strength of the *linear* relationship between two variables
  - Unit-less
  - Ranges between –1 and 1
- The closer to –1, the stronger the negative linear relationship
- The closer to 1, the stronger the positive linear relationship
- The closer to 0, the weaker any positive linear relationship

Correlation



Source image: https://www.investopedia.com/ask/answers/032515/what-does-it-mean-if-correlation-coefficient-positive-negative-or-zero.asp

Regression

How well a certain independent variable predict dependent variable?

Regression is a statistical procedure that determines the equation for the straight line that best fits a specific set of data.

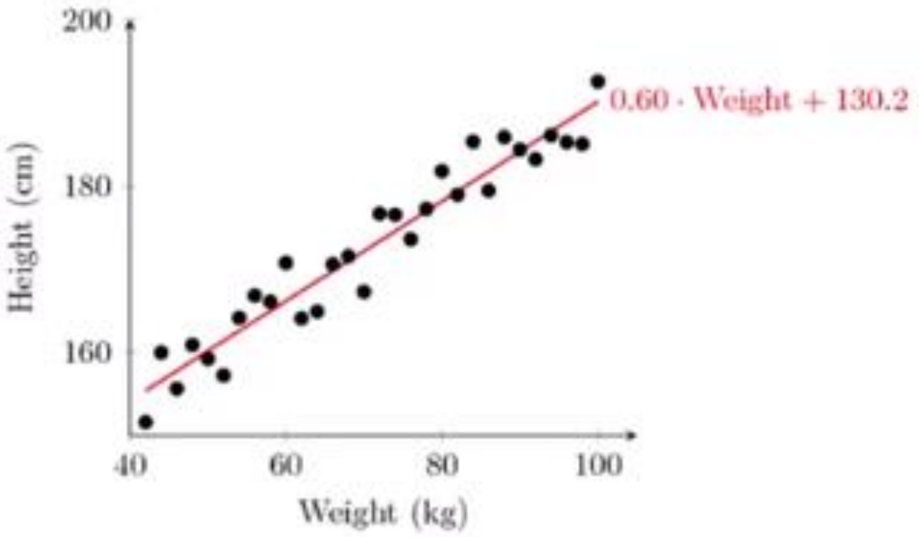
Linear Regression

Linear Regression

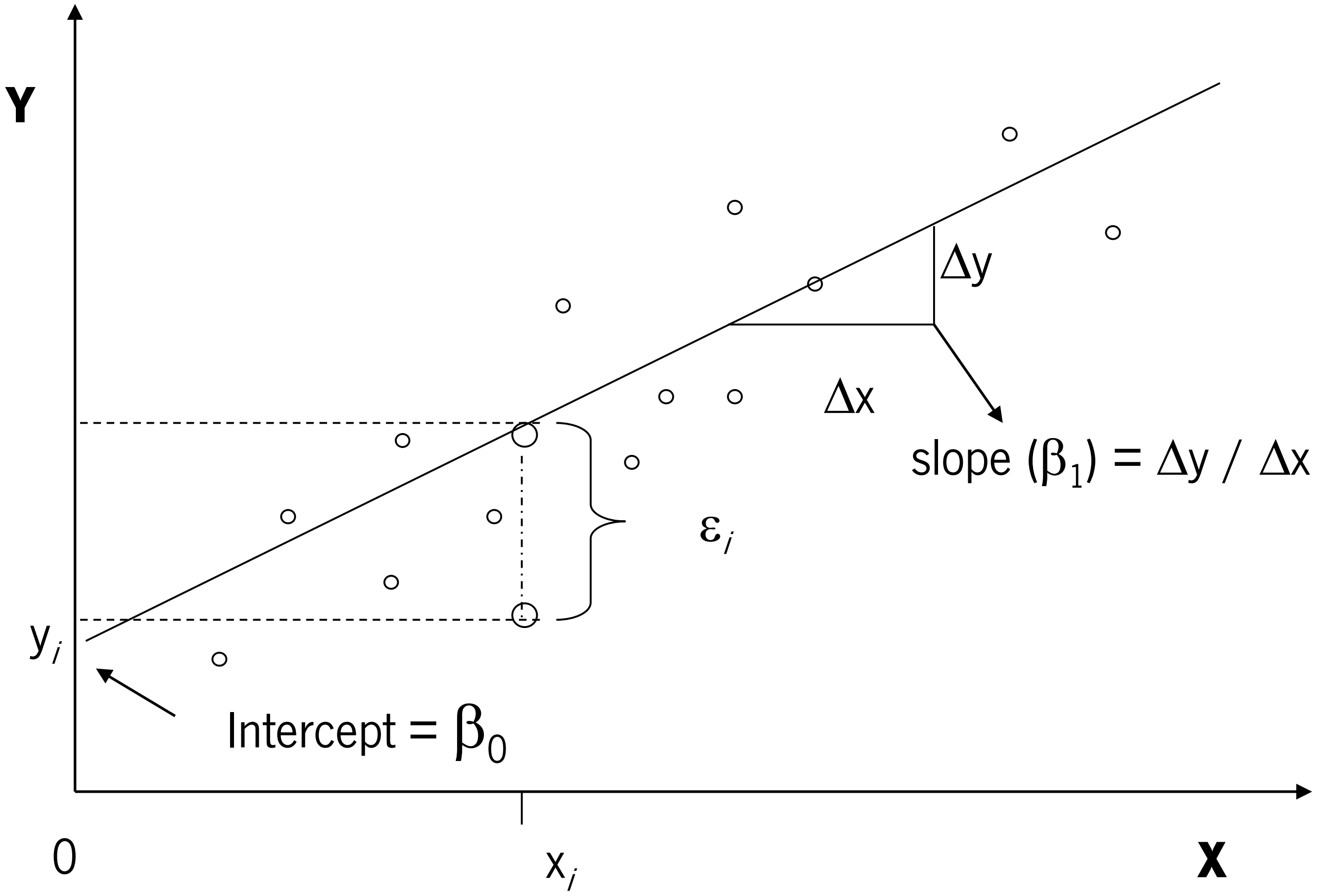
Linear Regression

Aims to predict the value of a outcome, Y, based on the value of an predictor variable, X.

- □ Fit a straight line into a data set of observations;
- □ Use this line to predict unobserved values.



Linear Regression

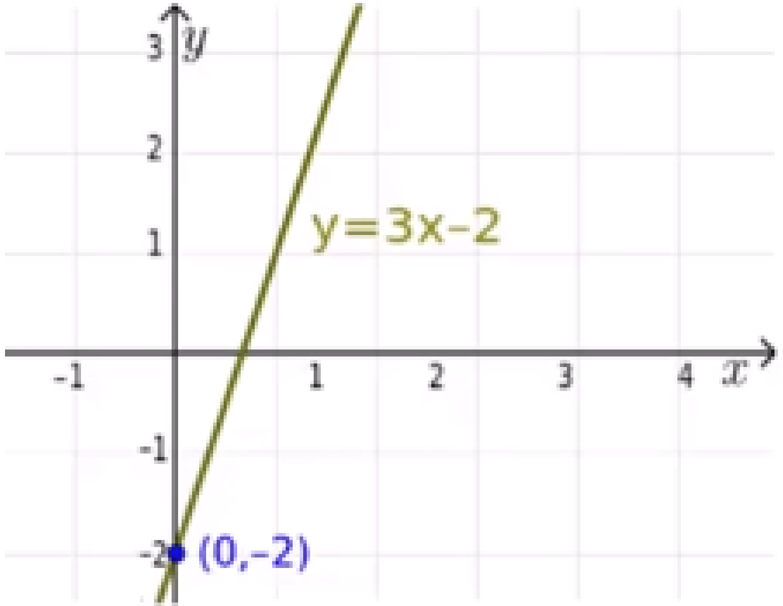


Model: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$

Linear Regression

How does it work?

- □ Usually using “least squares” - minimize the squared-error between each point and the line
- □ Follows the slope-intercept equation of a line: Y = m.x + b
  - ➢ m – slope: correlation between the two variables times the stand. Dev. In Y, all divided by the standard deviation in X.
  - ➢ b – y intercept: Intercept is the mean of Y minus the slope times the mean of X

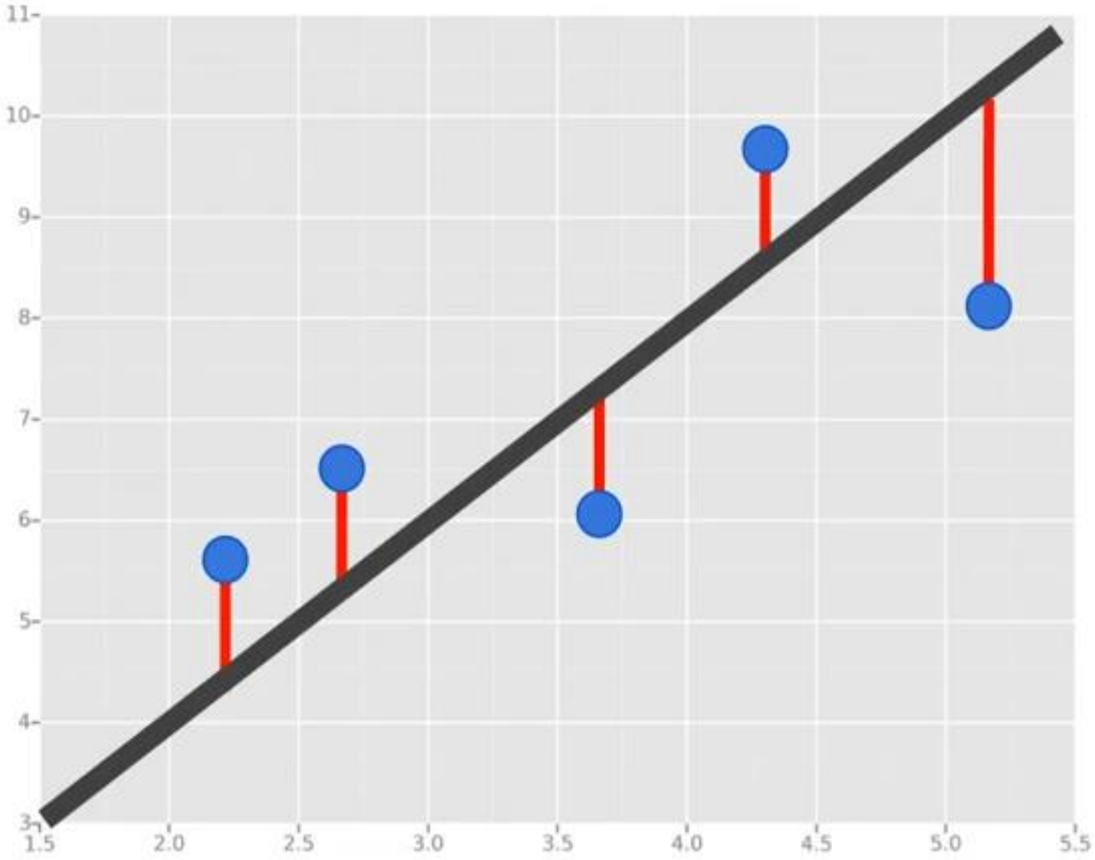


Linear Regression

How does it work?

- Least squares minimizes the sum of squared errors
  - y(i): true value
  - F(x(i), β): predicted value / fitted line

- Residuals for an observation is the difference between the observation (y-value) and the fitted line



$$r_i = y_i - f(x_i, \boldsymbol{\beta}).$$

The least-squares method finds the optimal parameter values by minimizing the sum, S, of squared residuals:

$$S = \sum_{i=1}^{n} {r_i}^2.$$

Linear Regression

Measuring error with r-squared

- How do we measure how well our line fits our data?
- R-squared (i.e., coefficient of determination) measures – “*the fraction of the total variation in Y that is captured by the model*”

Computing r-squared

$$1.0 - \frac{\textit{sum of squared errors}}{\textit{sum of squared variation from mean}}$$

- Ranges from [0-1]
- 0 is bad (none of the variance is captured)
- 1 is good (all of the variance is captured)

Multiple regression

- Multiple regression is used to determine the effect of a number of independent variables, x_1, x_2, x_3 etc, on a single dependent variable, y
- The different x variables are combined in a linear way and each has its own regression coefficient:

$$y = a_1x_1 + a_2x_2 + \ldots\ldots + a_nx_n + b + \varepsilon$$

- The a parameters reflect the independent contribution of each independent variable, x, to the value of the dependent variable, y.

Logistic Regression

Logistic Regression

The essential difference between these two is that Logistic regression is used when the dependent variable is binary in nature. In contrast, Linear regression is used when the dependent variable is continuous and nature of the regression line is linear.

Logistic Regression

- Logistic Regression as a method for Classification:
- Some examples of classification problems:
  - Spam vs “Ham” emails;
  - Loan Default (yes / no);
  - Disease Diagnosis.
- Above were all examples of Binary Classification.

Categorical Response Variables

Whether or not a person smokes

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

Binary Response

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

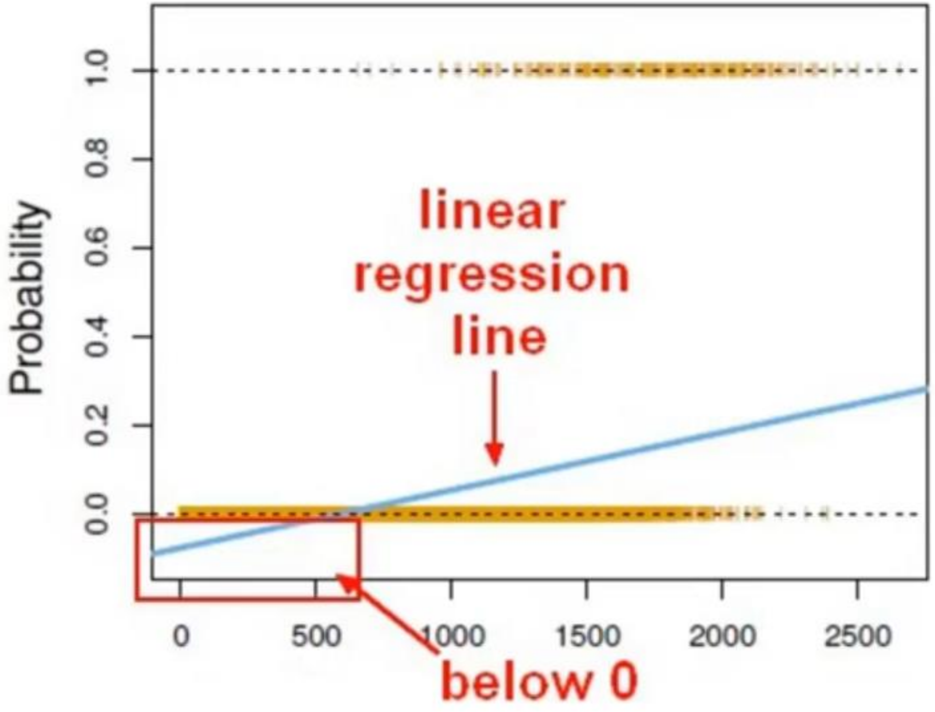
Logistic Regression

- Regression problems are normally used to predict a continuous value;
- Although the name may be confusing at first, logistic regression allows us to solve classification problems, where we are trying to predict discrete categories;
- The convention for binary classification is to have two classes: 0 and 1.

Logistic Regression

Logistic Regression

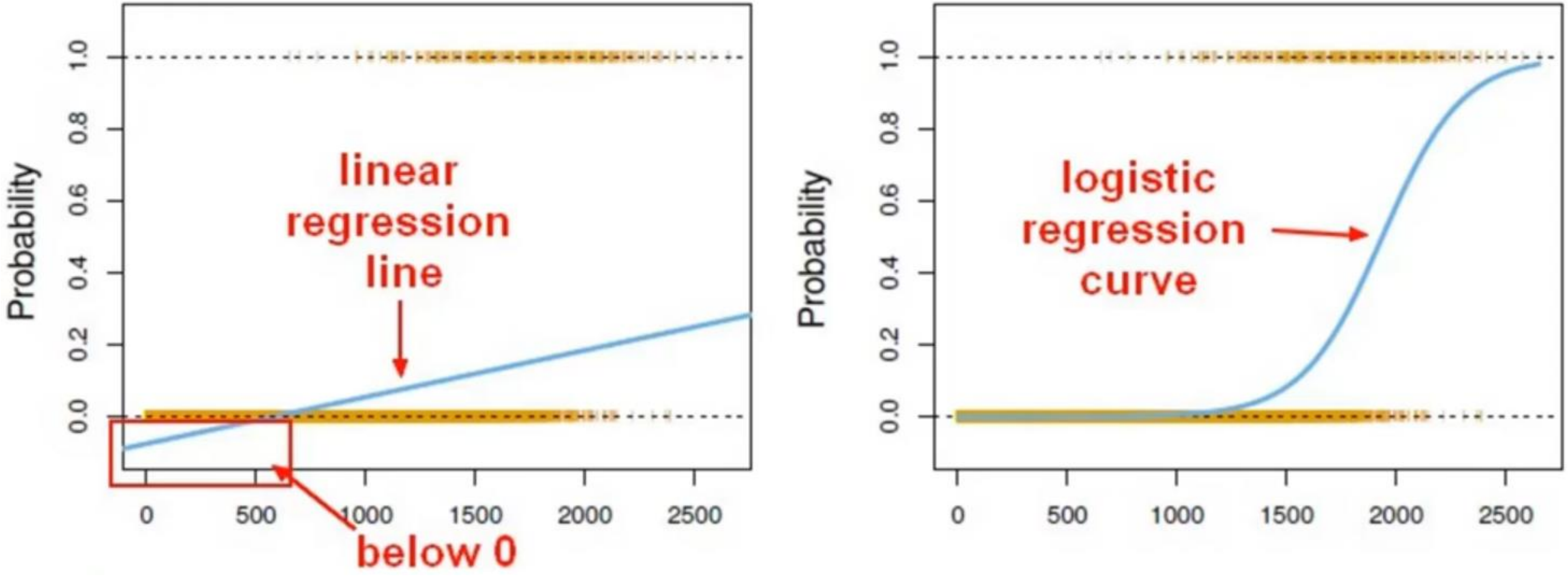
- We can’t use a normal linear regression model on binary groups. It won’t lead to a good fit.



Logistic Regression

Logistic Regression

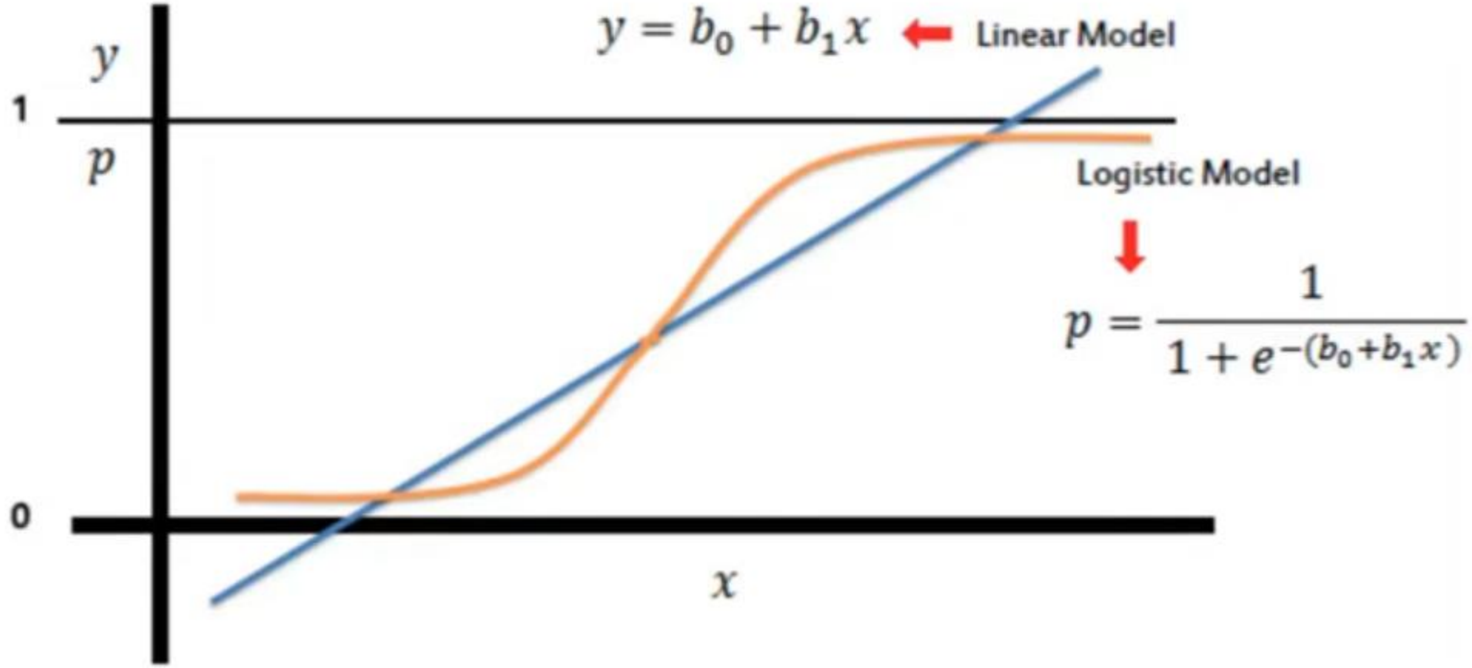
- Instead we can transform our linear regression into a logistic regression curve.



Logistic Regression

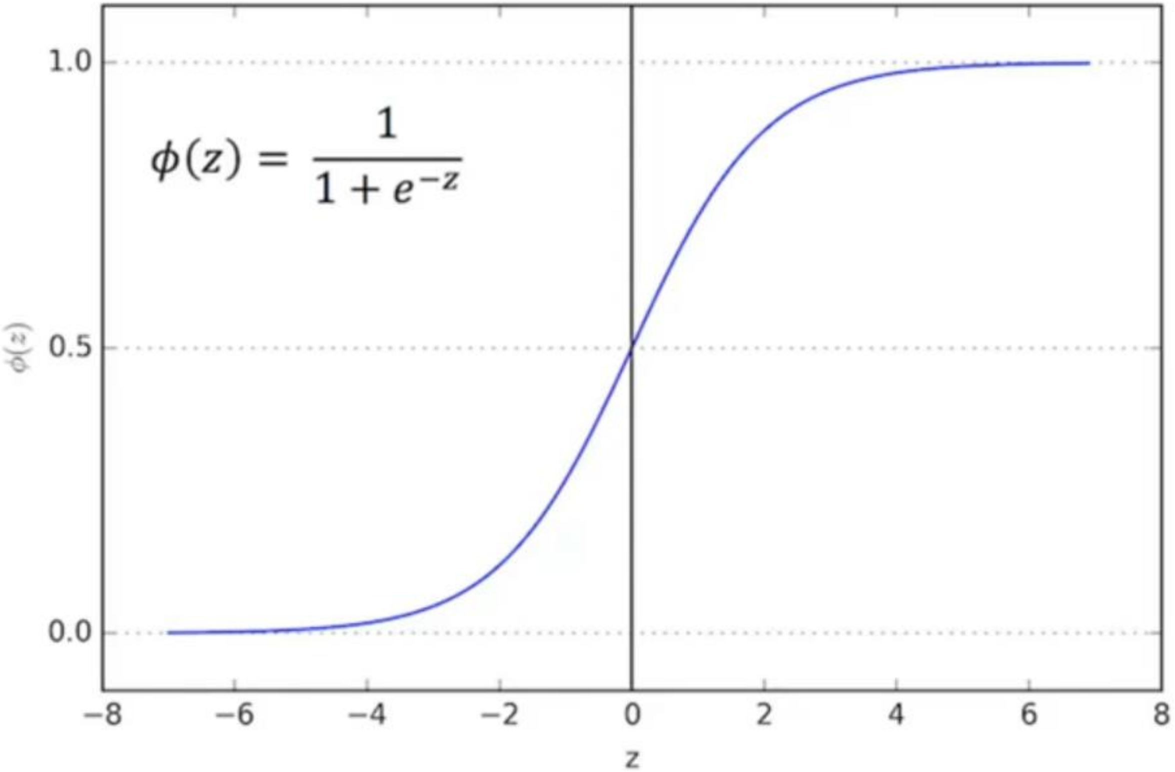
Logistic Regression

- We can take our Linear Regression Solution and place it into the Logistic Regression Function.



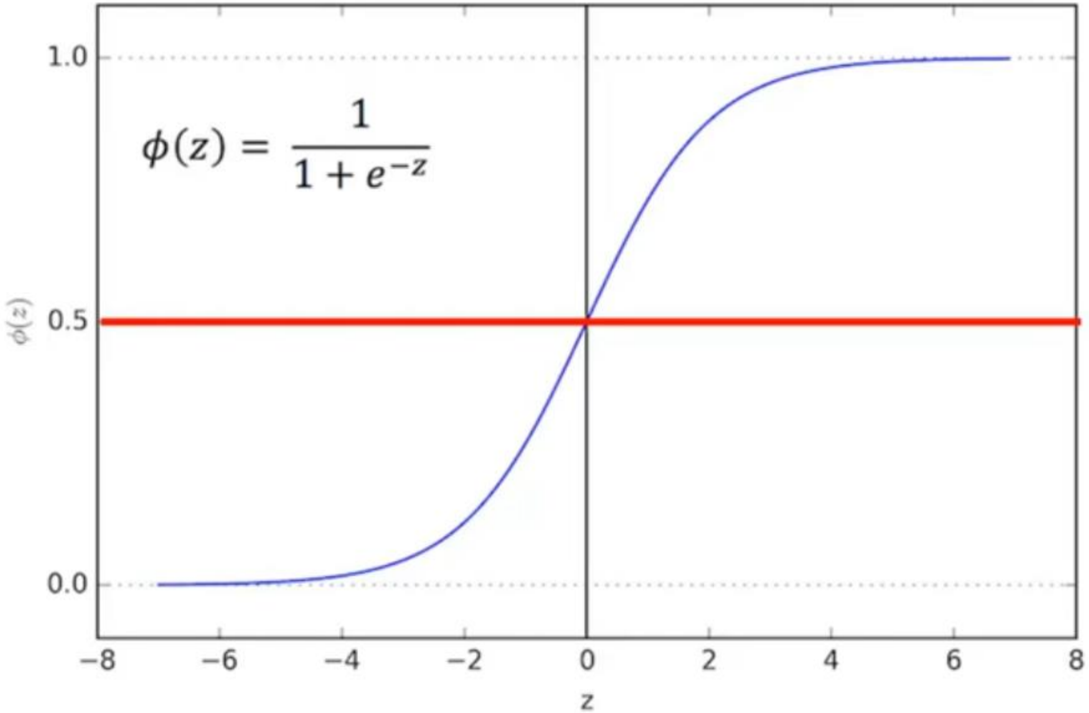
Logistic Regression

- □ The Sigmoid (i.e. Logistic) function takes in any value and outputs it between [0-1];
- □ This results in a probability from [0-1] of belonging into a class.



Logistic Regression

- We can set a threshold point at 0.5, defining:
  - Based off this probability, we assign a class
  - Predicted results below this threshold results into a class: 0
  - Predicted results above result results into a class: 1



Logistic Regression

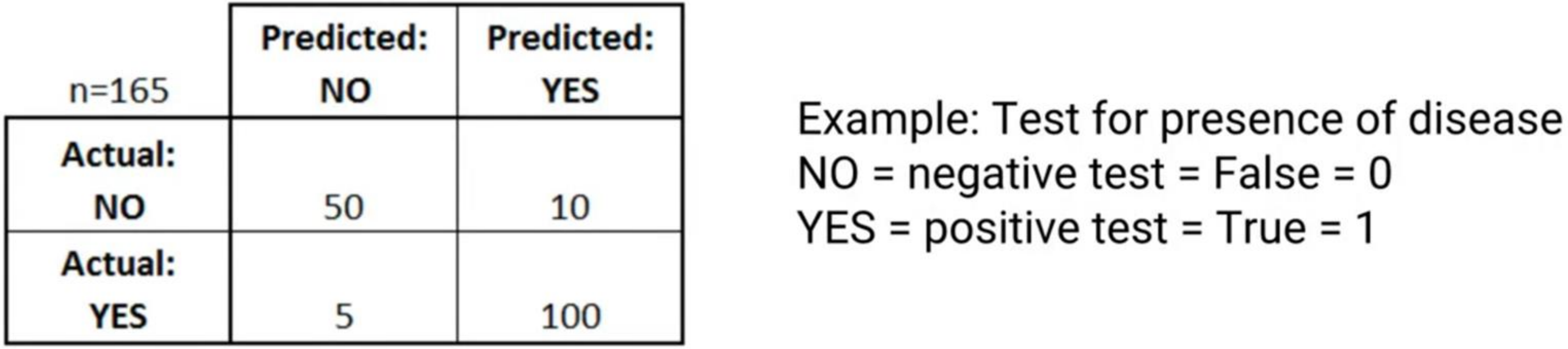
Logistic Regression

- □ After you train a classification model on some training data, you will evaluate your model’s performance on some test data
- □ You can use a confusion matrix to evaluate classification models

Logistic Regression

Logistic Regression

- A confusion matrix can be used to evaluate our model
- Example: Model evaluation on disease classifier



n=165	Predicted: NO	Predicted: YES
Actual: NO	50	10
Actual: YES	5	100

Example: Test for presence of disease
NO = negative test = False = 0
YES = positive test = True = 1

Main difference

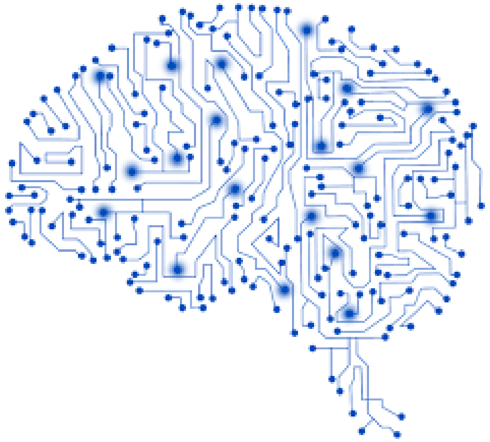
Linear and Logistic regression are the most basic form of regression which are commonly used.

- The essential difference between these two is that Logistic regression is used when the dependent variable is binary in nature;
- Linear regression is used when the dependent variable is continuous and nature of the regression line is linear;
- In Linear regression data is modeled using a straight line;
- In Logistic regression, the probability of some obtained event is represented as a linear function of a combination of predictor variables.

Bibliography

- Montgomery, Douglas C., Elizabeth A. Peck, and G. Geoffrey Vining. Introduction to linear regression analysis. John Wiley & Sons, 2021.
- Ranganathan, Priya, C. S. Pramesh, and Rakesh Aggarwal. "Common pitfalls in statistical analysis: logistic regression." Perspectives in clinical research 8.3, 2017.

University of Minho
School of Engineering



Dados e Aprendizagem Automática

Supervised Learning

Linear & Logistic Regression

DAA @ MEI/1º ano – 1º Semestre