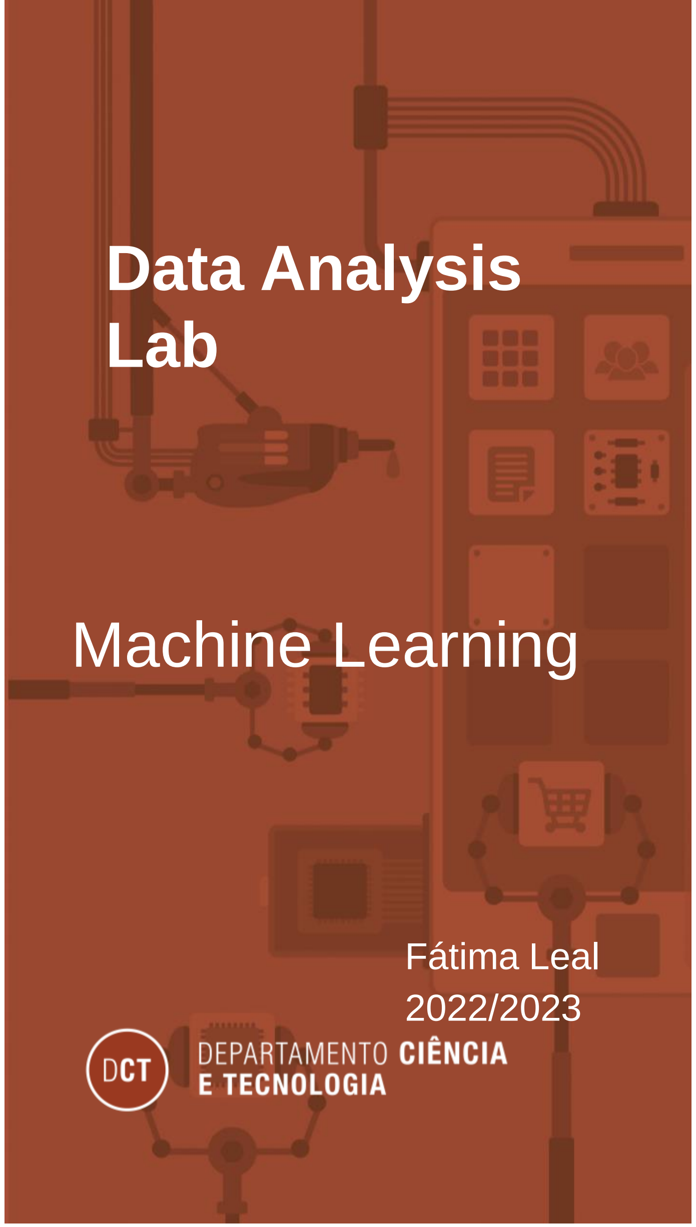


Data Analysis Lab

Machine Learning

Fátima Leal

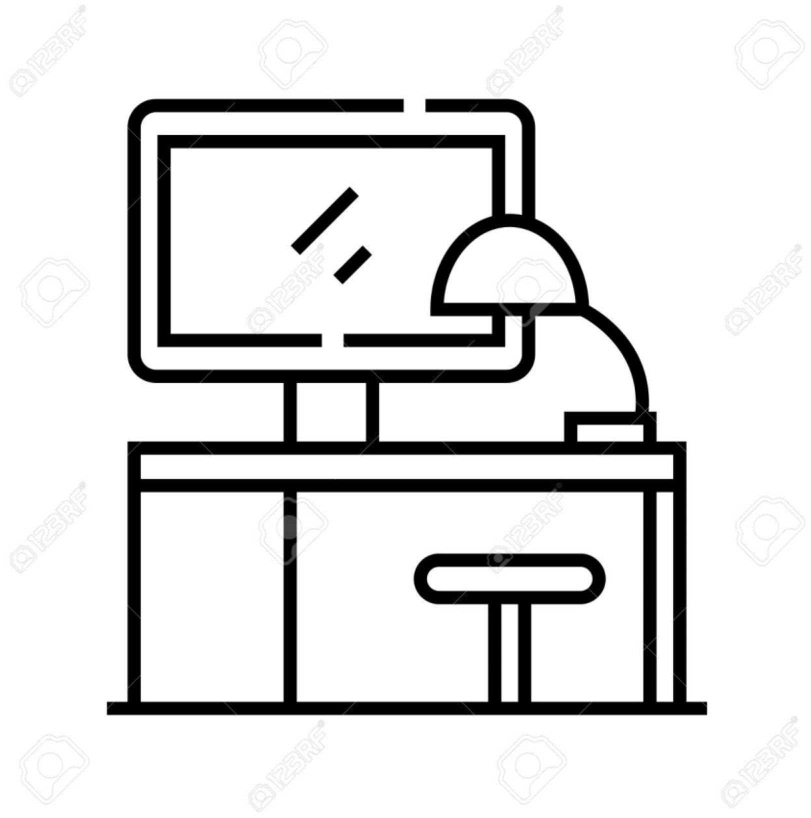
2022/2023

DEPARTAMENTO **CIÊNCIA E TECNOLOGIA**

UNIVERSIDADE PORTUCALENSE

Previous Lesson

- Introduction to data analysis
- Statistical measures
- Graph Analysis



Outline

- Machine Learning
  - Introduction
  - Supervised Learning
  - Training, validation, and test sets
  - Cross Validation

- Regression Models
  - Linear Models
  - Practical examples
  - Logistic Models
  - Practical examples

Introduction

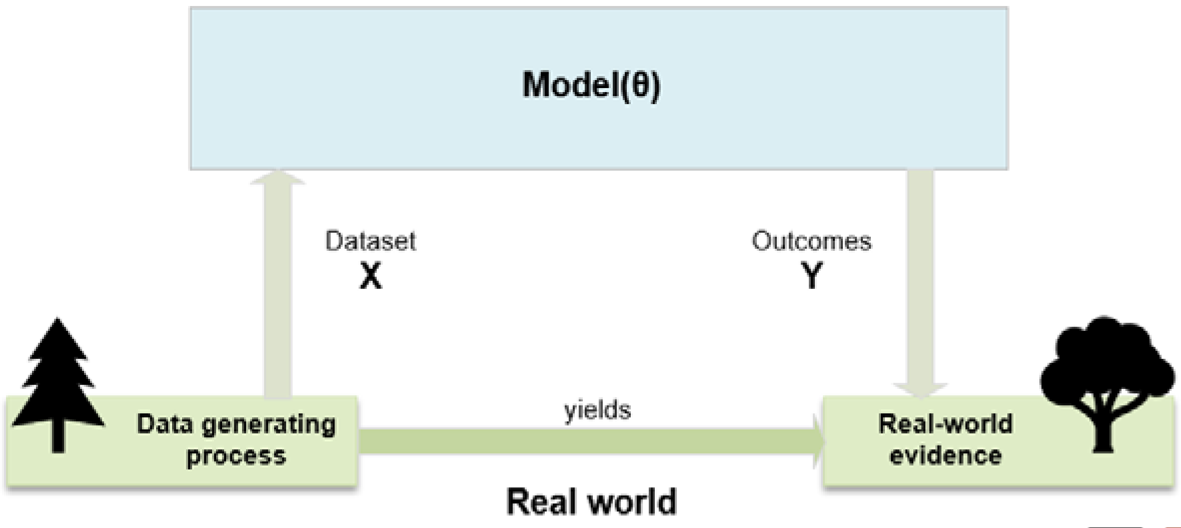
- Machine learning models are mathematical tools that allow us to uncover synthetic representations of external events, with the purpose of gaining better **understanding and predicting future behavior**.

- They create associations, find out relationships, discover patterns, generate new samples, and more, working with well-defined datasets, which are homogenous collections of data points.

Introduction

- We can think of a model as a gray box, where a vector input X extracted from a dataset is transformed into a vector output Y:

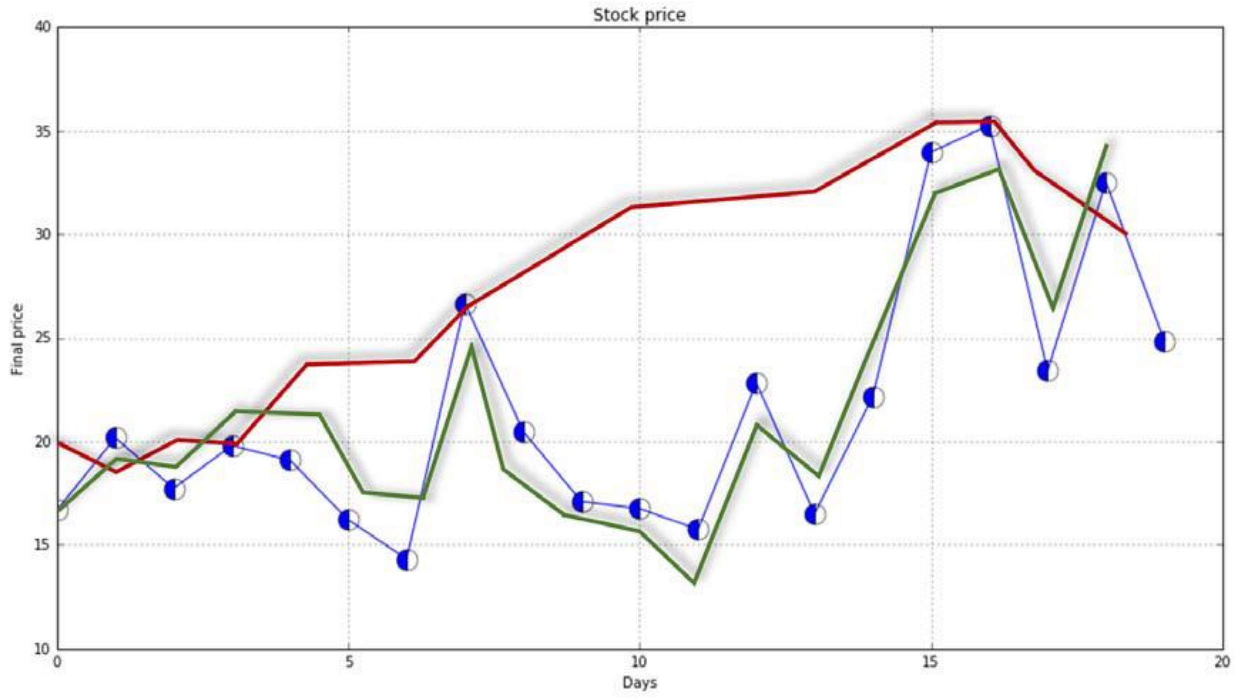
- The dataset is represented by data extracted from a real-world scenario, and the outcomes provided by the model must reflect the nature of the actual relationships. These conditions are very strong in logic and probabilistic contexts, where the inferred conditions must reflect natural ones.



Supervised Learning

- A supervised scenario is characterized by the concept of a teacher or supervisor, whose main task is to provide the agent with a precise measure of its error.

- With input and expected output, the agent can correct its parameters so as to reduce the magnitude of a global loss function.



DEPARTAMENTO **CIÊNCIA E TECNOLOGIA**

Supervised Learning

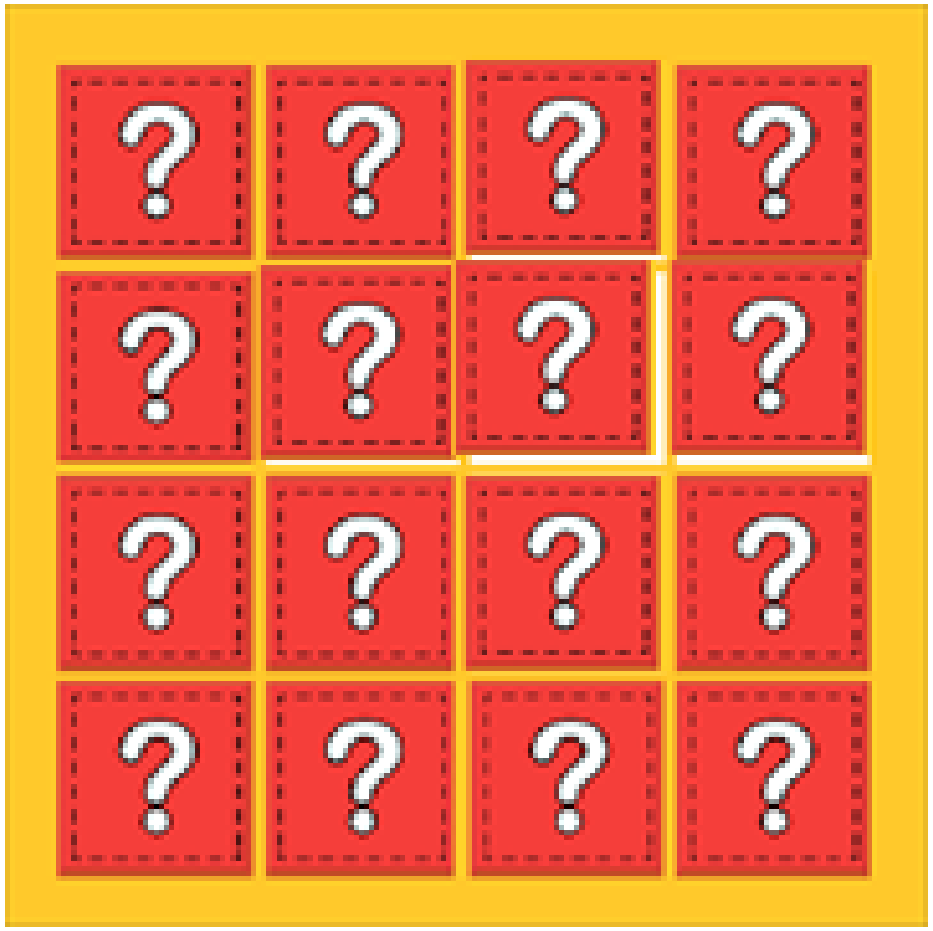
- After each iteration, if the algorithm is flexible enough and data elements are coherent, the overall accuracy increases and the **difference between the predicted and real value becomes close to zero**

- The goal is training a system that must also work with **samples never seen before**.

- It’s necessary to allow the model to develop a generalization ability and avoid a common problem called **overfitting**, which causes an overlearning due to an excessive capacity.

Supervised Learning - Overfitting

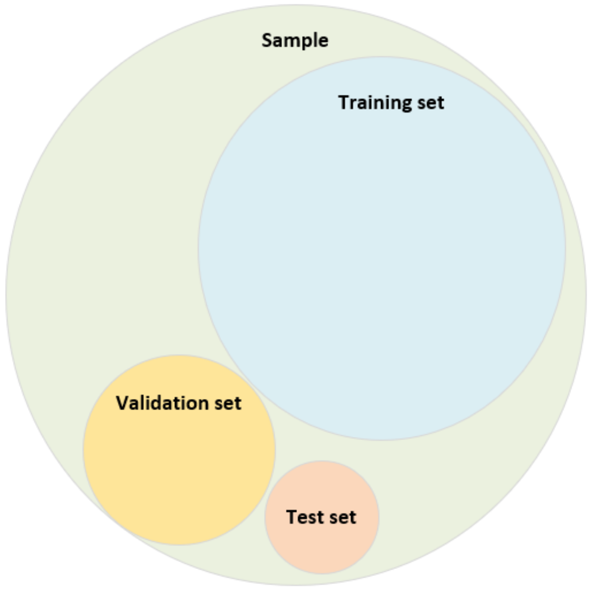
Supervised Learning - Overfitting



Where is the ducks?

Training, validation, and test sets

- It’s usually necessary to split the initial set X, together withY, each of them containing N independente and identically distributed elements sampled from p_{data}, into two or three subsets as follows:
  - **Training** set used to train the model
  - **Validation** set used to assess the score of the model without bias, with samples never seen before
  - **Test** set used to perform the final validation before moving to production



Training, validation, and test sets

- The choice of using two (training and validation) or three (training, validation, and test) sets is normally related to the specific context. In many cases, a single validation set, which is often called the test set, is used throughout the whole process.

- Depending on the nature of the problem, it’s possible to choose a split percentage ratio of 70 % – 30 %, which is a good practice in machine learning, where the datasets are relatively small, or a higher training percentage of 80%, 90%, or up to 99% for deep learning tasks where the numerosity of the samples is very high.

- Shuffling the sets is always good practice, in order to reduce the correlation between samples.

Training, validation, and test sets

Example:

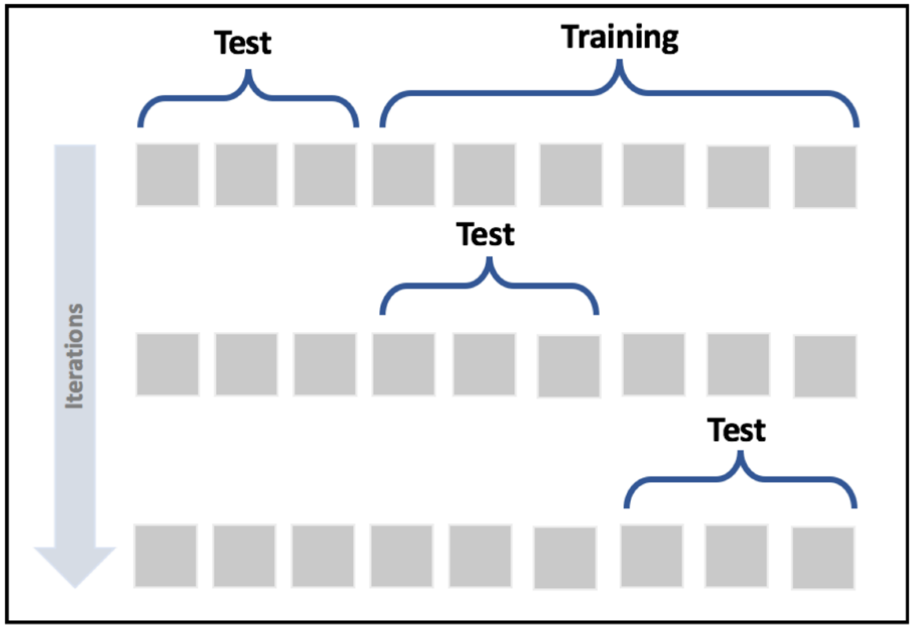
Data frame with dataset variables

```
from sklearn.model_selection import train_test_split
import pandas as pd

df=pd.read_csv("data.csv")

X_train, X_test, Y_train, Y_test = train_test_split(df.Input,
                                                    df.output, test_size=0.3)
```

Data frame with target variable

Cross-validation

- A method to detect the problem of wrongly selected test sets is provided by the cross-validation (CV) technique.
- The idea is to split the whole dataset X into a moving test set and a training set made up of the remaining part. The size of the test set is determined by the number of folds, so that during k iterations, the test set covers the whole original dataset.



Cross-validation

Example

```
from sklearn.model_selection import cross_val_score

clf = Model...

scores = cross_val_score(clf, X, y, cv=5)

scores or scores.mean() or scores.std()
```

Machine Learning tasks

- **Prediction** is a learning task which is able to predict value in several contexts
- **Classification** assigns data to categories or classes
- **Examples**:
  - Sales number
  - House Prices
  - Animal category
  - Name of flower
  - Cancer categories
  - Sudent performance
  - Diabetes in pregnants
  - Human or non-human
  - Recommended or non-recommended

Regression Models

- A regression is a model that associates an input vector, $\bar{x} \in \mathbb{R}$ with one or more continuous dependent variables y $\in \mathbb{R}$.
- In a general scenario, there’s no explicit dependence on time, even if regression models are often employed to model time series.
- Linear Regression - Predictions
- Logistic Regression - Classifications
- Some Linear models require normalisation

Linear Regression

- Linear models are the simplest parametric methods and always deserve the right attention, because many problems, even intrinsically non-linear ones, can be easily solved with these models.

- Consider a dataset of real-values vectors: $X = \{\overline{x_1}, \overline{x_2}, \ldots.., \overline{x_n}\}$ where $\overline{x_n} \in \mathbb{R}^2$.

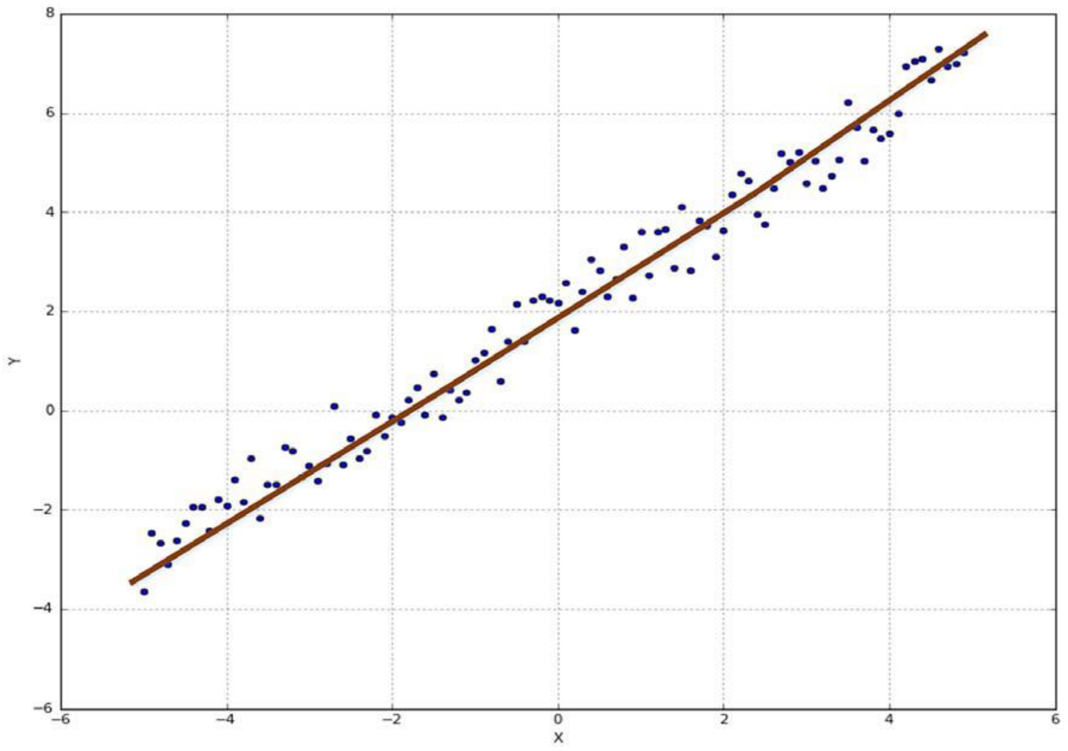
- Each input vector is associated with a real value yi: $\mathrm{Y} = \{\overline{y_1}, \overline{y_2}, \ldots.., \overline{y_n}\}$ where $\overline{y_i} \in \mathbb{R}^2$.

- A linear model is based on the assumption that it's possible to approximate the output values through a regression process based on the rule: $y = \alpha_0 + \sum_{i=1}^{m} \alpha_i x_i\}$

Linear Regression

- In the following figure, there's a plot with a candidate regression function:

y = x +2+n



DEPARTAMENTO **CIÊNCIA E TECNOLOGIA**

Example: Linear Regression

- For this example, we will use the Boston dataset, which contains data about the housing and price data in the Boston area.
- What’s the data?
- http://lib.stat.cmu.edu/datasets/boston
- What’s the goal of machine learning here?

Variables Boston Data set

CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT	% lower status of the population
MEDV	Median value of owner-occupied homes in $1000's

Example: Linear Regression

```
from sklearn import datasets
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression

boston = datasets.load_boston()
X_train, X_test, Y_train, Y_test = train_test_split(boston.data,
boston.target, test_size=0.3)

lr = LinearRegression(normalize=True)

lr.fit(X_train, Y_train)
print(lr.score(X_test,Y_test))
```

Exercise: Linear Regression

- Use the dataset available on Moodle ”USA Housing”
- Do an exploratory/statistical analysis of the dataset
- Divide the dataset in train and test
- Apply a linear regression model
- Obtain the mean and std of the cross-validation scores

- Predict the Price of a house with:
  - Avg. Area Income: 79550
  - Avg. Area House Age: 5.68
  - Avg. Area Number of Rooms: 8
  - Avg. Area Number of Bedrooms: 4.09
  - Area Population: 23087

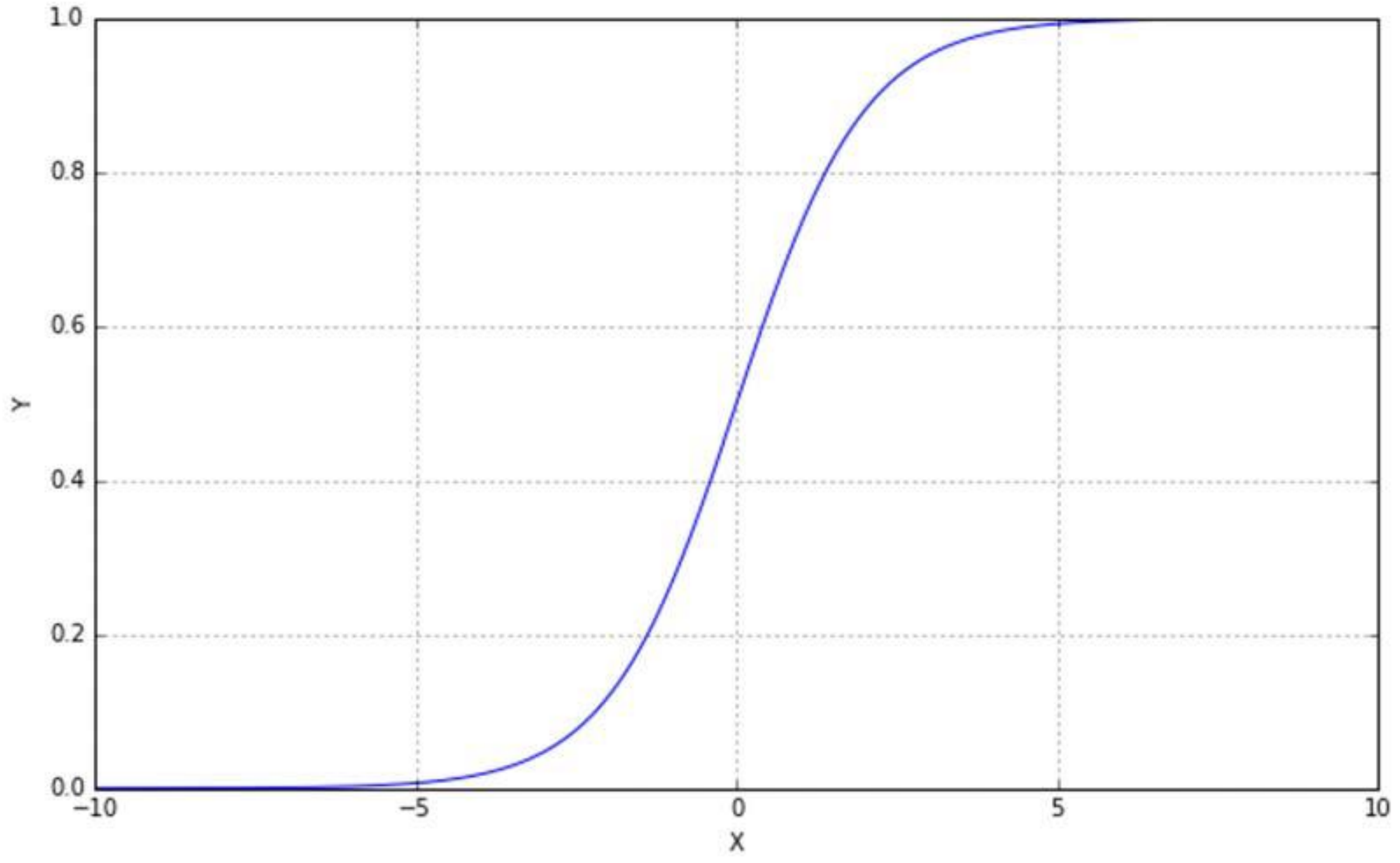
Logistic Regression

- This is a **classification** method which is based on the probability for a sample to belong to a class.

- As our probabilities must be continuous in R and bounded between (0, 1), it's necessary to introduce a threshold function to filter the term z.

- The name logistic comes from the decision to use the sigmoid (or logistic) function: $\sigma(z) = \frac{1}{1+e^{-z}}$ which becomes $\sigma(\bar{x}, \bar{w}) = \frac{1}{1+e^{-\bar{x}.\bar{w}}}$

Logistic Regression



Logistic Regression

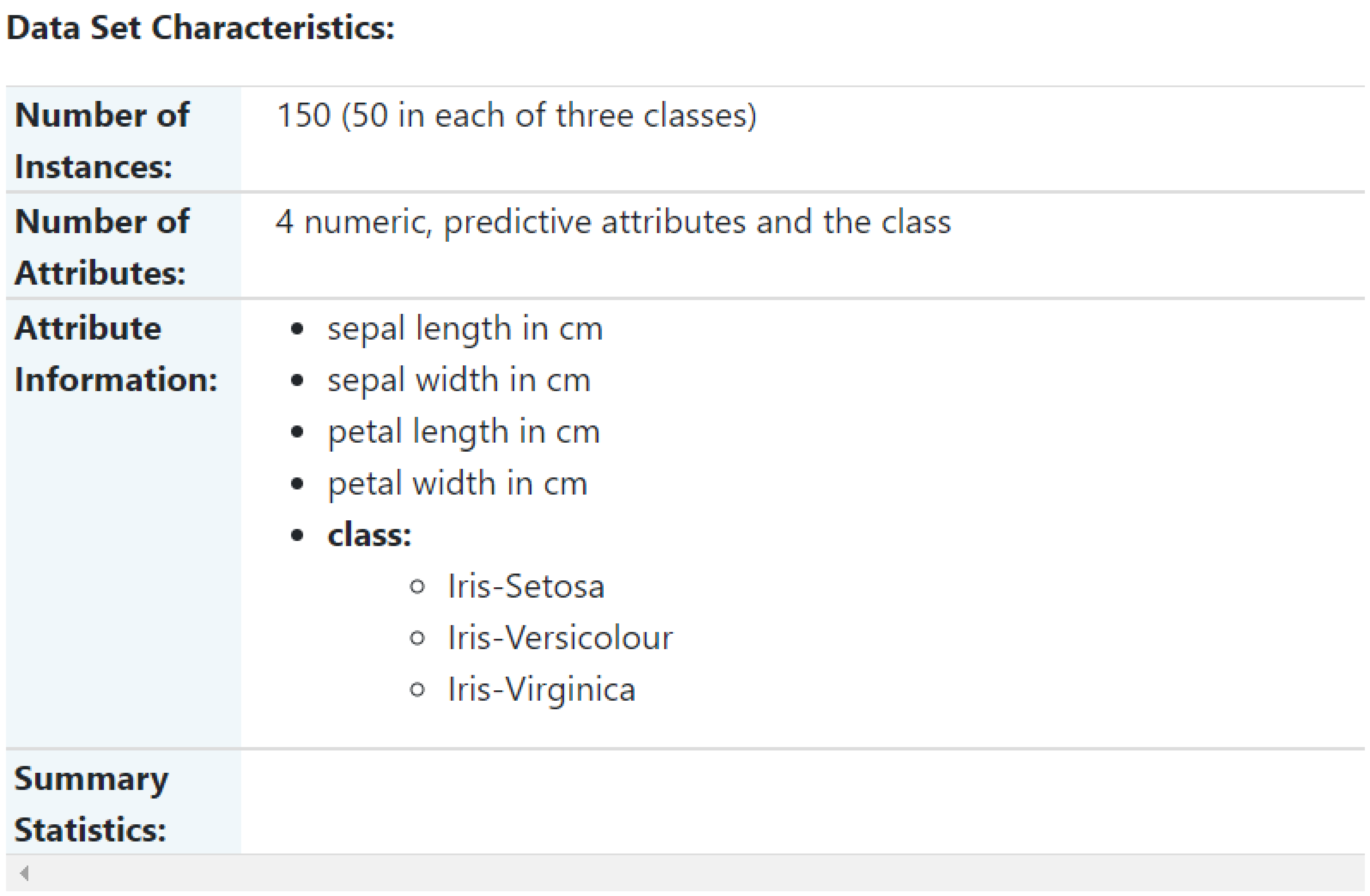
- We can define the probability for a sample to belong to a class (from now on, we'll call them 0 and 1 as $P(y|\bar{x}) = \sigma(\bar{x};\ \overline{w})$

- Finding the optimal parameters is equivalent to maximizing the log-likelihood given the output class: $\mathrm{L}(\overline{w}; y) = logP(y|\overline{w})$

Logistic Regression - Examples

- For this exemple we will use iris dataset which contains data about flowers.
- Whats the data?
- https://scikit-learn.org/stable/datasets/toy_dataset.html
- What’s the goal of machine learning here?

Logistic Regression - Examples



Data Set Characteristics:

Number of Instances:	150 (50 in each of three classes)
Number of Attributes:	4 numeric, predictive attributes and the class
Attribute Information:	• sepal length in cm • sepal width in cm • petal length in cm • petal width in cm • **class:** ○ Iris-Setosa ○ Iris-Versicolour ○ Iris-Virginica
Summary Statistics:	

Logistic Regression - Example

LogisticRegression parameters sklearn

```
from sklearn.datasets import load_iris
from sklearn.linear_model import LogisticRegression
from sklearn.model_selection import train_test_split

X, y = load_iris(return_X_y=True)
X_train, X_test, Y_train, Y_test = train_test_split(X,y, test_size=0.5)
clf = LogisticRegression(random_state=10)
clf.fit(X_train, Y_train)

print(clf.score(X_test,Y_test))
```

Exercise

1. Use the dataset available on Moodle ”Admission prediction”
2. Do a exploratory/statistical analysis of the dataset
3. Divide the dataset in train and test
4. Apply a logistic regression model
5. Obtain the scores

6. Predict de admission of a student with:
    1. GRE Score: 330
    2. TOEFL Score: 100
    3. University Rating: 4
    4. SOP: 5
    5. LOR:4
    6. CGPA:9
    7. Research: 1

UNIVERSIDADE
PORTUCALENSE

Do conhecimento à prática.