Intro to Machine Learning Workshop

Day 1

Dr.Reem Alotaibi

February 20th 2022

Outline

Outline

Introduction

Workshop Description

- The workshop is a 3-day.
- The workshop will cover introduction to machine learning using R (by Dr.Reem Alotaibi).

Outline

Machine Learning

Making Sense of Data

Definition

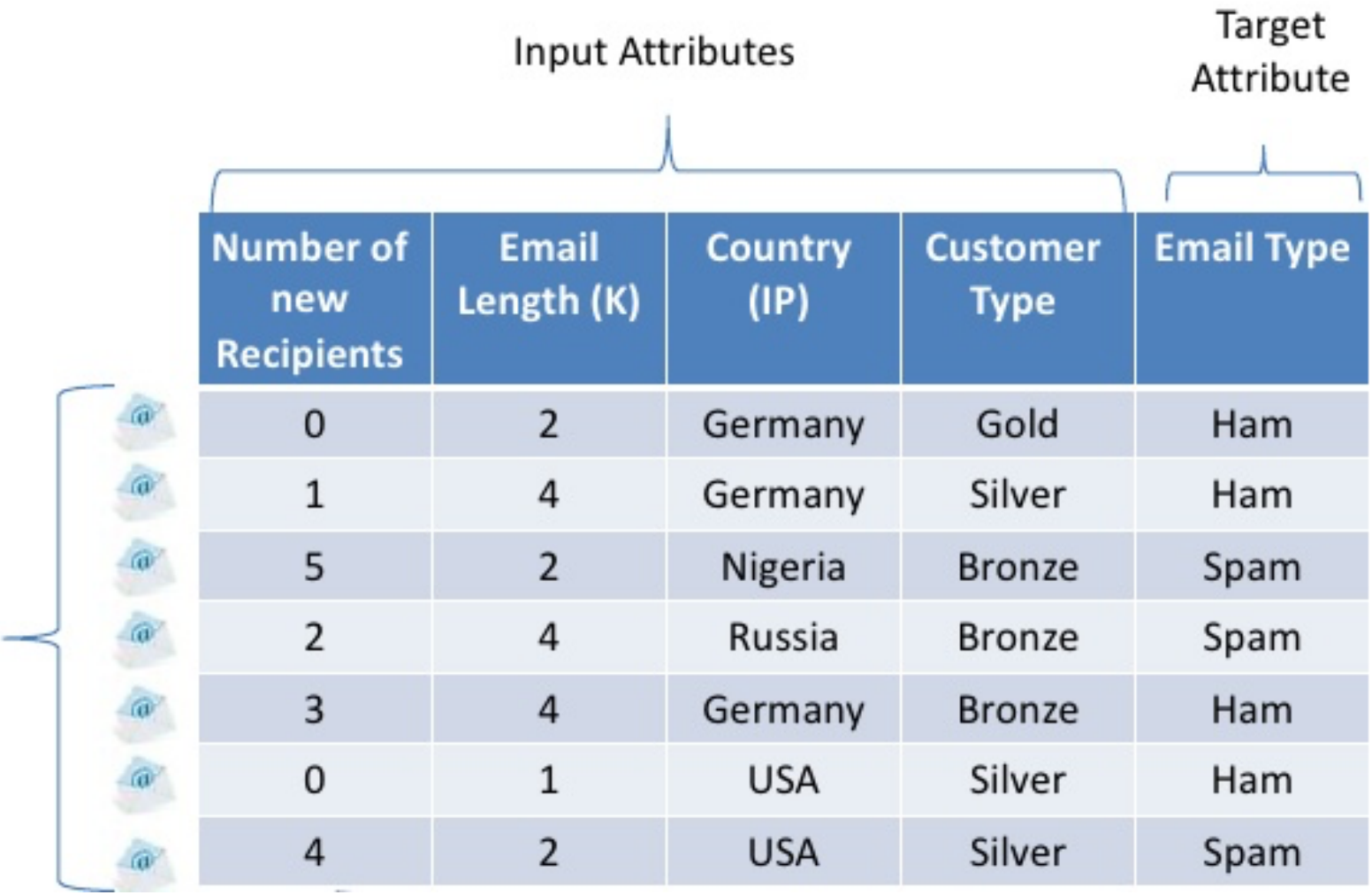
- A branch of artificial intelligence.
- Algorithms that learn from data and previous experience.

Motivating Example: Spam Filtering

- **Task:** Spam all emails that the user dose not want.
- **Experience:** A database of previous emails that were labelled by the user.

Machine Learning

Making Sense of Data



Number of new Recipients	Email Length (K)	Country (IP)	Customer Type	Email Type
0	2	Germany	Gold	Ham
1	4	Germany	Silver	Ham
5	2	Nigeria	Bronze	Spam
2	4	Russia	Bronze	Spam
3	4	Germany	Bronze	Ham
0	1	USA	Silver	Ham
4	2	USA	Silver	Spam

Machine Learning Tasks

	Supervised Learning	**Unsupervised Learning**
Discrete	Classification • Trees	Clustering • K-means
Continues	Regression • Linear regression	Dimensionality Reduction • PCA

Machine Learning Tasks

Classification

Input space

Features, attributes, variables, covariate

Training set

	Age	Gender	Glucose	Diabetics
1	65	male	120	Yes
2	35	female	180	Yes
3	55	female	150	Yes
4	37	male	90	No
5	25	female	105	No
6	56	male	125	?

Output space

Class, label, target, dependent variable

Classification: Categorical or discrete target

Machine Learning Tasks

Regression

Input space

Features, attributes, variables, covariate

Training set

	Month	Temp C	Ice cream sales SAR/Day
1	Jan	15	501
2	Feb	20	550
3	March	25	600
4	April	27	900
5	May	30	1050
	June	35	?

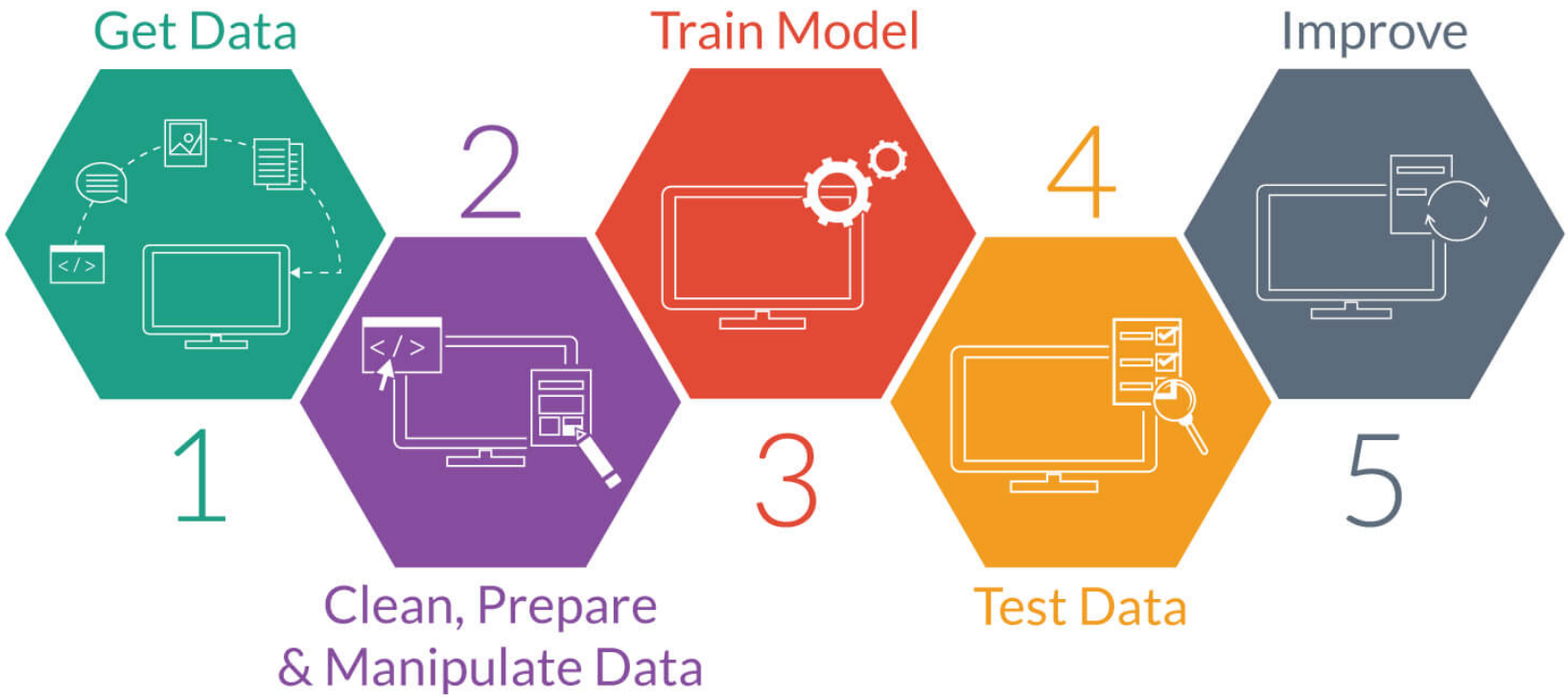
Output space

Target

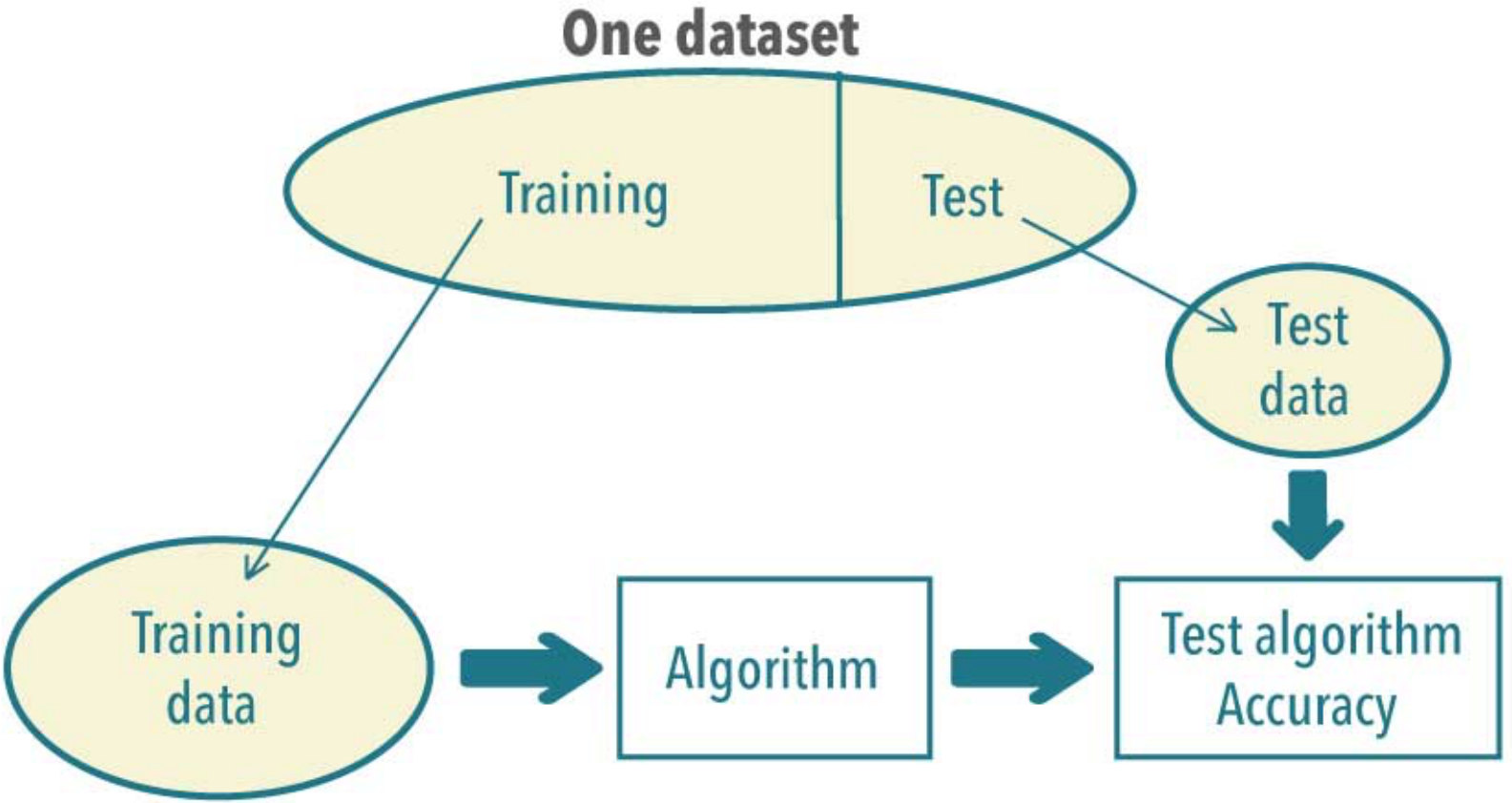
dependent variable

Regression: Numerical or continues target

Machine Learning Process



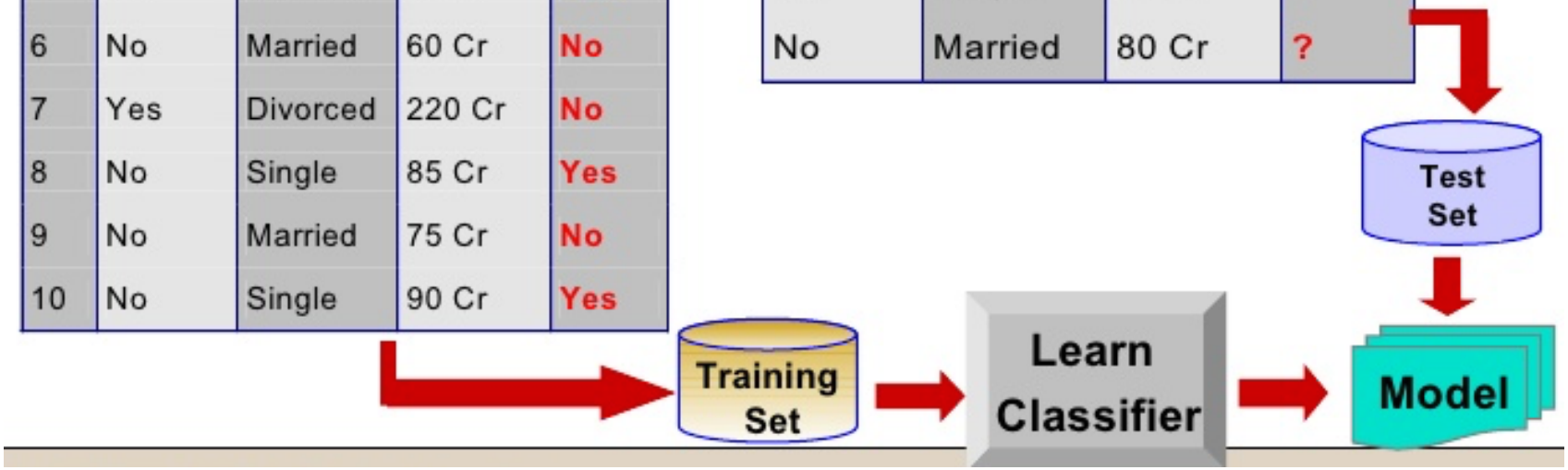
Training VS Test Data



Training VS Test Data

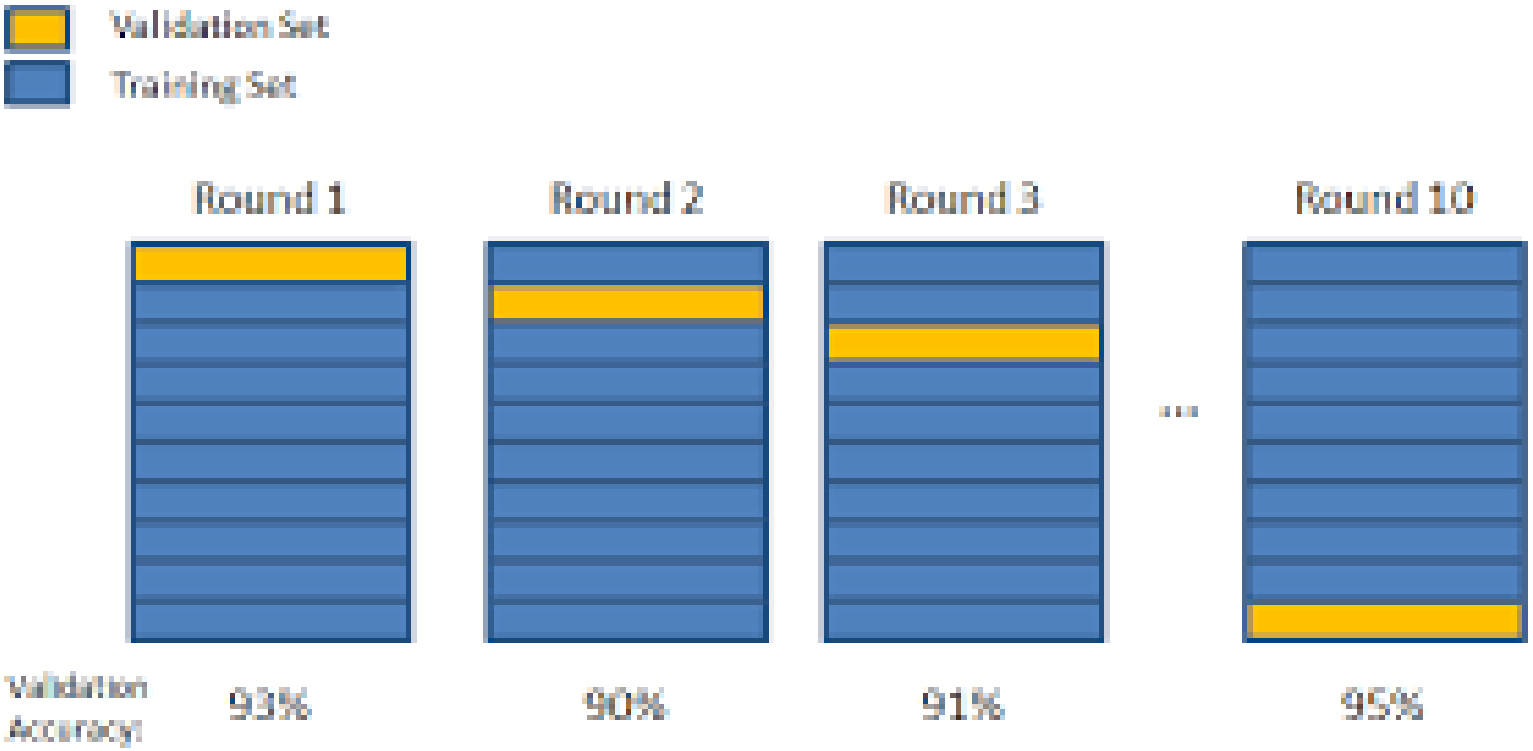
Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125 Cr	No
2	No	Married	100 Cr	No
3	No	Single	70 Cr	No
4	Yes	Married	120 Cr	No
5	No	Divorced	95 Cr	Yes
6	No	Married	60 Cr	No
7	Yes	Divorced	220 Cr	No
8	No	Single	85 Cr	Yes
9	No	Married	75 Cr	No
10	No	Single	90 Cr	Yes

Refund	Marital Status	Taxable Income	Cheat
No	Single	75 Cr	?
Yes	Married	50 Cr	?
No	Married	150 Cr	?
Yes	Divorced	90 Cr	?
No	Single	40 Cr	?
No	Married	80 Cr	?



K-Cross Validation

Example: 10-cross validation



Final Accuracy = Average(Round 1, Round 2, ...)

Machine Learning Applications

Machine Learning Applications

Ranking

Machine Learning | Coursera
https://www.coursera.org/learn/machine-learning
About this course: **Machine learning** is the science of getting computers to act without being explicitly programmed. In the past decade, **machine learning** has ...

Machine Learning: What it is and why it matters | SAS
https://www.sas.com/en_sa/insights/analytics/machine-learning.html
Machine learning is a method of data analysis that automates analytical model building. Using algorithms that iteratively learn from data, **machine learning** allows computers to find hidden insights without being explicitly programmed where to look.

What Is The Difference Between Artificial Intelligence And Machine ...
https://www.forbes.com/.../what-is-the-difference-between-artificial-intelligence-and-...
Dec 6, 2016 - Artificial Intelligence (AI) and **Machine Learning** (ML) are two very hot buzzwords right now, and often seem to be used interchangeably. ... Artificial Intelligence is the broader concept of **machines** being able to carry out tasks in a way that we would consider "smart".

The 10 Algorithms Machine Learning Engineers Need to Know
https://www.kdnuggets.com/2016/08/10-algorithms-machine-learning-engineers.html
It is no doubt that the sub-field of **machine learning** / artificial intelligence has increasingly gained more popularity in the past couple of years. As Big Data is the ...

Machine Learning | Udacity
https://www.udacity.com/course/machine-learning--ud262
Machine Learning is a graduate-level course covering the area of Artificial Intelligence concerned with computer programs that modify and improve their ...

Machine Learning | Microsoft Azure
https://azure.microsoft.com/en-us/services/machine-learning-studio/
Get started now with Azure **Machine Learning** for powerful cloud-based analytics, now part of Cortana Intelligence Suite.

Machine Learning Applications

Abnormality Detection: Credit Card Fraud

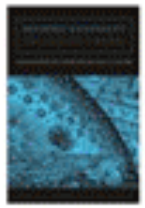
Machine Learning Applications

Collaborative Filtering

Customers Who Bought This Item Also Bought

Reckoning with Risk: Learning to Live with Unce... by Gerd Gigerenzer
★★★★☆ (8) £6.49

Gut Feelings: The Intelligence of the Unconscious by Gerd Gigerenzer
£10.27



Bounded Rationality: The Adaptive Toolbox (Dahl... by G Gigerenzer
£20.95

What Do Customers Ultimately Buy After Viewing This Item?

68% buy
Simple Heuristics That Make Us Smart (Evolution & Cognition)
£18.99

17% buy
Gut Feelings: Short Cuts to Better Decision Making
£6.74

9% buy
Influence: The Psychology of Persuasion ★★★★☆ (12)
£7.09

Play with Machine Learning

Datasets

- UCI: https://archive.ics.uci.edu/ml/datasets.php
- KEEL: http://sci2s.ugr.es/keel/datasets.php
- Kaggle: https://www.kaggle.com/datasets
- LIBSVM: https://www.csie.ntu.edu.tw/∼cjlin/libsvmtools/datasets/
- U.S. open data: https://www.data.gov/
- Saudi open data: http://data.gov.sa/en/home

Languages

- R
- Python (Scikit-learn for ML, Theano for deep learning)
- MATLAB

Play with Machine Learning

Online Platforms

- Kaggle: https://www.kaggle.com/
- OpenML: https://www.openml.org

Tools with GUI

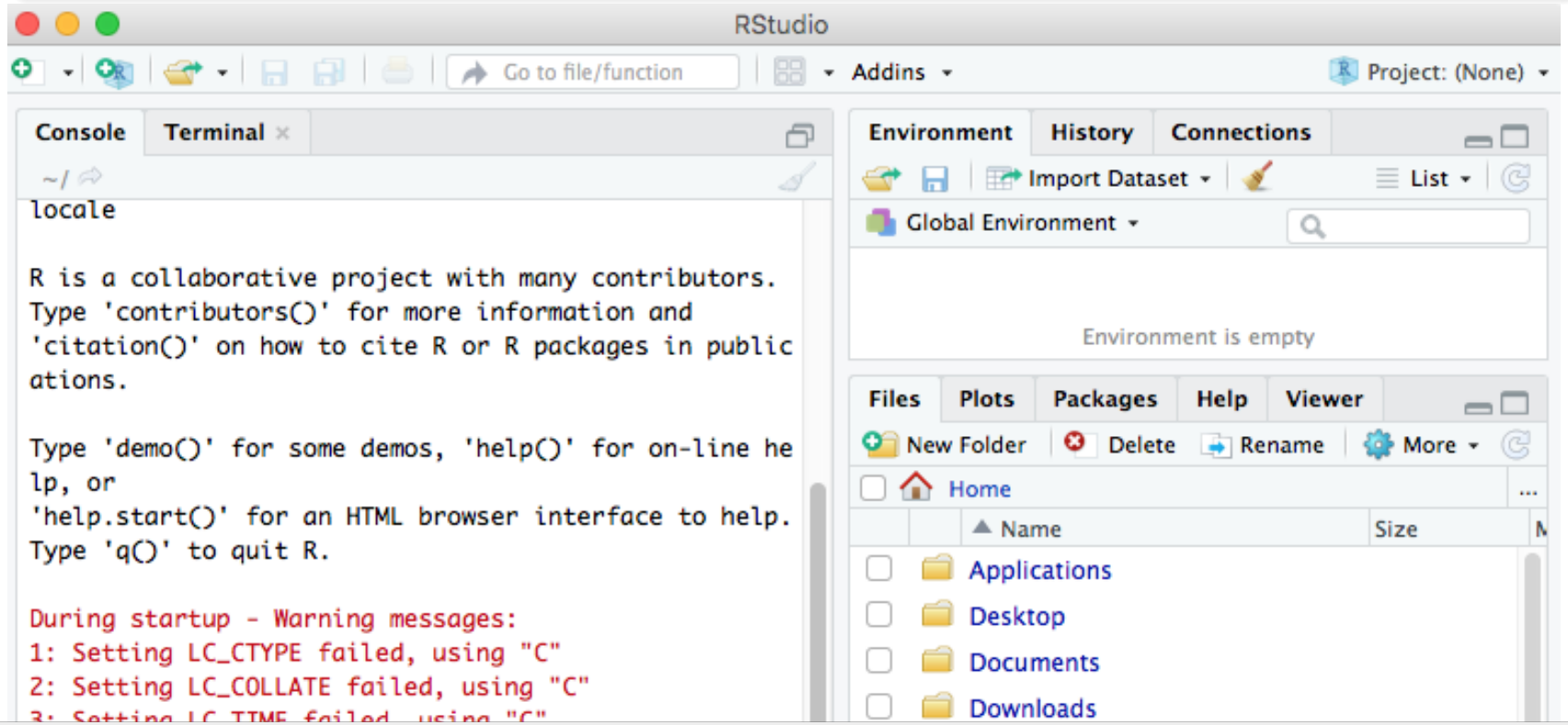
- Weka
- RapidMiner
- Orange
- Neuro Solution

Outline

R for Machine Learning

Getting Started

- The R Project website is http://www.r-project.org/.
- RStudio is a set of integrated tools designed to help you be more productive with R.



R for Machine Learning

Installing and loading packages

- install.packages("package name")
- To load the package type: library(package name)

Example: e1071 package to use SVMs

- install.packages("e1071") // only once
- Library(e1071) //each time you start R

Basic Functions

Dataset

Loading dataset from R or PC

- A list of datasets available in R $\rightarrow$ data()
- Load a particular dataset from R $\rightarrow$ data(iris)
- Load dataset from your pc $\rightarrow$ myData$<-$read.csv("/Users/reem/Desktop/iris.csv")

Basic Functions

Dataset

Loading dataset from url

- library(curl)
- urlfile $<-$'https://archive.ics.uci.edu/ml/machine-learning-databases/auto-mpg/auto-mpg.data'
- test $<-$read.csv(curl(urlfile))

Basic Functions

Dataset

Summarizing the dataset

- View dataset type name → iris
- Preview the first 6 rows → head(iris)

```
> iris
    Sepal.Length Sepal.Width Petal.Length Petal.Width     Species
1            5.1         3.5          1.4         0.2      setosa
2            4.9         3.0          1.4         0.2      setosa
3            4.7         3.2          1.3         0.2      setosa
4            4.6         3.1          1.5         0.2      setosa
5            5.0         3.6          1.4         0.2      setosa
6            5.4         3.9          1.7         0.4      setosa
```

Basic Functions

Dataset

Summarizing the dataset

- Find the dimensions of the dataset → dim(iris)
- A quick overview of the dataset → str(iris)

```
> str(iris)
'data.frame':   150 obs. of  5 variables:
 $ Sepal.Length: num  5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...
 $ Sepal.Width : num  3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...
 $ Petal.Length: num  1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...
 $ Petal.Width : num  0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...
 $ Species     : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...
>
```

Basic Functions

Dataset

Summarizing the dataset

- Display the summary of the dataset → summary(iris)

```
> summary(iris)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
 Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica :50
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
>
```

Basic Functions

Partitioning Dataset

Train-Test partitioning

- Use 75% of data to train and the remaining for testing the model.

```
1  library(caret)
2  validation_index <- createDataPartition(iris$Species, p=0.75, list=FALSE)
3  # use the remaining 80% of data to training and testing the models
4  train_split <- iris[validation_index,]
5  # select 20% of the data for validation
6  test_split <- iris[-validation_index,]
```

Basic Functions

Partitioning Dataset

10-fold Cross Validation

- Split the data into 10 folds and use these folds to split the data.

```
1  # 10-fold cross validation
2  folds <- createFolds(iris$Species, k=10)
3  str(folds)
```

3:11 (Top Level)

Console Terminal

~/

```
$ Fold01: int [1:15] 8 17 23 31 49 61 67 79 84 93 ...
$ Fold02: int [1:15] 4 11 18 21 29 58 74 80 81 97 ...
$ Fold03: int [1:15] 2 14 25 44 48 59 70 71 75 98 ...
$ Fold04: int [1:15] 16 19 28 34 47 53 55 72 73 95 ...
$ Fold05: int [1:15] 6 27 36 38 40 64 78 90 91 94 ...
$ Fold06: int [1:15] 5 9 33 37 46 60 63 82 89 92 ...
$ Fold07: int [1:15] 3 20 35 42 45 52 54 56 68 100 ...
$ Fold08: int [1:15] 7 12 13 15 30 57 65 85 86 87 ...
```

Outline

Data Pre-processing

Basic Methods

- “center“: subtract mean from values.
- “scale“: divide values by standard deviation.
- “center“ and “scale“: standardize your data (mean=0, standard deviation=1).
- “range“: normalize values.
- “pca“: transform data to PCA.

Data Pre-processing

Example: Mean-centering

- The center method calculates the mean for an attribute and subtracts it from each value.The resulting variable will have a zero mean.

```
1  # calculate the pre-process parameters from the dataset
2  preprocessParams<-preProcess(iris[,1:4], method="center" )
3  # summarize transform parameters
4  print(preprocessParams)
5  # transform the dataset using the parameters
6  transformed <- predict(preprocessParams, iris[,1:4])
7  # summarize the transformed dataset
8  summary(transformed)
```

4:17 (Top Level)

Console Terminal

~/

```
> summary(transformed)
  Sepal.Length        Sepal.Width         Petal.Length      Petal.Width
 Min.   :-1.54333   Min.   :-1.05733   Min.   :-2.758   Min.   :-1.0993
 1st Qu.:-0.74333   1st Qu.:-0.25733   1st Qu.:-2.158   1st Qu.:-0.8993
```

Data Pre-processing

Example: Standardiztion

- The standardiztion method transforms the variables in such a way they have a zero mean and standard deviation 1.

```
10  # calculate the pre-process parameters from the dataset
11  preprocessParams<-preProcess(iris[,1:4], method=c("center", "scale"))
12  # summarize transform parameters
13  print(preprocessParams)
14  # transform the dataset using the parameters
15  transformed <- predict(preprocessParams, iris[,1:4])
16  # summarize the transformed dataset
17  summary(transformed)
```

16:27 (Top Level)

Console Terminal

~/

```
 Sepal.Length       Sepal.Width        Petal.Length       Petal.Width
Min.   :-1.86378   Min.   :-2.4258   Min.   :-1.5623   Min.   :-1.4422
1st Qu.:-0.89767   1st Qu.:-0.5904   1st Qu.:-1.2225   1st Qu.:-1.1799
Median :-0.05233   Median :-0.1315   Median : 0.3354   Median : 0.1321
```

Outline

Data Visualisation

Univariate Plots

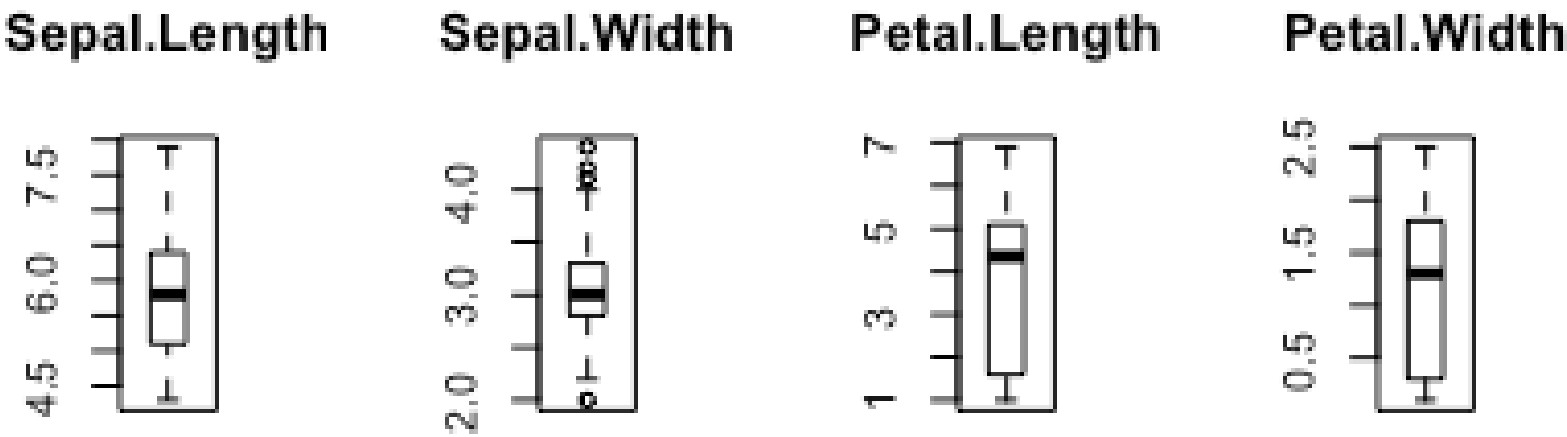
BoxPlot

- It gives you a clearer idea of the distribution of the input attributes.

```
1  # split input and output
2  x <- iris[,1:4]
3  y <- iris[,5]
4
5  # boxplot for each attribute
6  par(mfrow=c(1,4))
7  for(i in 1:4)
8  {
9    boxplot(x[,i], main=names(iris)[i])
10 }
```

Data Visualisation

BoxPlot Example

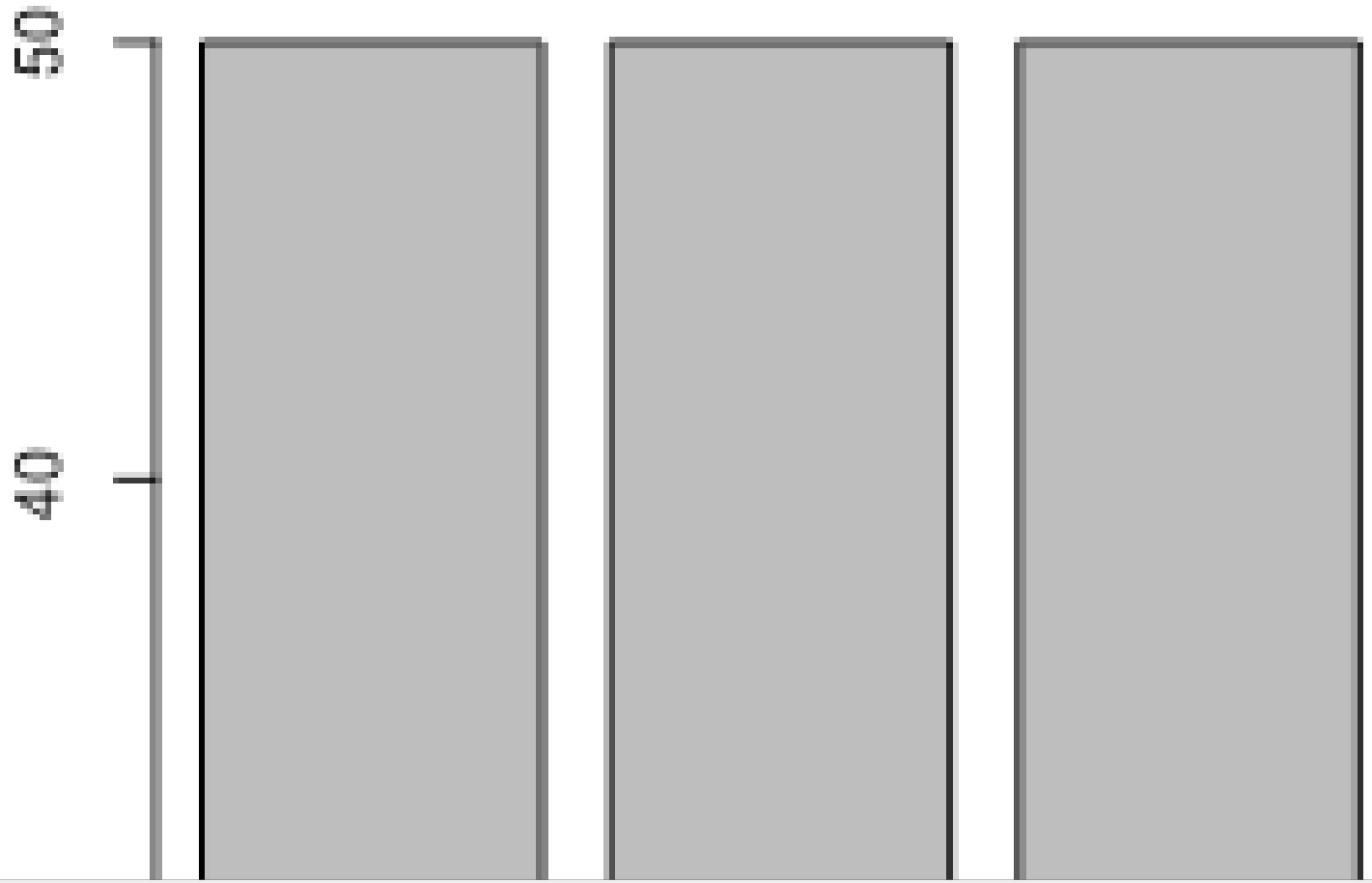


Data Visualisation

Univariate Plots

BarPlot

- Plot class distributions → plot(y)



Data Visualisation

Univariate Plots

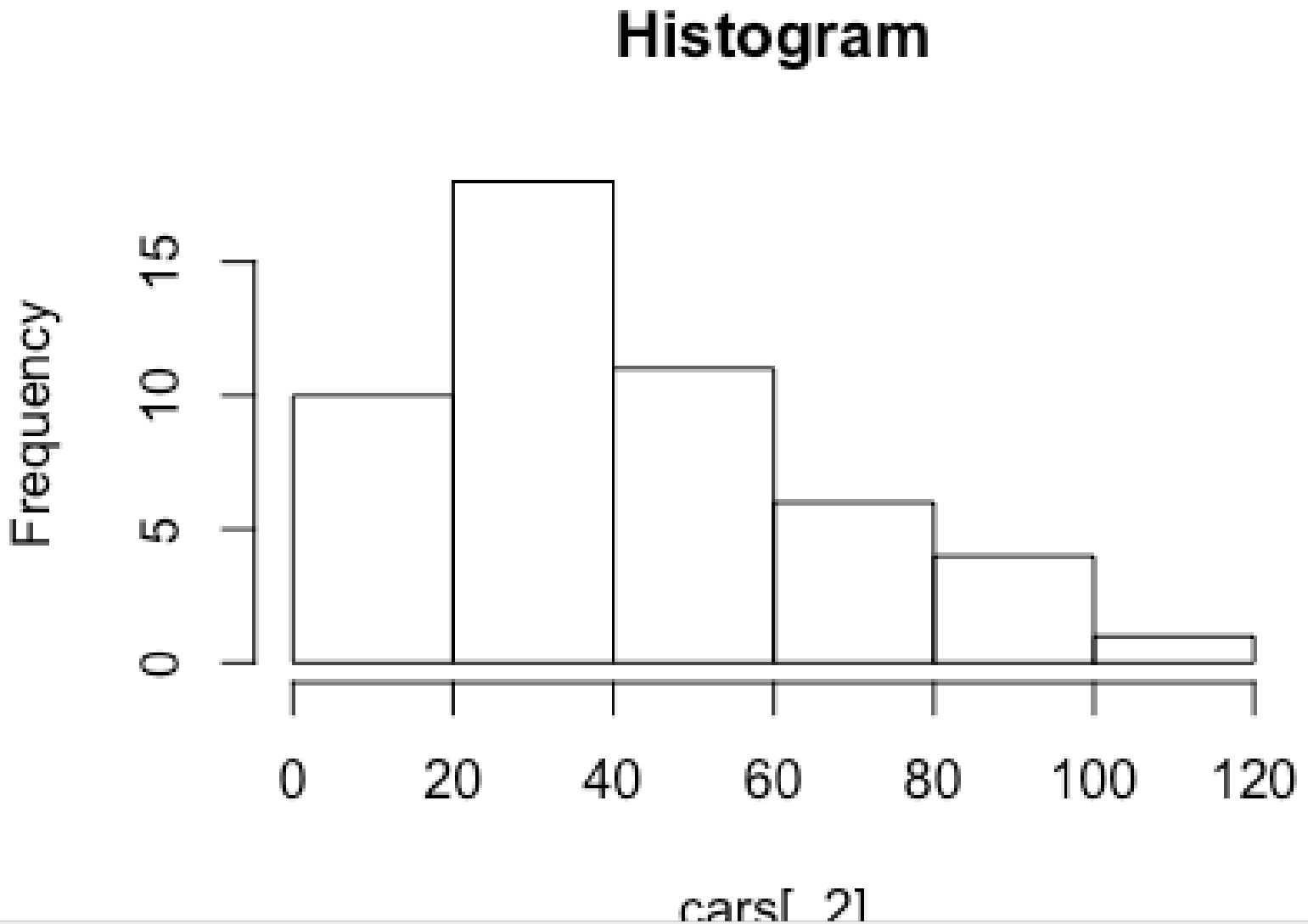
Histogram

- Plot class distributions → hist(y)

```
1  hist(cars[,2], main="Histogram")
```

Data Visualisation

Histogram Example



Data Visualisation

Multivariate Plots

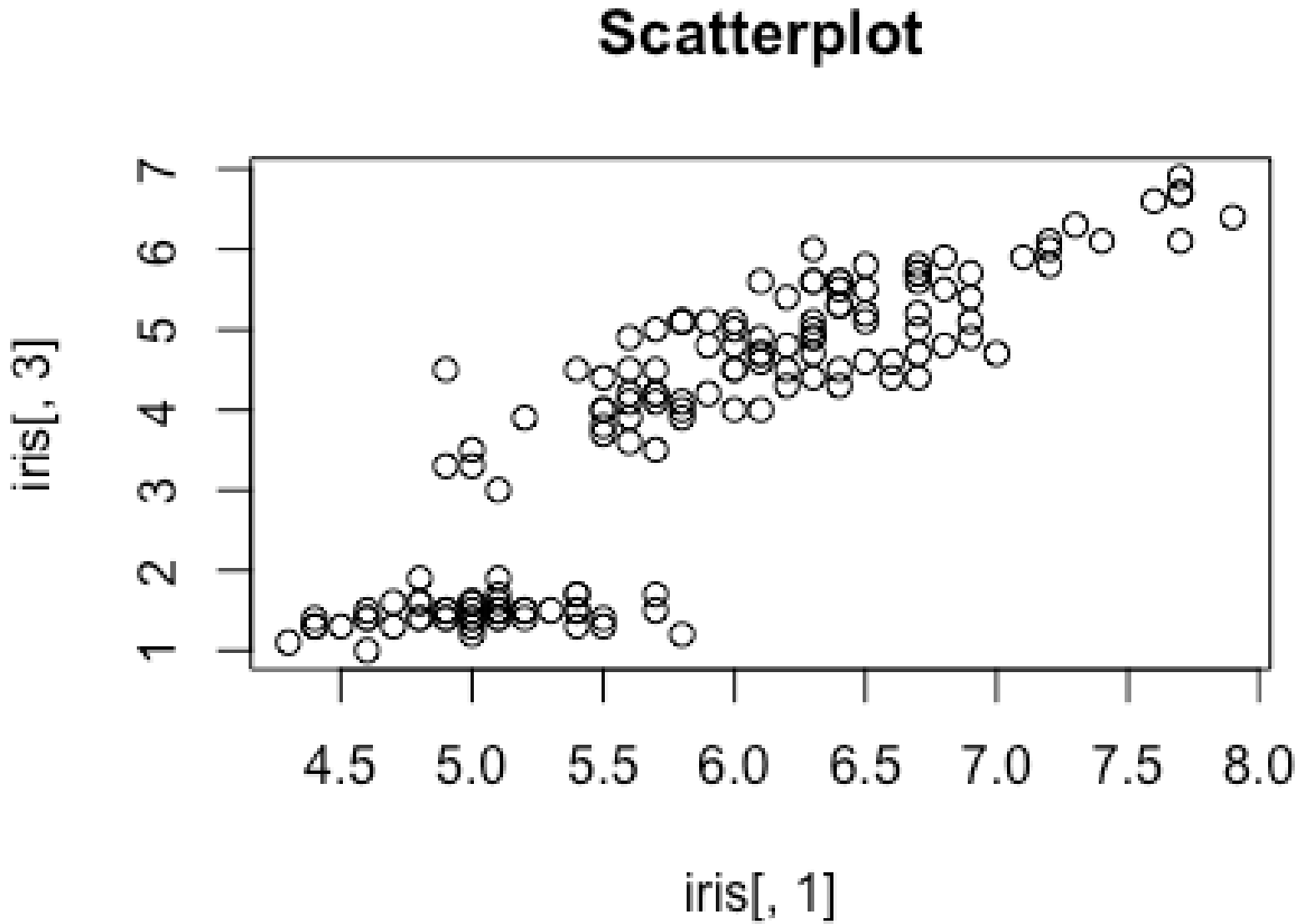
Scatter Plots

- It shows the interaction between attributes.

```
1  plot(iris[,1], iris[,3], main="Scatterplot")
```

Data Visualisation

Scatterplot Example



Outline

Summary

Wrap-up

- step-by-step to start your first machine learning project using R.
- qplot is the simplest choice if you are dealing with input vectors.
- ggplot and ggplot2 for data frames.
- ?FunctionName in R to get help.

Day 2

- Try some machine learning algorithms.