**What is Factor in R?**

Factors are variables in R which take on a limited number of different values; such variables are often referred to as categorical variables.

In a dataset, we can distinguish two types of variables: **categorical** and **continuous**.

* In a categorical variable, the value is limited and usually based on a particular finite group. For example, a categorical variable can be countries, year, gender, occupation.
* A continuous variable, however, can take any values, from integer to decimal. For example, we can have the revenue, price of a share, etc..

**Categorical variables**

R stores categorical variables into a factor. Let's check the code below to convert a character variable into a factor variable. **Characters are not supported in machine learning algorithm, and the only way is to convert a string to an integer.**

factor(x = character(), levels, labels = levels, ordered = is.ordered(x))

Arguments:

- x: A vector of data. Need to be a string or integer, not decimal.

- Levels: A vector of possible values taken by x. This argument is optional. The default value is the unique list of items of the vector x.

- Labels: Add a label to the x data. For example, 1 can take the label `male` while 0, the label `female`.

- ordered: Determine if the levels should be ordered.

Let's create a factor data frame.

# Create gender vector

gender\_vector <- c("Male", "Female", "Female", "Male", "Male")

class(gender\_vector)

# Convert gender\_vector to a factor

factor\_gender\_vector <-factor(gender\_vector)

class(factor\_gender\_vector)

**Output:**

## [1] "character"

## [1] "factor"

It is important to transform a **string** into factor when we perform Machine Learning task.

A categorical variable can be divided into **nominal categorical variable** and **ordinal categorical variable**.

**Nominal categorical variable**

A categorical variable has several values but the order does not matter. For instance, male or female categorical variable do not have ordering.

# Create a color vector

color\_vector <- c('blue', 'red', 'green', 'white', 'black', 'yellow')

# Convert the vector to factor

factor\_color <- factor(color\_vector)

factor\_color

**Output:**

# **## [1] blue red green white black yellow**

**## Levels: black blue green red white yellow**

**From the factor\_color, we can't tell any order.**

**Ordinal categorical variable**

**Ordinal categorical variables do have a natural ordering. We can specify the order, from the lowest to the highest with order = TRUE and highest to lowest with order = FALSE.**

**We can use summary to count the values for each factor.**

**# Create Ordinal categorical vector**

**day\_vector <- c('evening', 'morning', 'afternoon', 'midday', 'midnight', 'evening')**

**# Convert `day\_vector` to a factor with ordered level**

**factor\_day <- factor(day\_vector, order = TRUE, levels =c('morning', 'midday', 'afternoon', 'evening', 'midnight'))**

**# Print the new variable**

**factor\_day**

**Output:**

**## [1] evening morning afternoon midday**

**midnight evening**

**## Levels: morning < midday < afternoon < evening < midnight**

**# Append the line to above code**

**# Count the number of occurence of each level**

**summary(factor\_day)**

**Output:**

**## morning midday afternoon evening midnight**

**## 1 1 1 2 1**

**R ordered the level from 'morning' to 'midnight' as specified in the levels parenthesis.**

**Continuous variables**

**Continuous class variables are the default value in R. They are stored as numeric or integer. We can see it from the dataset below. mtcars is a built-in dataset. It gathers information on different types of car. We can import it by using mtcars and check the class of the variable mpg, mile per gallon. It returns a numeric value, indicating a continuous variable.**

**dataset <- mtcars**

**class(dataset)**

**Output**

**## [1] "numeric"**