EDX Data Science Project HarvardX

Miami Kelvin

2019-06-17

Introduction to machine learning

The Data Science course on EDX by Harvard has introduced me to some core data science concepts, i am certainly not yet a seasoned data scientist but with the little knowledge i have received from the course i will attempt to demonstrate machine learning through this ML capstone project. This ML projects offer you a a slightly naive approach to training and testing a Machine Learning algorith to categorise some images.

Problem Statement

At my current job, we recently engaged a suplier to digitize all our client documents. one of the requirements was to categorise documents based on type and unit and at the same time embed metadata elements. Their approach was to have human workers look through each and every document and categorize it, the project was for 140,000 files and on average each file has 30 different document types varying in content and structure which can be 1 of multiple types(approximately 42 not counting emails and unstructured documents). I saw this as a total waste of time and resources, this is something that could be done more efficiently using a well trained algorithem. While I am miles away from this, this will be a great way to learn. Bingo! We have a great problem to implement a classification algorithm, this would have been a brillian project but taking a few steps back i realized i could not proceed with that given the timelines and the project requirements that data is available for peer review, so i went for the next BIG thing, Iris.

Dataset overview

There are very many datasets out in the wild however for a learning algorithm for image recognition and simple enough for a learner new to R and machine learning, the Iris data set is perfect.

```r
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00  
```

Getting and installing dependencies

Performing data cleaning and wrangling

Summarise the dataset we will use

Creating a model

Split the data to Training and Test sets

Training the Model

Evaluating the overal sensitivity and specificity of the model

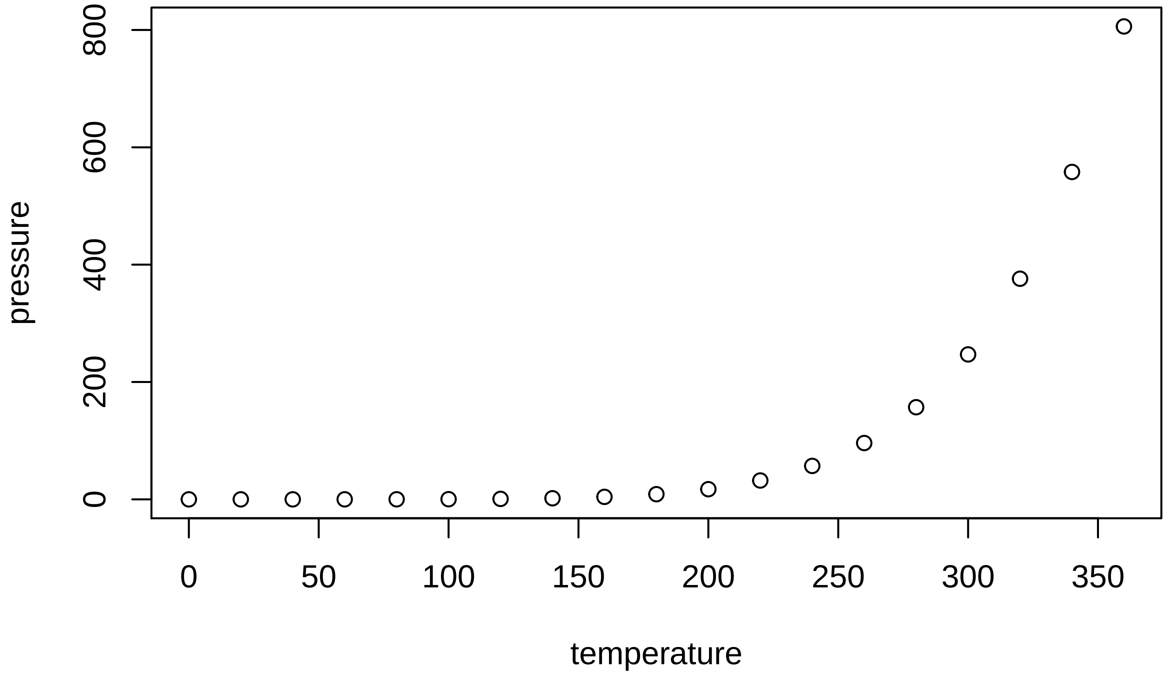
Testing themodel on the Test Set

Evaluating accuracy of the test set

Including Plots

You can also embed plots, for example:

```
plot(pressure)
```



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.