Introduction to Statistical Learning

Muhammad Zaid
Computer Vision and Machine Learning Engineer, FidelAI

FOCUS ON ME INSTEAD OF SLIDES!!

Not Much of Coding Today!

What is Statistical Learning

- Statistical learning refers to a vast set of tools for understanding data

- Sister Field of Machine Learning with different emphasis and approaches

- Heard of **Data Driven Companies ?** Statistical learning is the **key** which drives companies through data.

Some Use Cases of Statistical Learning

- Predict whether a patient, hospitalized due to a heart attack, will have a second heart attack. The prediction is to be based on demographic, diet and clinical measurements for that patient.

- Predict the price of a stock in 6 months from now, on the basis of company performance measures and economic data.

- Churn prediction for telecommunication companies based on user’s data

Categories of Learning

- Supervised Learning : Data consists of both input and output and the task is to learn a mapping from input to output. (**y = f(x))**
  - Many examples around!

- Unsupervised Learning : Data only consists of input variables and the task is to find any structure within the data. No output/response variable is present.
  - Cluster Analysis, Dimensionality Reduction etc (**f(x))**

Supervised Learning

- Regression : When output variable is continuous/real.

  Example : predicting stock price

- Classification : When output variable is discrete

  Example : Email Spam prediction

Some Important Terminologies

- Terminologies can sometimes turn out to be really confusing

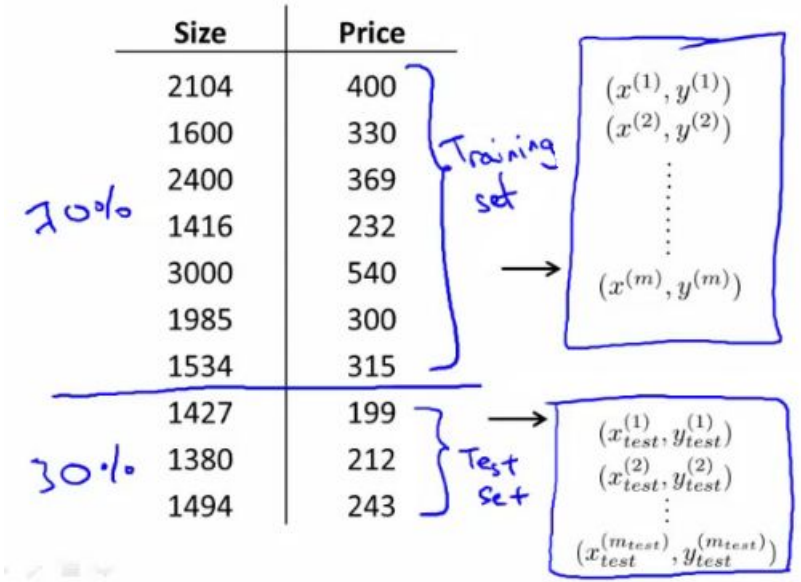
- Following terms have to be understood before getting started :
  - Number of Observations/Examples/Samples : n
  - Number of features/predictors/variables : p
  - Input Feature Matrix : X
  - Output variable vector : y

Training and Testing Set

- Data is usually divided into **Training** and **Testing** Sets

- Training set is the subset of data on which our model will be trained! It will be paired with a set of labels in case of Supervised learning.

- Testing set is remaining subset of the data! We check the accuracy of the model on the test (unseen) data.


Size
Price
2104
400
1600
330
2400
369
1416
232
3000
540
1985
300
1534
315
70%
Training set
$(x^{(1)}, y^{(1)})$
$(x^{(2)}, y^{(2)})$
$(x^{(m)}, y^{(m)})$
1427
199
1380
212
1494
243
30%
Test set
$(x_{test}^{(1)}, y_{test}^{(1)})$
$(x_{test}^{(2)}, y_{test}^{(2)})$
$(x_{test}^{(m_{test})}, y_{test}^{(m_{test})})$

Problem Statement

●Given the data set of heights and weights , train a model so that it could predict weight on the basis of heights.

● Supervised or Unsupervised ?

●Classification or Regression ?

Weight : Output variable

Height : Input variable

Understand this Equation

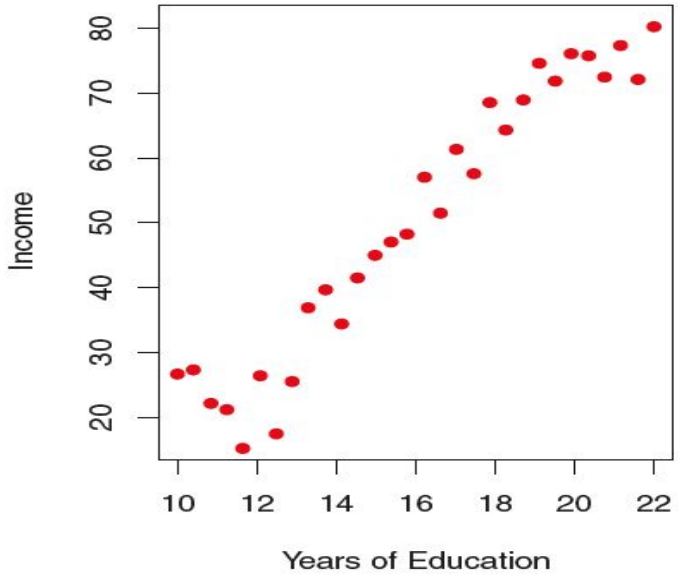
$$Y = f(X) + \text{error}$$

Linear Regression

Introduction

- A supervised algorithm used for regression.
- Linear Regression tries to fit a line to the data set
- Line implies a **linear** relationship between the output and input variable
- Sounds Confusing ? Let’s go through an example!

- Say we want to understand the relationship between **Years of Education** and **Income**. We have the following data set.



- Equation of Line : y=mx+c (m is the slope and c is y-intercept)

- For our case : Income = m (Years of Education) + c

- Linear regression tries to find out the values of Slope and intercept of a line such that it fit the data.

ACKNOWLEDGEMENTS

I would like to pay my humble gratitude to the following people , who are **True Mentors** for me:

- Dr Tahir Syed
- Sadaf Iqbal Behlim
- Muhammad Suleman
- Yameen Malik