Basics of Statistical Learning

Notations

- X --> Predictors, Independent variables
- Y --> Response, Dependent variables
  Input X ---> Y Output

$$X = (X_1, X_2 ... Xp) \to Y$$

  General form:

$$Y = f(x) + \varepsilon$$

- $f(x)$: Fixed unknown function
- ε : Error term
  We use Statistical Learning to estimate $f(x)$

Reasons to estimate $f(x)$

Prediction

$$\hat{Y} = \hat{f}(x)$$

- $\hat{f}$ estimate of f
- $\hat{Y}$ result of the prediction

There is two types of errors :

1. Reducible Error $\hat{f} \to f$
2. Irreducible Error Y is a Function of ε too

$$E(Y - \hat{Y})^2 = E(f(x) + \varepsilon - \hat{f}(x))^2$$
$$= (f(x) - \hat{f}(x)) + Var(\varepsilon)$$

- $f(x) - \hat{f}(x)$ is the reducible Error
- $Var(\varepsilon)$ is the variance of the irreducible Error

Note: In Prediction we treat $\hat{f}$ as a black-box, What matter more is that we get $\hat{Y}$ as close to Y as we can, Only the prediction matter

Inference

Here $\hat{f}$ cannot be treated as a black-box, We want to know it exact form because we may be interested in answering these questions :

- Which $predictors\ Xi$ are associated with the $response\ Y$, The ones with the most impact on the $response$?
- What is the relationship between the $response\ Y$ and each $predictor\ X_i$?
- Can the relation between Y and each $predictor X_i$ be written as a Linear equation or its more complex then that ?
    - Because a Linear relationship is more interptiable so we always seek that

How do we Estimate F?

- We always assume we observe n data points
- The observation x_{ij} represent the value of the jth predictor
- The response y_i represent the value of the ith observation
- So $\{(x_1, y_1)(x_2, y_2) \ldots (x_i, y_i)\}$ will be our Training Data with $x_i = (x_{i1}, x_{i2} ... x_{ip})$, with $variables$ or fields $1 \to p$
  We want to find $\hat{f}$ such that $Y \approx \hat{f}(x)$ for any observation (X, Y) -> unseen data

There is two types of approaches :

Parametric Methods

This method require two steps:

1. make assumption about the Functional Form or Shape of f
   most simple approach is that f is Linear Linear Regression

$$f(x) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots \beta_p X_p$$

   $f(x)$ is so Simple we only need to estimate $\beta_0 .. \beta_p$
2. Training or Fitting the model to estimate β_p such as:

$$Y \approx \beta_0 + \beta_1 X_1 + \ldots \beta_p X_p$$

Disadvantage:

- The model we choose usually wont match the unknown form of f if the model is too far true from of f our estimations will be poor
- We can get around that by choosing more flexible model that fits many forms of f
- More flexible model requires more parameters which can lead to Overfitting
- The more flexible model be the more complex it become

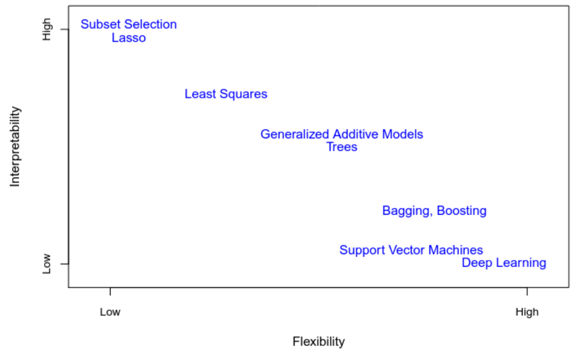
Non-Parametric Methods:

- Seek an estimate of f that get as close as possible to the Data points
- No assumptions needed
- f Fit wider range of possible shapes and forms
- Avoid the danger of not Fitting

Disadvantage:

- Since the non-parametric methods don't reduce the number of parameters
- A very big number of Observations data points is need (Way more than usual) to estimate f

Trade off : Prediction Accuracy vs Model interpretability



- Interpretability helps when inference cause its easy to understand the relationship between Y and X
- When Prediction is our goal more flexible models fits more
- **Note** : Sometimes a less flexible models gives better predictions than more flexible ones, it all depends on the type of relationship between Y and X

Supervised Versus Unsupervised Learning :

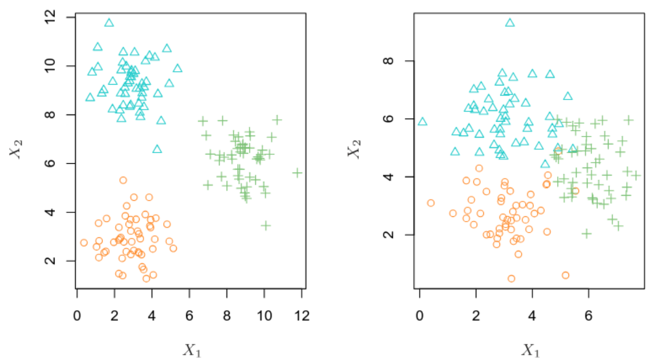
Supervised Learning:

- For each Observation x_i there is a Response, Measurement y_i
- We want to fit a model that Observation <--> Response
- With the aim to predict future unseen responses Prediction
  Or better understanding of the relationship between Observations n and the response y Inference

Unsupervised Learning :

Deals with more challenging situations

- We have the Observation x_i without the Response y
- Seek to understand the relationships between the variables or the Observations



Regression Vs Classification Problems :

It all depend on the type of $\text{Variables } 1 \rightarrow p$ either

- Quantitative
  Regression Problems
- Qualitative (Categorical)
  Classification Problems
- So Selecting the Statistical Learning Method depends on --> The Response variable type (quan/quali)
- The predictors X variable type are way less important on the decision