em lyon business school

Session 1 – Main Concepts

A high level overview of the main concepts used in Machine & Statistical Learning

Reference :

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An Introduction to Statistical Learning : With Applications in Python* (1st ed. 2023 edition). Springer.

Statistical Learning versus Machine Learning

em lyon business school

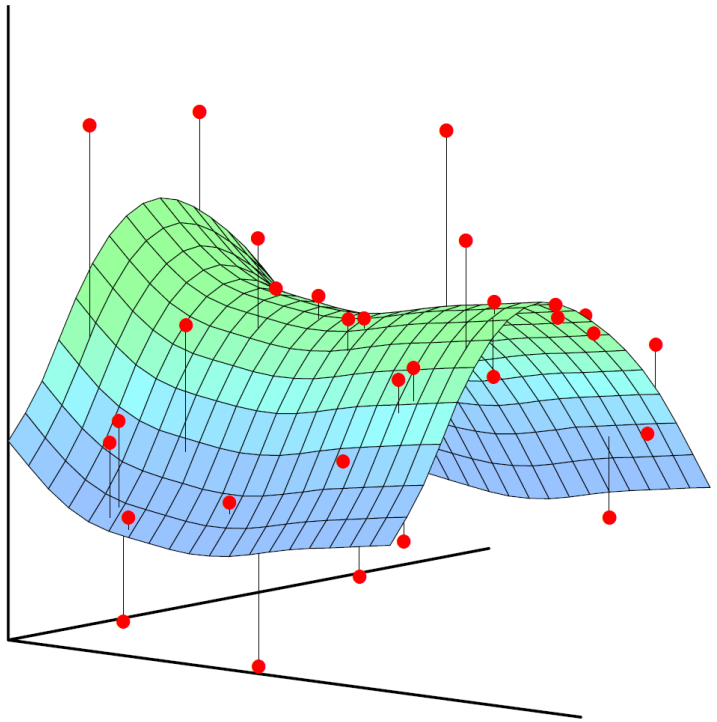
- **Machine learning arose as a subfield of Artificial Intelligence.**
  - Leaning more towards computer science.
  - An algorithmic approach.

- **Statistical learning arose as a subfield of Statistics.**
  - Leaning more towards mathematics and statistics.
  - A modeling approach.

- **There is much overlap : both fields focus on supervised and unsupervised problems:**
  - Machine learning has a greater emphasis on large scale applications and prediction accuracy.
  - Statistical learning emphasizes models and their interpretability, and precision and uncertainty.

- **But the distinction has become more and more blurred, and there is a great deal of "cross-fertilization".**

- **Machine learning has the upper hand in Marketing!**



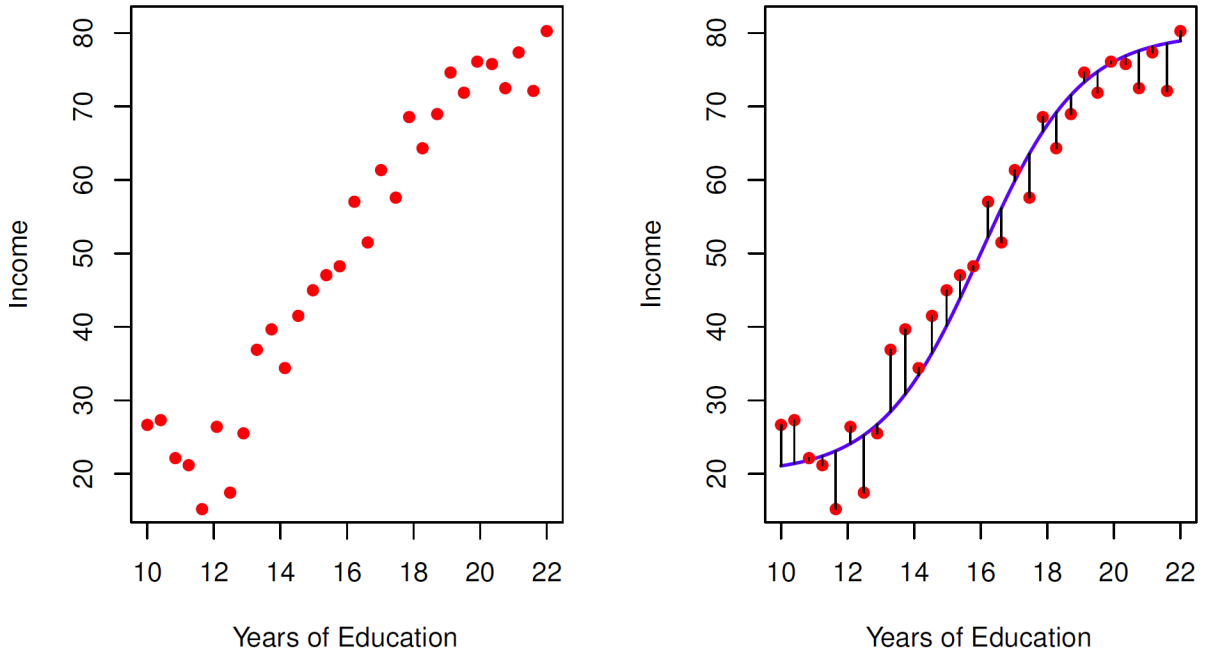
What Is Statistical Learning?

- **Starting point**
  - Several input variables $\boldsymbol{X} = [X_1, X_2, \cdots, X_p]$
    - Inputs :: Predictors :: Features :: Independent variables
    - Each predictor X_j has n *data points*
  - One output variable Y *(with n data points also)*
    - Output :: Outcome :: Response :: Dependent variable
  - $\boldsymbol{X}$ and Y are given by the (observed) data
  - Some relationship exists between $\boldsymbol{X}$ and Y

$$Y = f(\boldsymbol{X}) + \epsilon$$

    - ϵ: a random error term with $E(\epsilon) = 0$
    - f: systematic information $\boldsymbol{X}$ provides about Y

- **Goal**
  - Estimating f from the data



In essence, statistical learning refers to a set of approaches for estimating f (James et al, 2023, p.17)

Why estimate f ?

- **Goal : Prediction**
  - Is this newly admitted patient likely to have a prolonged stay ?
  - What is the mostly likely rate of turnover in our organization ?
  - Give your own example…
- **Find $\hat{f}$ – an estimate of f – where**
  - $\hat{Y} = \hat{f}(X)$ represents the vector of predicted values
  - The overall (aggregated) prediction error between Y and $\hat{Y}$ is minimized
- **Example : minimize $E\left[(Y - \hat{Y})^2\right]$**
  - Assuming f and X fixed :

$$E\left[(Y - \hat{Y})^2\right] = \underbrace{\left[f(X) - \hat{f}(X)\right]^2}_{\text{Reducible error}} + \underbrace{Var(\varepsilon)}_{\text{Irreducible error}}$$

- **Goal : Inference / Explainability**
  - What factors (medical predictors) are most predictive of a prolonged stay ?
  - What factors (organizational predictors) are most predictive of the turnover rate ?
  - Give your own example…
- **Relationship between the outcome and the predictors**
  - Type or nature of the relationship
  - Strength of the relationship

Focus on **prediction performance only** raises the issue of the ***Black Box Problem***.
Focus on the **explainability alone** raises the issue of ***Prediction Reliability***.

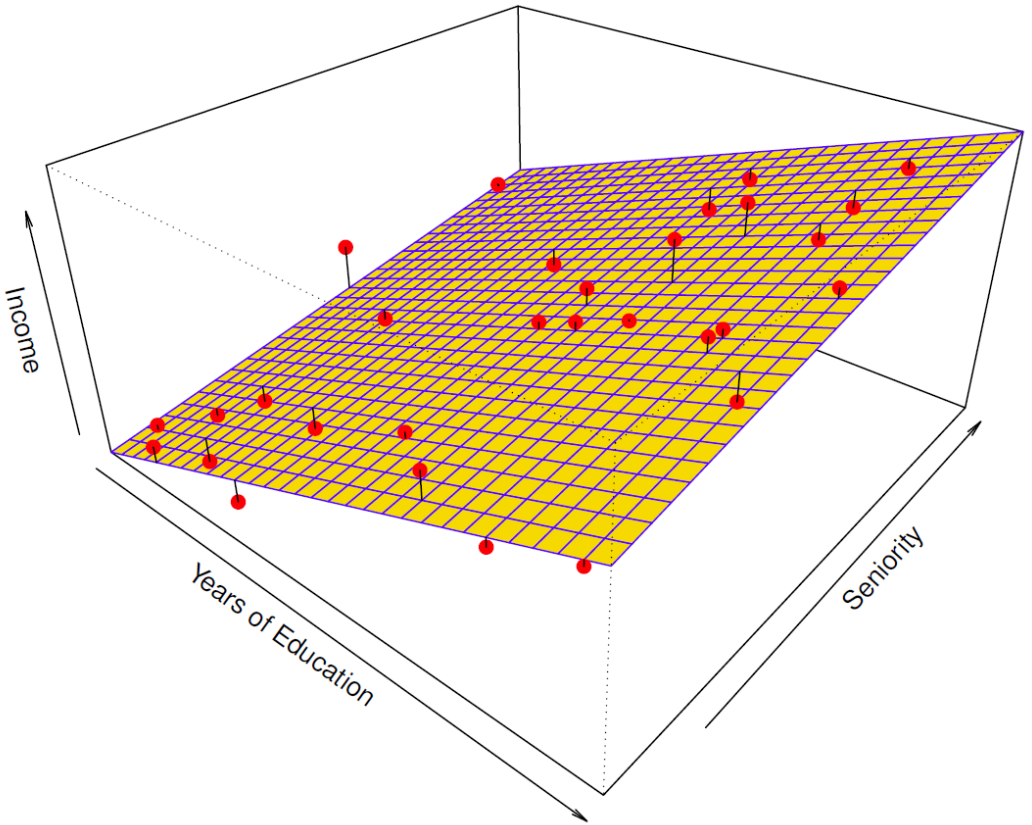
How do we estimate f ? Part 1

em lyon business school

- **Parametric methods**
  - We make explicit assumptions on the functional relationship between the outcome and the predictors.
  - The problem of estimating f is reduced down to estimating a set of parameters.
  - Rely on (statistical) modeling
  - Example :
    - $Y = f(\boldsymbol{X}) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$
    - The problem of estimating f is reduced to estimating $[\beta_0, \beta_1, \ldots, \beta_p]$

- **Upsides**
  - Simplifies the estimation to a reduced number of parameters

- **Downsides**
  - The model $\hat{f}$ will not match the true f
    - More flexible models may lead to overfitting (the model picks as much noise as it picks information)

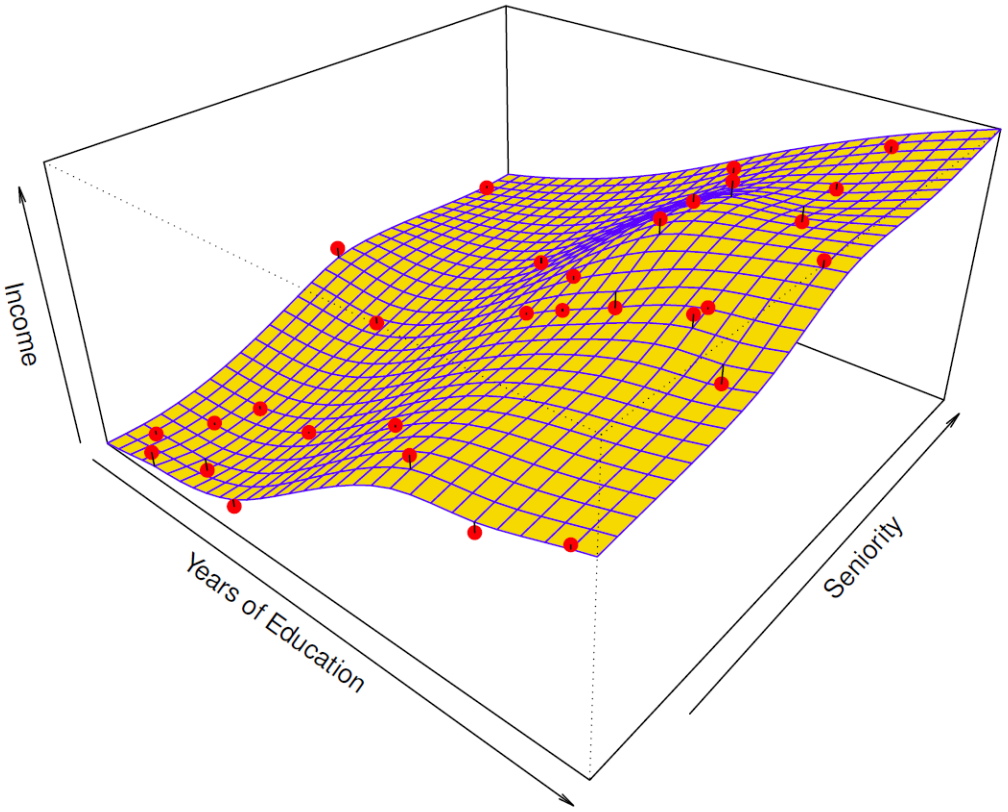


How do we estimate f ? Part 2

- **Non-Parametric methods**
  - No explicit assumptions on the functional relationship between the outcome and the predictors.
  - Seek an estimate of f as close to the datapoints as possible.
  - Rely on an algorithmic approach
  - Example :
    - $Y = kNN(X) : the\ k - nearest\ neighbors$
    - kNN estimates each datapoint based on the values of the k-nearest neighbors

- **Upsides**
  - Have the potential to accurately fit a wider range of possible shapes for f

- **Downsides**
  - Large number of observations required to obtain an accurate estimate for f
  - May lead easily to *overfitting*



Supervised Learning

em lyon business school

- **Supervised Learning**
  - Given some observed data $(\boldsymbol{X}, Y)$
    - Of n *data pair points* $(\boldsymbol{X}_{i.}, Y_i)$
    - And p *variables* $\boldsymbol{X} = [X_1, X_2, \dots, X_p] = [x_{i,j}]$
    - $\boldsymbol{X}$: input (features) is associated to Y : output (response)
    - We can learn Y from $\boldsymbol{X}$ through the relationships in the n *data points* $(X_{i.}, Y_i)$
  - Given a new set $\boldsymbol{\chi} = (X_{\mu.})$ of m *data points*
    - We can predict the corresponding (Y_μ)
    - Based on the information learned in the n *data points* $(\boldsymbol{X}_{i.}, Y_i)$

$$\begin{bmatrix} X_1 & X_2 & \dots X_j \dots & X_p & Y \\ x_{1,1} & x_{1,2} & \dots x_{1,j} \dots & x_{1,p} & y_1 \\ \vdots & & \ddots & & \vdots \\ x_{i,1} & & \dots x_{i,j} \dots & & y_i \\ & & \dots & & \\ x_{n,1} & x_{n,2} & \dots x_{n,j} \dots & x_{n,p} & y_n \end{bmatrix}$$

If the outcome (or output) Y is quantitative, the supervised learning is called a ***regression***
If it is categorical, it is called a ***classification***

Give your own example of regression and classification

Unsupervised Learning

em lyon business school

- **Unsupervised Learning**
  - Given some observed data (X)
    - Of n *data points* $(X_{i.})$
    - As p *variables* $\boldsymbol{X} = [X_1, X_2, \dots, X_p] = [x_{i,j}]$
    - There is no given output in the data
  - We look for patterns of similarity
    - Either between the observations (rows)
      - Usually by comparing the « distances » between observations
    - Or between the columns (variables)
      - Usually by looking at « angles » between columns
  - Then aggregating those closest in distance (angles)
    - Finding sensible interpretations for each group of observations or of variables

$$\begin{bmatrix} X_1 & X_2 & \dots X_j \dots & X_p \\ x_{1,1} & x_{1,2} & \dots x_{1,j} \dots & x_{1,p} \\ \vdots & & \ddots & \\ x_{i,1} & x_{i,2} & \dots x_{i,j} \dots & \\ & & \dots & \\ x_{n,1} & x_{n,2} & \dots x_{n,j} \dots & x_{n,p} \end{bmatrix}$$

Assessing model performance

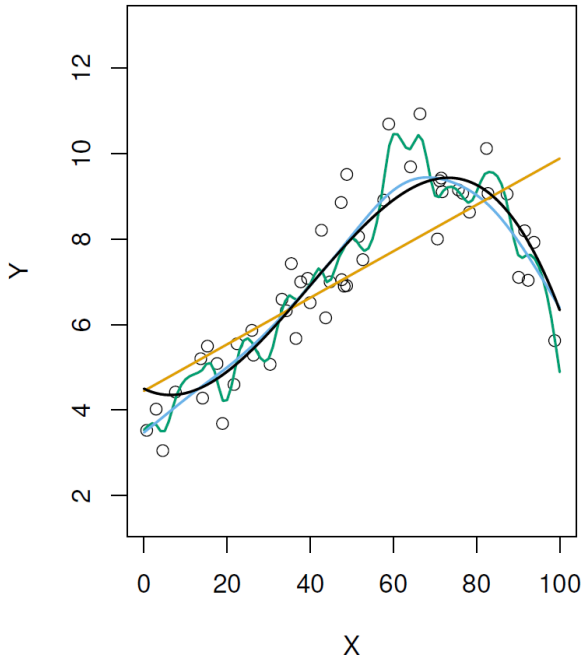
- **There is no model that is the best under all circumstances**
  - Performance depends on the model and the data
  - Performance should be compared :
    - Between models
    - On the same dataset
    - The dataset used to train each model and to estimate their performance should not be the same

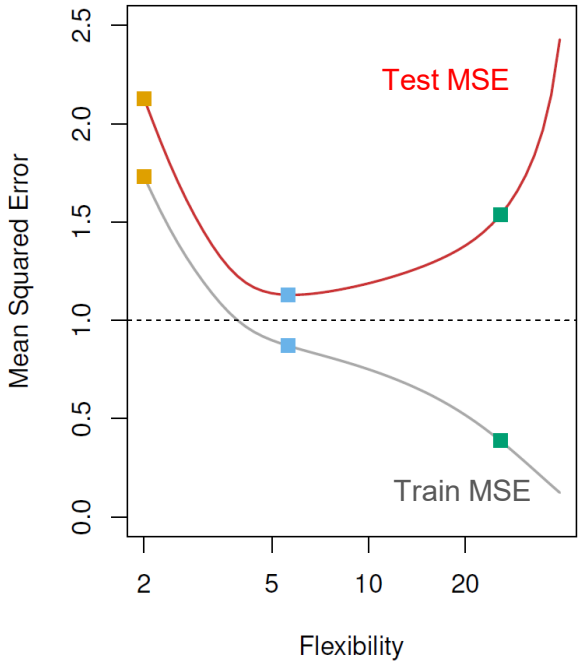
- **Simplest way to estimate model performance : aggregated error of prediction**
  - Measuring performance of a regression model
    - Aggregated distance between actual and predicted

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2$$

  - Measuring classification performance
    - Aggregated counts of correct classification

$$Accuracy = \frac{1}{n}\sum_{i=1}^{n} I(y_i = \hat{y}_i)$$





When the performance of a model (here measured in MSE) is **much higher on the test data** than on the training data, we are ***overfitting*** the data

Bias-Variance Trade-Off

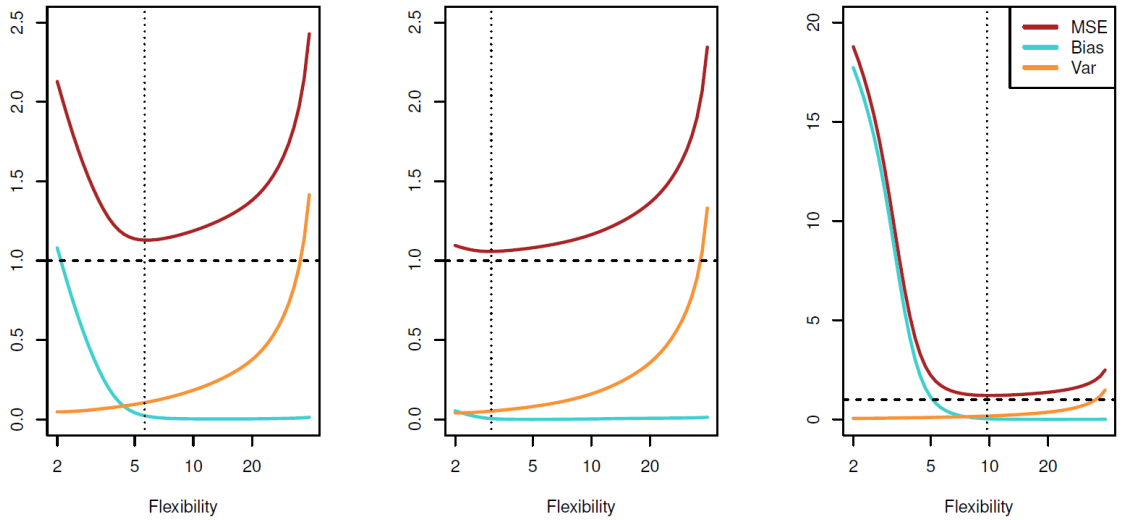
- **Bias**
  - Error introduced by approximating a real-life problem
    $$bias(\hat{Y}) = E(\hat{Y}) - Y$$
  - More flexible model tends to result in less bias (less prediction errors)

- **Variance**
  - Amount by which $\hat{f}$ would change if we estimated it from another dataset
  - More flexible model tends to result in higher variance (less reliability)

- **Expected SE**

$$E[(Y_0 - f(X_0))^2] = Var\left(\hat{f}(X_0)\right) + \left[bias\left(f(X_0)\right)\right]^2 + Var(\varepsilon)$$

  - The goal is to find a model that minimizes the bias and the variance simultaneously



The minimum value for the red curve represents the ***Bias-Variance Trade Off*** for respectively : **medium** flexibility, **low** flexibility and **high** flexibility

Class Exercise

- **Hands on…**
  - Apply the principles discovered in this session on the <Flourishing> dataset