Introduction to Machine Learning

Franck JAOTOMBO

Introduction to Data Analysis

- In Data Science, before you do anything, you first start by exploring your dataset.
- The basic steps in data exploration are following:
  1. Univariate Data Analysis
  2. Bivariate Data Analysis
  3. (Modeling)

Step 1 - Univariate Data Analysis

1. If the variables are categorical.
   1. Generate the summary table for each variable.
   2. Plot their Pie Chart
   3. Plot their Bar Chart

2. If the variables are quantitative.
   1. Generate the frequency table for each variable
   2. Plot their histogram
   3. Plot their boxplot

Step 2 – Bivariate Data Analysis

1. If the variables are both categorical.
   1. Generate the contingency table
   2. Check the significance of their relationship with the chi-square test & provide Cramer's V
   3. Plot their side-by-side bar charts
   4. Plot their stacked bar charts

2. If the variables are both quantitative.
   1. Compute the correlation (table)
   2. Check the significance of their relationship with the correlation test & provide the r value
   3. Plot their scatter plot (matrix)

3. If the variables are mixed categorical & quantitative.
   1. Compute the anova table
   2. Check the significance of the difference in values between groups
   3. Plot the grouped boxplots

Modeling : supervised

- Which variables should be selected as the ***outcome*** (target, response, dependent variable) or variable to be explained from the others (***predictors***, features, independent variables)?
- The answer should be justified with theoretical, managerial or statistical arguments

- The goal is to ***find a function*** that captures in the best possible way the relationship between the outcome and the predictors
- This process is called "modeling" and statistical learning is one way of addressing it

- One goal of modeling is thus to explain the variability (or variance) in the outcome from the predictors.
- This approach to modeling is associated with "**supervised learning**" in statistical learning.

Modeling : unsupervised

- The variance may also be explained by the existence of subgroups of individual instances or subgroup heterogeneity.
- The process to account for these subgroups is associated with “**unsupervised learning**” in statistical learning.

When the number of variables is too numerous, the relationship between the outcome and the predictors can become too complex.

- To reduce this complexity and to simplify interpretability, dimension reduction is recommended.
- The set of tools to reduce dimensions is also associated with “**unsupervised learning**” in statistical learning.

Hypothesis Testing : A review

Correlation Test

$H_0\colon\ \rho = 0$

$H_1\colon\ \rho \neq 0$

$$Statistic\ of\ the\ test\ \ - Student(n-2) : t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Chi-square Test

$H_0\colon\ \chi^2 = 0$

$H_1\colon \chi^2 > 0$

$$Statistic\ of\ the\ test\ \ - Chi\ square[(r-1)(c-1)]\colon X^2 = \sum_{i,j} \frac{\left[n_{ij} - \frac{r_i c_j}{n}\right]^2}{\frac{r_i c_j}{n}}$$

Anova Test

$H_0\colon\ \mu_1 = \mu_2 = \cdots = \mu_c$

$H_1\colon All\ the\ \mu_j\ are\ not\ equal$

$$Statistic\ of\ the\ test\ \ - Fisher\ (c-1, n-c) : F_{stat} = \frac{MSB}{MSW}$$

Linear Correlation Test (Pearson)

- We want to test if there is a significant association between two continuous variables

- The covariance between two variables X and Y indicates if there is an association between the variation of the two variables around their respective means

$$cov(X,Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- Correlation is a standardized measure of the covariance

$$r = \frac{cov(X,Y)}{s_X s_Y} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}}$$

s_X and s_Y are the respective standard deviations of X and Y

Chi square

- **If the row and column variables are independent:** $E(n_{ij}) = \frac{r_i c_j}{n}$

$$X^2 = \sum_{i,j} \frac{\left[n_{ij} - \frac{r_i c_j}{n}\right]^2}{\frac{r_i c_j}{n}}$$

	X_1	X_j	X_c	$Total$
Y_1	n_{11}		n_{1c}	r_1
Y_i		n_{ij}		r_i
Y_l	n_{r1}		n_{rc}	r_l
$Total$	c_1	c_j	c_c	n

- r_i and c_j indicate respectively the total (marginal) frequency of row i and column j
- X^2 follows a Chi Square distribution with a degree of freedom = $(r-1) * (c-1)$
  - where r **= number of modalities on rows** and c**= number of modalities on columns**
  - We need only to compare X^2 with the threshold values of the Chi Square (χ^2) distribution

- **Effect size :** $\phi = \sqrt{\frac{X^2}{n}}$ **and** $V_{Cramer} = \sqrt{\frac{X^2}{n.\min[(r-1),(c-1)]}}$

Analysis of Variance

$$SST = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(X_i^j - \bar{\bar{X}}\right)^2$$

$$MST = \frac{SST}{n-1}$$

$$SSB = \sum_{j=1}^{c} n_j\left(\bar{X}_j - \bar{\bar{X}}\right)^2$$

$$MSB = \frac{SSB}{c-1}$$

$$SSW = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\left(X_i^j - \bar{X}_j\right)^2$$

$$MSW = \frac{SSW}{n-c}$$

$$Grand\ mean: \ \bar{\bar{X}} = \sum_{j=1}^{c}\sum_{i=1}^{n_j}\frac{X_i^j}{n}$$

$$Group\ mean: \bar{X}_j = \sum_{i=1}^{n_j}\frac{X_i^j}{n_j}$$

Exercise

- **Explore the <Flourishing> dataset**
  - Read the instructions in the <Flourishing_Case.docx> document

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

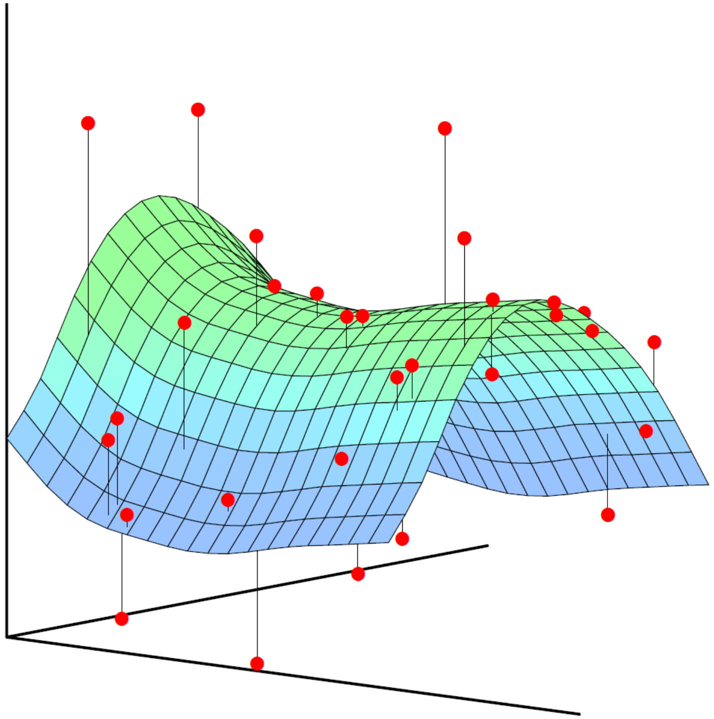
- **Machine learning arose as a subfield of Artificial Intelligence.**
  - Leaning more towards computer science.
  - An algorithmic approach.

- **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!**



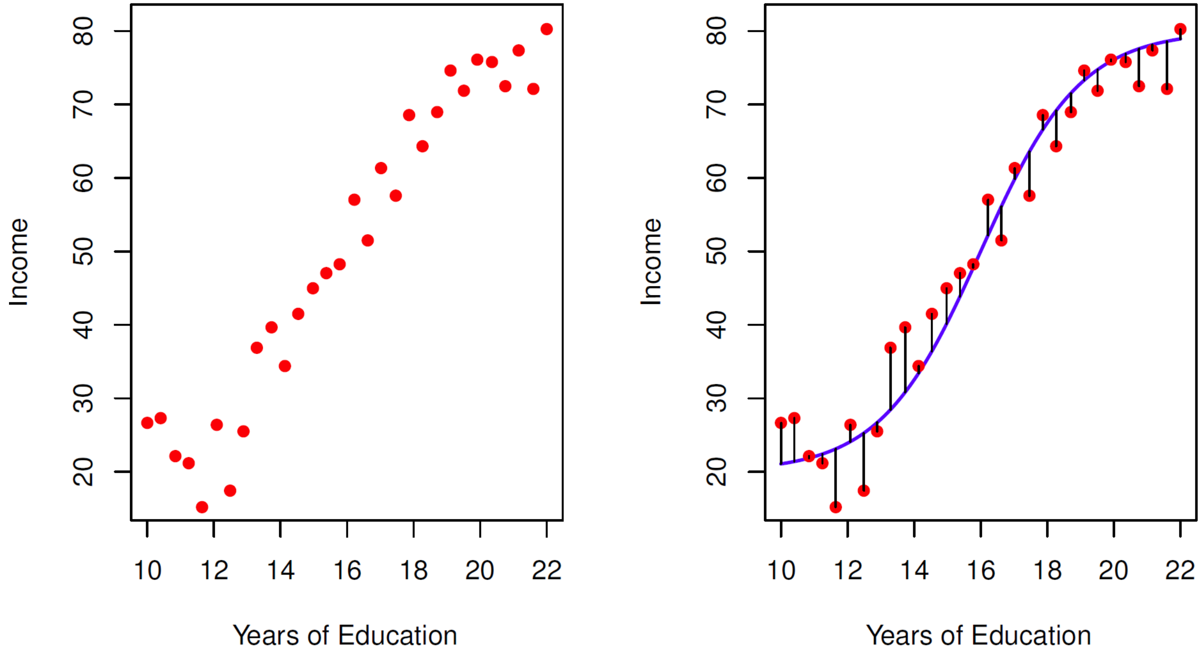
What Is Statistical Learning?

- **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 X and Y

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

- ϵ: a random error term
- 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 $E\left(Y - \hat{Y}\right)^2$**
  - Assuming f and X fixed :

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

Reducible error 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

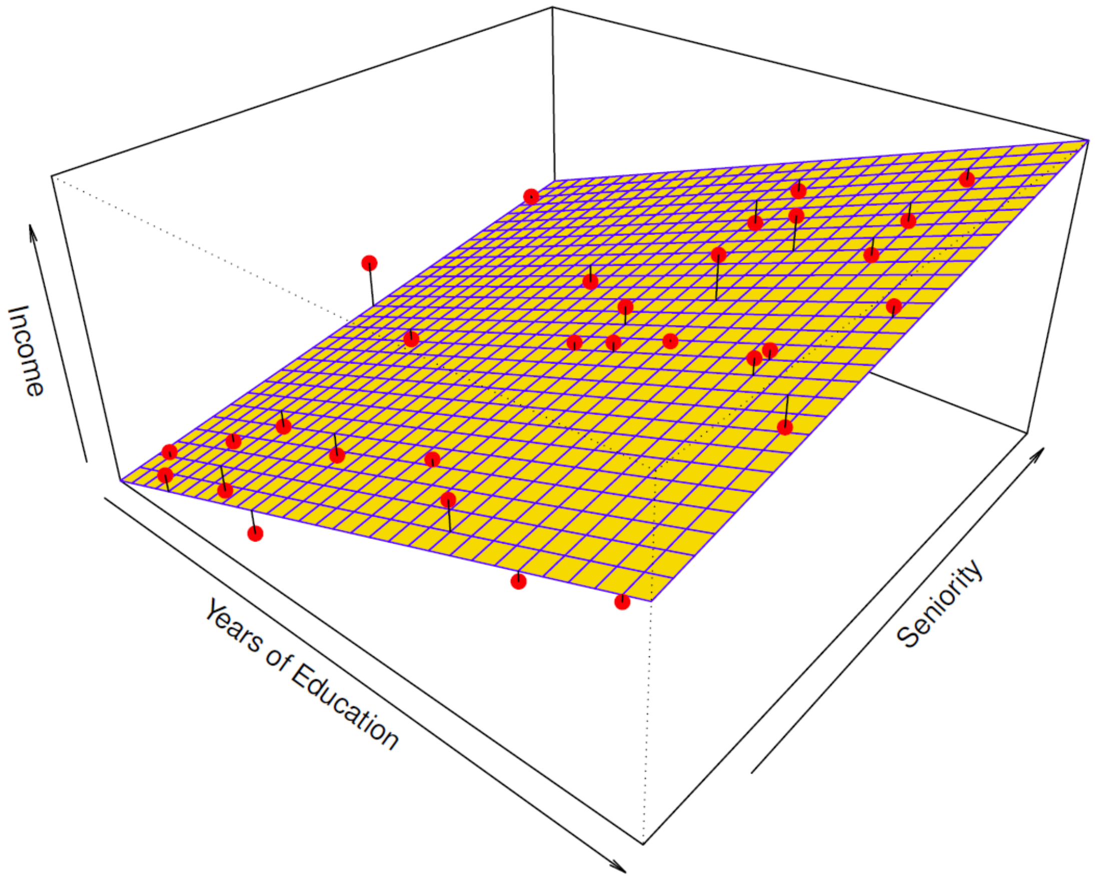
Focus on **prediction performance only** raises the issue of the ***Black Box Problem***.
Focus on the **explainability alone** raises the issue of ***Prediction Reliability***.

How do we estimate f ? Part 1

- **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 :
    - $Y = f(X) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p$
    - The problem of estimating f is reduced to estimating [β_0, β_1,…, β_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

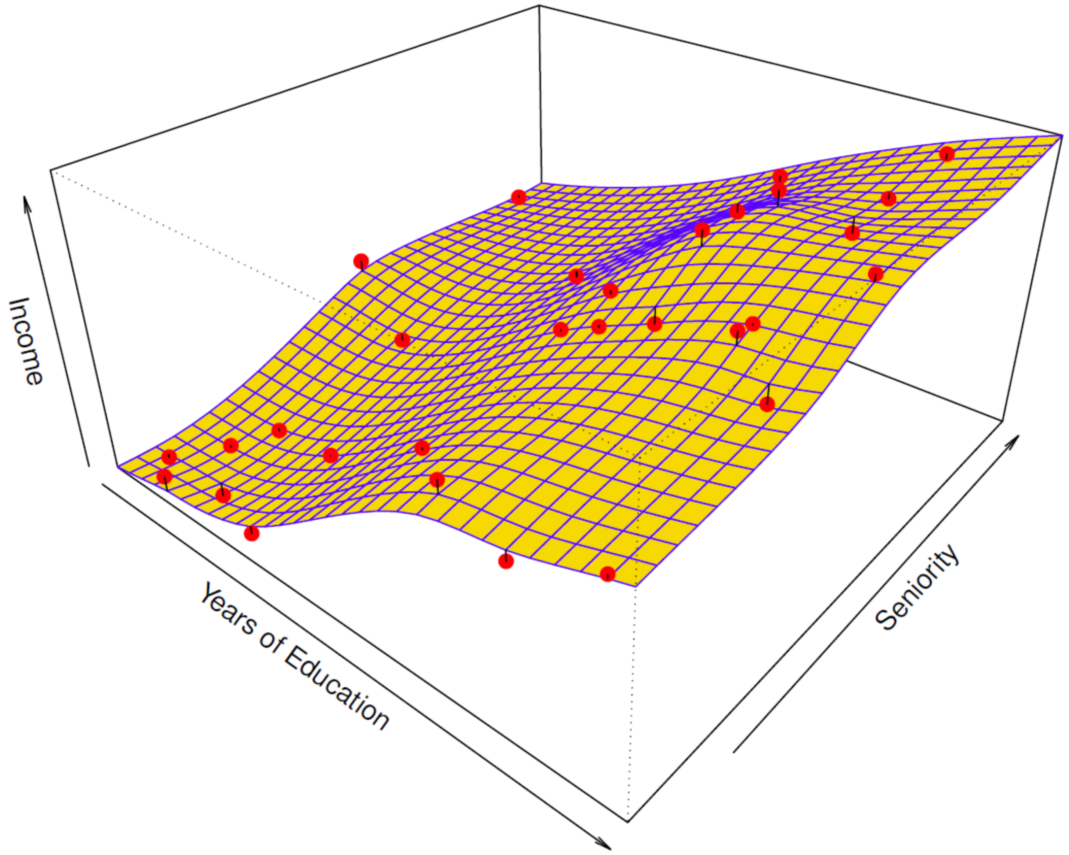


How do we estimate f ? Part 2

- **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

- **Supervised Learning**
  - Given some observed data (X, Y)
    - Of $n\ data\ pair\ points\ (x_i, y_i)$
    - And $p\ variables\ \boldsymbol{X} = [X_1, X_2, \ldots, X_p] = \left[x_{i,j}\right]$
    - X : input (features) is associated to Y : output (response)
    - We can learn Y from $\boldsymbol{X}$ from 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 $n\ data\ points\ (x_i, y_i)$

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

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 you own example of regression and classification

Unsupervised Learning

- **Unsupervised Learning**
  - Given some observed data (X)
    - Of $n\ data\ points\ (x_i)$
    - As $p\ variables\ \boldsymbol{X} = [X_1, X_2, \dots, X_p] = \left[x_{i,j}\right]$
    - 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 « distances » between columns
  - Then aggregating those closest in distance
    - Finding sensible interpretations for each group of observations or of variables

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

Assessing model performance

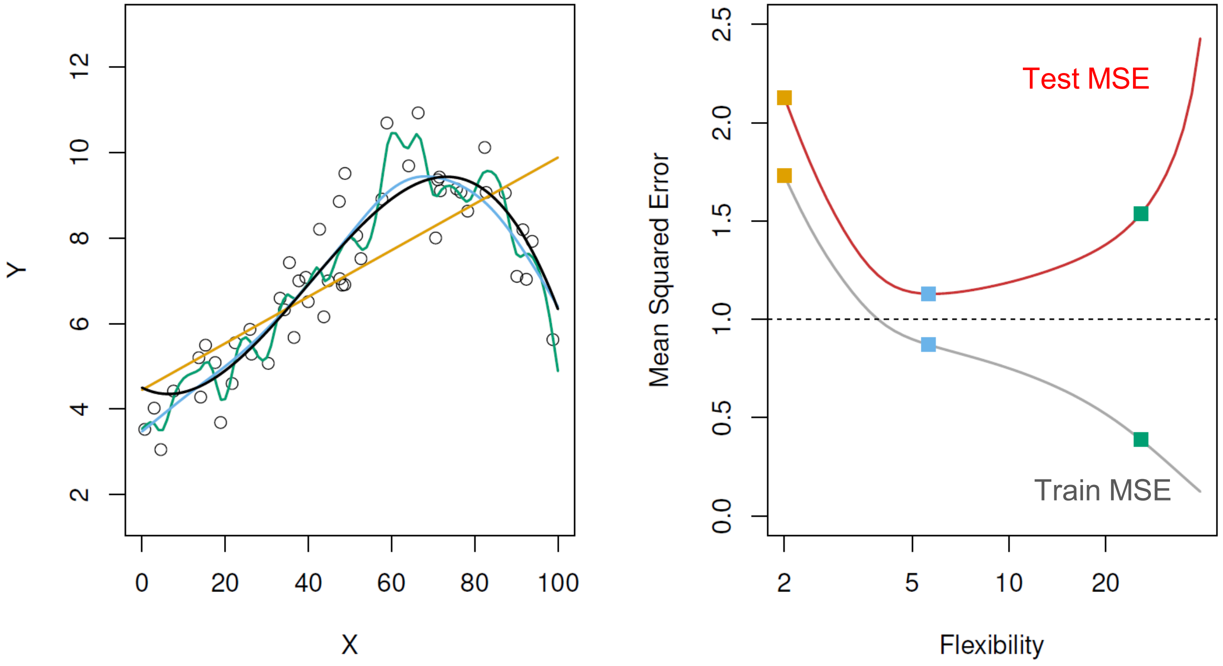
- **There is no model that is the best under all circumstances**
  - Performance depends on the model and the data
  - Performance should be compared :
    - Between models
    - On the same dataset
    - The dataset used to train each model and to estimate their performance should not be the same

- **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

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



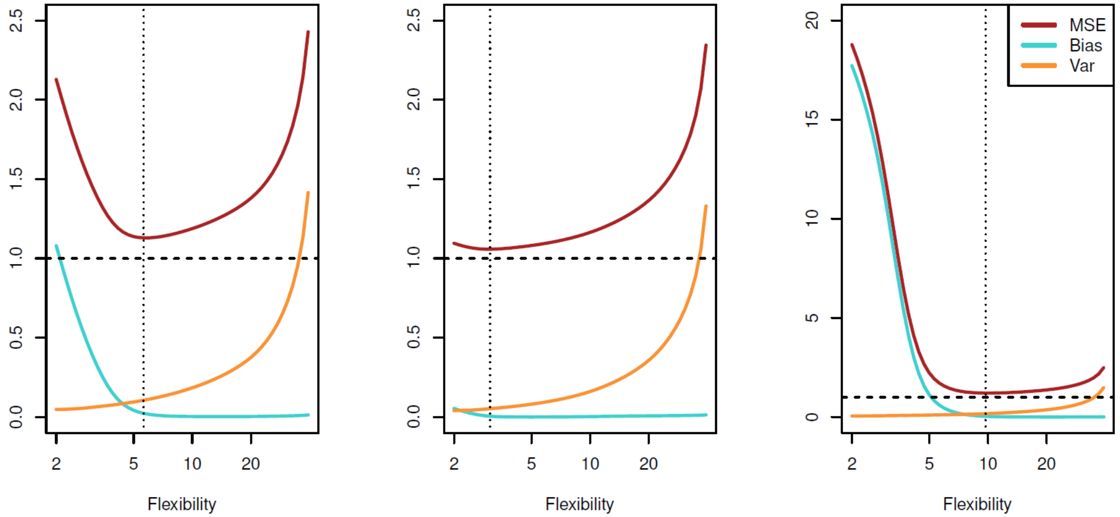
When the performance of a model (here measured in MSE) is **much higher on the test data** than on the training data, we are ***overfitting*** the data

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 MSE**

$$E(y_0 - f(x_0))^2 = Var\left(\hat{f}(x_0)\right) + \left[bias(f(x_0))\right]^2 + Var(\varepsilon)$$

- 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

Session 2 – The Machine Learning Workflow – Part 1

Explore all the main steps involved in a Machine & Statistical Learning workflow

Preparing a « rich » dataset

- **Importing with pandas**
  - Save the `mini_victoria.txt` file
  - Check the data in a text editor such as Notepad++ or Visual Studio Code
  - Import it using pandas
  - Print a comprehensive summary

- **The price variables are not recognized as quantitative**
  - Make the necessary pre-processing to read them as such
    - Create a function that removes the « $ » symbol for the USD currencies and replaces all others by missing values
    - Apply it on each of the price columns
  - Check again

- **Are there some missing data ?**
  - If so, make an assessment of the missing data in the whole dataset
  - Suggest some ways of handling them

Pandas Exercise

1. Write the lines of code to provide the name of the cheapest product
2. Write the lines of code to count the number of products with available size equal to '38A’
3. Write the lines of code to list and count the type and color of the most expensive products containing 'sport bra'

Recoding Categorical Data

- **Machine Learning models cannot process non numerical data**
  - Whether the data is made of sounds, of images, or text, it must be numerically encoded before ML processing
  - Categorical data must also be numerically encoded
    - Either through dummy coding [or onehot encoding]
    - Or through label encoding
    - As long as the variable does not have too many categories, onehot encoding is the preferred option
    - When label encoding is necessary, then some ML models may not be appropriate

- **Onehot encoding (dummy coding)**
  - Given a categorical data with k categories
    - **Onehot encoding** : create $k\ binary\ variables$ – one for each category – where the $value\ = 1$ for that category and $0\ otherwise$
    - **Dummy coding (stastistics approach)** : same as onehot encoding but $we\ select\ a\ reference\ category.$

      We do not create a binary variable for the reference category. We create $k-1\ binary\ variables$ for the others
  - The statistical approach is not easy to interpret in ML. It makes sense only in linear models

- **Label encoding**
  - When a variable has too many categories, onehot encoding becomes unpractical
  - Then we use label encoding
    - Each category is replaced by an integer
    - This generates an information that was not initially present in data : the order between the numbers
    - Some ML models may then become unsuitable for modeling

Import the `Credit.csv` dataset

- Recode all the categorical variables using sklearn onehotencoder and pandas get_dummies
- Compare your results

Import the `mini_victoria.txt` dataset

- Which categorical variables should be onehot encoded ?
- Which categorical variables should be label encoded ?
- Recode all categorical data appropriately

Handling Missing Data

- **Basic methods**
  - Replacing missing values with a central tendency estimate
    - For quantitative data :
      - the mean or the median (be careful…!)
    - For nominal data:
      - the mode
    - For ordinal data
      - the mode or the median

- **Intermediate methods**
  - Replacing missing values through interpolation
    - Only for quantitative data
      - usually, linear interpolation is good enough
    - Categorical data should be onehot encoded

- **Advanced Methods**
  - Replacing missing values through modeling
    - Iterative & Multiple imputation
    - Nearest neighbors
    - Machine Learning imputation

Import the `Credit.dat` dataset

- It should contain some missing values

Import the `Credit.csv` dataset

- It is the original dataset before introducing missing values

Explore the different ways of imputing missing values

- Which ones are the best in retrieving the value of the original dataset ?
- Use an appropriate performance metrics to compare your results

Selecting Features & Targets

- **Targets**
  - Should be selected based on the type of data science problem to be addressed…
    - If the target is quantitative we have a regression problem
    - If the target is categorical we have a classification problem
    - If there is no relevant target we have an unsupervised learning problem
- **Features**
  - In real-life data science problems, we are usually given datasets where the features are not readily exploitable.
    - We may have too many features
    - We may have too few features
    - Both can be a cause of serious difficulty
  - When they are exploitable, not all of them are relevant to the study.
    - And most of the time, they need to be prepared (pre-processed) before they can be suitable for Machine Learning

In the `Default.csv` data, we want to predict whether a client is going to default on his(her) credit

- Assess the target and the type of problem

In the `Credit.csv` data, we want to predict the balance of a client’s account

- Assess the target and the type of problem

We want also to understand the different profiles of client based on all available information in the `Credit.csv` data

- Assess the target and the type of problem

We want to predict brand name in the `mini_victoria.txt` data

- Discuss which features would be relevant for the analysis and what type of pre-processing would be required

Complementary class exercise & homework

- Complete the exercises in the following notebook
  - < session_2_ML_workflow_stu.ipynb>