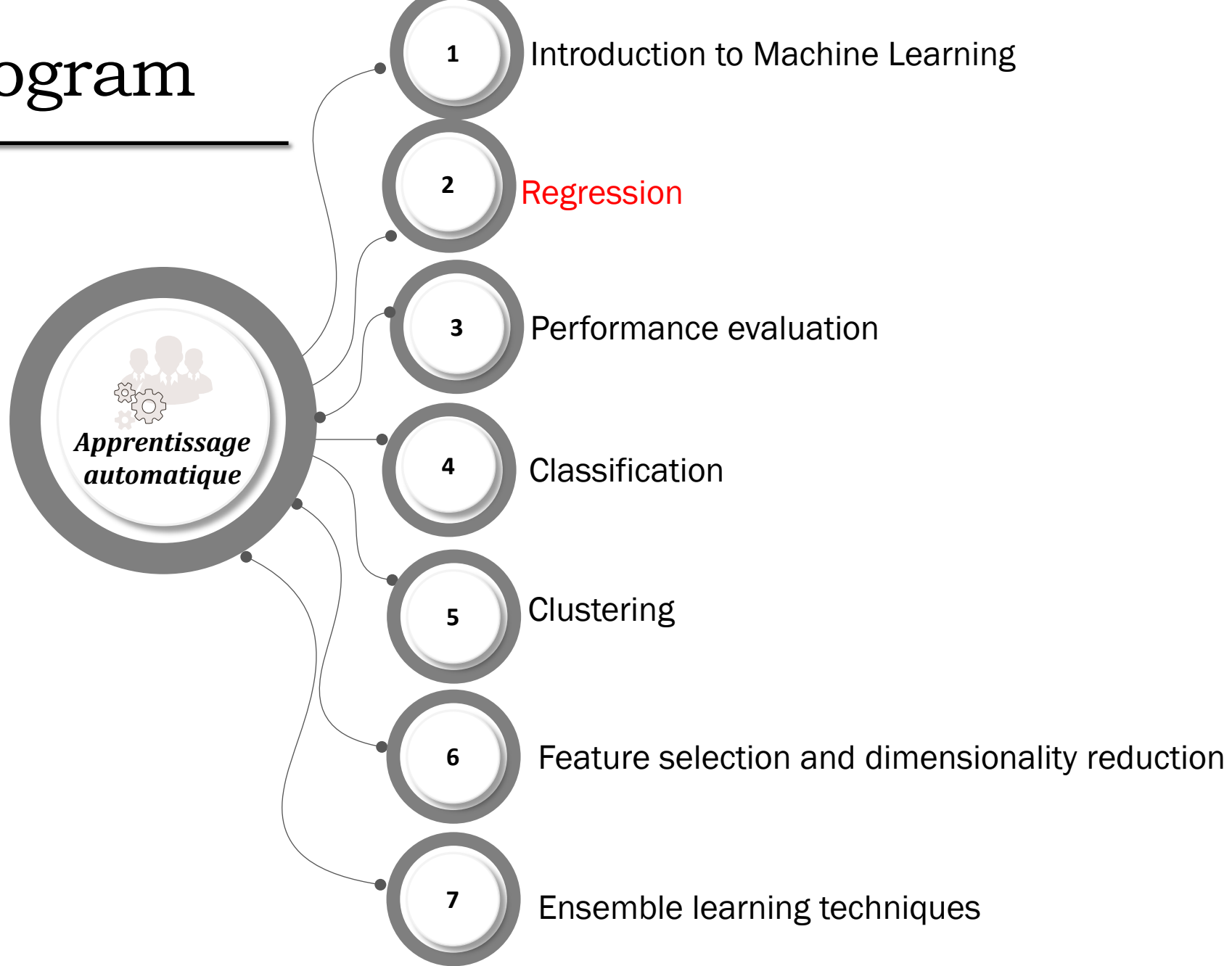


Machine Learning

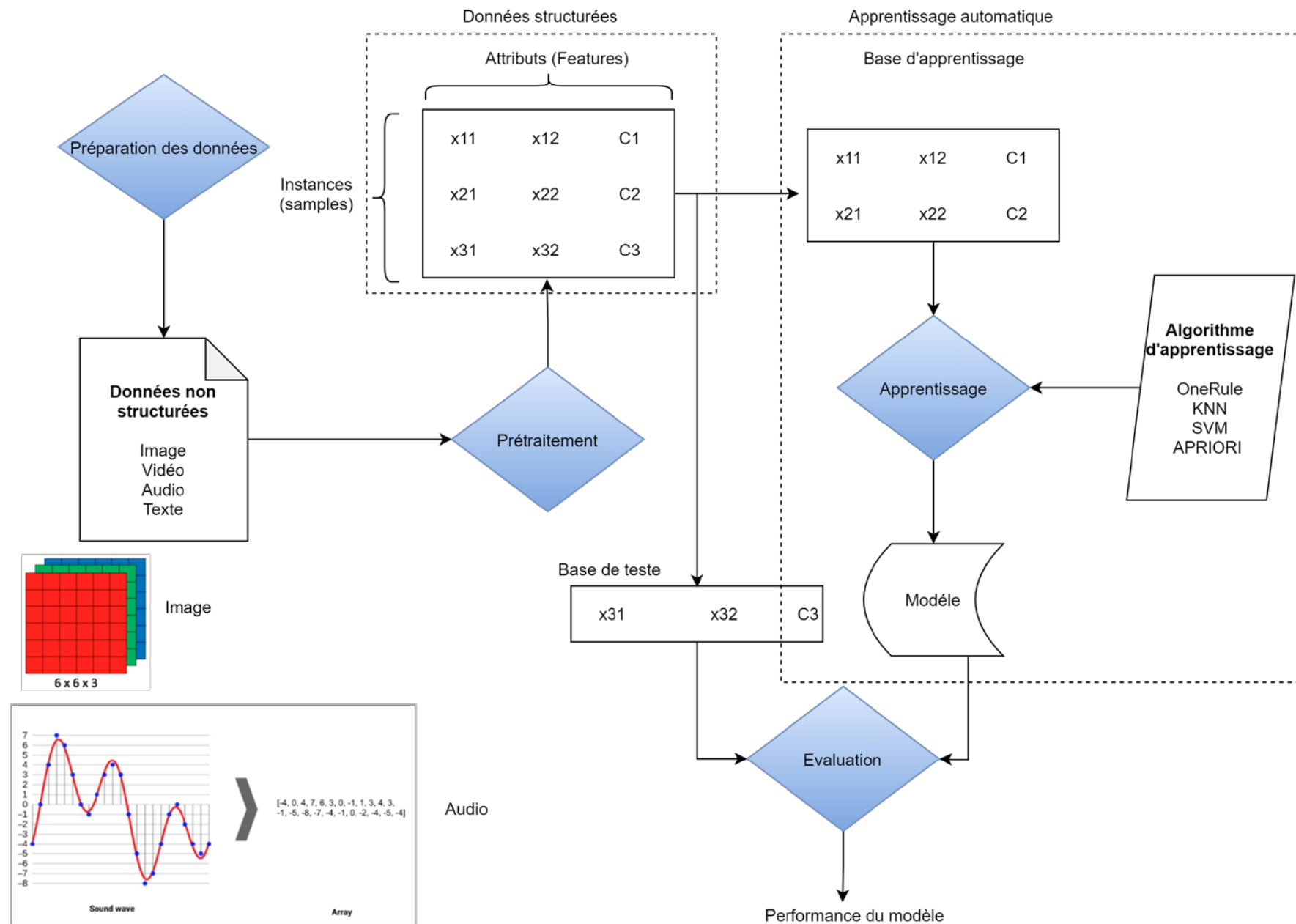
*Intelligence Artificielle et
Sciences de Données
(IASD)*

DR N. DIF

Program



Rappel



1. LINEAR AND NON LINEAR REGRESSION



Let's see first a real-world application of regression

Self driving car



Predict the appropriate angle at each moment t



$$f\left(\text{image of a road curve}\right) = -116.60^\circ$$


For more details, see:

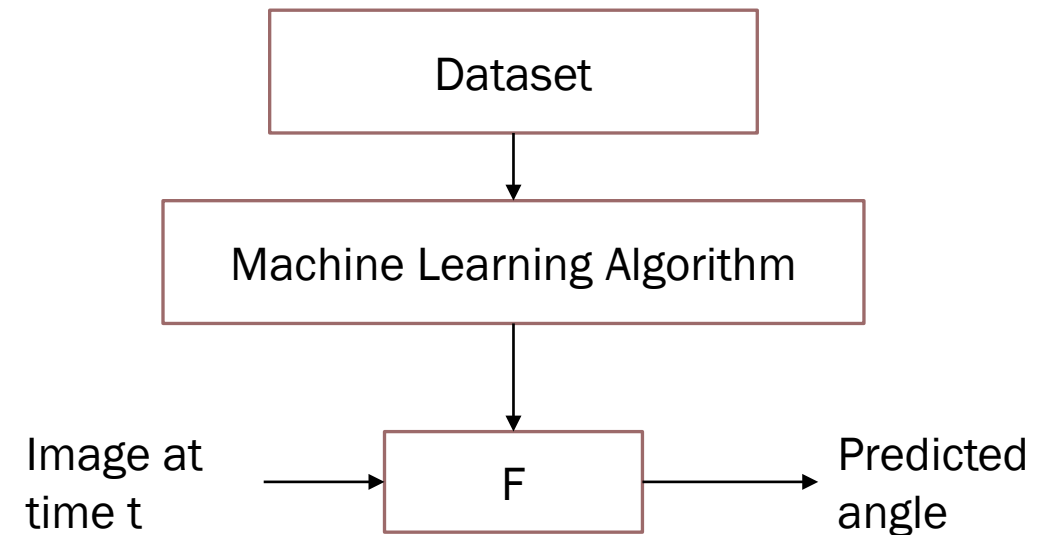
<https://www.youtube.com/watch?v=H0igiP6Hg1k>

1. LINEAR AND NON LINEAR REGRESSION



Let's see first a real-world application of regression

DATA	
x (INPUT DATA)	y (TARGET OUTPUT DATA)
	0.86°
	-99.82°
	-144.20°



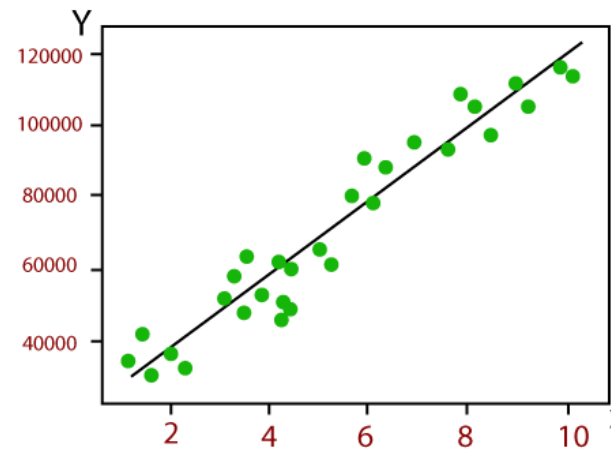
1. LINEAR AND NON LINEAR REGRESSION



How can we represent the function F that design the model? How can we extract knowledge from the training samples?

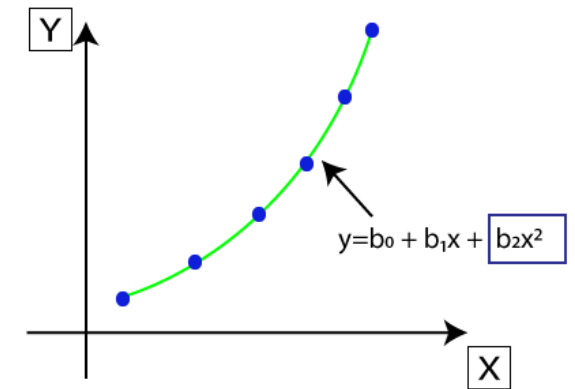
1. LINEAR AND NON LINEAR REGRESSION

- Statistical method that belongs to supervised learning problems.
- Regression models are specifically designed for datasets with numerical values.
- The class is represented in a continuous numerical form (e.g., temperature, age, salary, price).
- It explains the relation between a dependent variable (target) Y and one or more independent variables X_j ($j=1, \dots, q$) of a numerical type. The relation expressed between the two variables can be linear or nonlinear.



Régression linéaire

A straight line that separates the data points



Régression non linéaire

A graph that separates the data points

1. LINEAR AND NON LINEAR REGRESSION

- The selection of the regression algorithm type depends on the dataset's characteristics. In some situations, it becomes unfeasible to separate the data points using a straightforward straight line. In such cases, nonlinear regression is employed.

1. **Linear Regression:** estimation of monthly rent (loyer) based on the apartment surface, estimation of salary.
2. **Non linear Regression:** estimation of population growth over time.

- Linear regression is categorized into simple and multiple linear regression

Simple linear regression: uses a single independent variable ($Y_t = \beta_0 + \beta_1 X_t + \varepsilon_t, t = 1, 2, \dots, m$), Y_t is the dependent variable at time t,

Multiple linear regression : ($Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \dots + \beta_k X_{kt} + \varepsilon_t, t = 1, 2, \dots, m$) uses multiple independent variables.

X_t is the independent variable at time t, β_i are the parameters of the model, ε_t is the model's random error, m is the number of observations, and k is the number of independent variables.

2. SIMPLE LINEAR REGRESSION

- In a simple linear regression model, there is one dependent variable (endogenous) that is explained by a single independent variable (exogenous).
- Parameters β_0 and β_1 are estimated using the Ordinary Least Squares (OLS) method.
- Estimation is obtained by minimizing the sum of squared errors as follows:

$$\text{Min} \sum_{t=1}^m \varepsilon_t^2 = \text{Min} \sum_{t=1}^m (Y_t - \beta_0 - \beta_1 X_t)^2$$

To reach a minimum, the derivatives with respect to β_0 and β_1 must be equal to zero.

2. SIMPLE LINEAR REGRESSION

$$\frac{\partial S}{\partial \beta_0} = 0 \Leftrightarrow 2 \sum_{t=1}^m (Y_t - \beta_0 - \beta_1 X_t)(-1) = 0 \rightarrow \sum_{t=1}^m Y_t = m\beta_0 + \beta_1 \sum_{t=1}^m X_t \quad \dots (1)$$

$$\frac{\partial S}{\partial \beta_1} = 0 \Leftrightarrow 2 \sum_{t=1}^m (Y_t - \beta_0 - \beta_1 X_t)(-X_t) = 0 \rightarrow \sum_{t=1}^m Y_t X_t = \beta_0 \sum_{t=1}^m X_t + \beta_1 \sum_{t=1}^m X_t^2 \quad \dots (2)$$

From (1), we obtain :

$$\begin{cases} \widehat{\beta}_0 = \frac{\sum_{t=1}^m Y_t - \widehat{\beta}_1 \sum_{t=1}^m X_t}{m} \\ \widehat{\beta}_0 = \bar{Y} - \beta_1 \bar{X} \end{cases}$$

By replacing β_0 in (2), we obtain :

$$\begin{cases} \widehat{\beta}_1 = \frac{\sum_{t=1}^m X_t Y_t - \bar{Y} \sum_{t=1}^m X_t}{\sum_{t=1}^m X_t^2 - \bar{X} \sum_{t=1}^m X_t} \\ \widehat{\beta}_1 = \frac{\sum_{t=1}^m X_t Y_t - \bar{Y} \sum_{t=1}^m X_t}{\sum_{t=1}^m X_t^2 - m \bar{X}^2} \\ \widehat{\beta}_1 = \frac{\sum_{t=1}^m X_t Y_t - m \bar{X} \bar{Y}}{\sum_{t=1}^m X_t^2 - m \bar{X}^2} \\ \widehat{\beta}_1 = \frac{\sum_{t=1}^m (Y_t - \bar{Y})(X_t - \bar{X})}{\sum_{t=1}^m (X_t - \bar{X})^2} \end{cases}$$

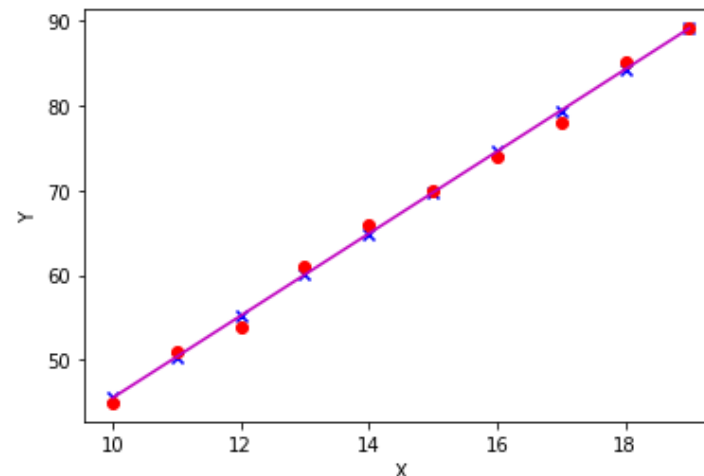
2. SIMPLE LINEAR REGRESSION

For the dataset bellow, determine the expected value Y for X=21

X	10	11	12	13	14	15	16	17	18	19	145
Y	45	51	54	61	66	70	74	78	85	89	-
XY	450	561	648	793	924	1050	1184	1326	1530	1691	10157
X ²	100	121	144	169	196	225	256	289	324	361	2185
\bar{X}	14.5				\bar{Y}	67.3					

$$\begin{cases} \widehat{\beta}_1 = \frac{10157 - 67.3 * 145}{2185 - 10 * (14.5)^2} = 4.8303 \\ \widehat{\beta}_0 = 67.3 - 4.8303 * 14.5 = -2.7394 \end{cases}$$

The predicted model is characterized by a good performance, as the distance between predicted and true values is minimal



Regression model: purple.
Predicted values: blue cross.
Actual values: red bubble

X	10	11	12	13	14	15	16	17	18	19
Y	45	51	54	61	66	70	74	78	85	89

2. SIMPLE LINEAR REGRESSION

Simple linear regression can only address problems that are linearly separable and characterized by a single attribute or feature, for example: predicting house prices based on surface. However, in most cases, a dataset is characterized by multiple attributes, such as predicting house prices based on surface and the type of the house. In such cases, multiple linear regression algorithms are employed.

Is the house id necessary for prediction? Why?

Sr. No.	Details	Price	Bedrooms	Bathrooms
1	23534368	221456	3	2
2	89756456	321234	4	3
3	45767857	134000	2	2
4	25756756	214679	3	1
5	23445466	213245	3	1

3. MULTIPLE LINEAR REGRESSION

The house price prediction problem is characterized by 3 features(details or surface (X_1 , Bathrooms X_2 , and Bedrooms X_3 , price Y):

- $F(x) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3$ to generalize $F(x) = \theta^T X = \theta_0 + \theta_1 X_1 + \dots + \theta_n X_n$, where

n is the number of features.

m is the number of samples.

The parameters to fix : $\theta_0, \theta_1, \dots, \theta_n$

Purpose : predict parameters $\theta_0, \theta_1, \dots, \theta_n$ by minimizing **the cost function (convex function)**:

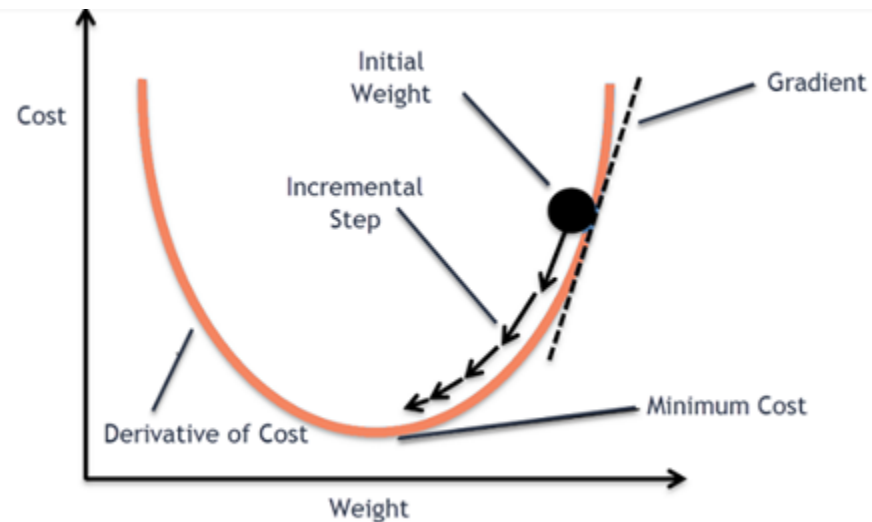
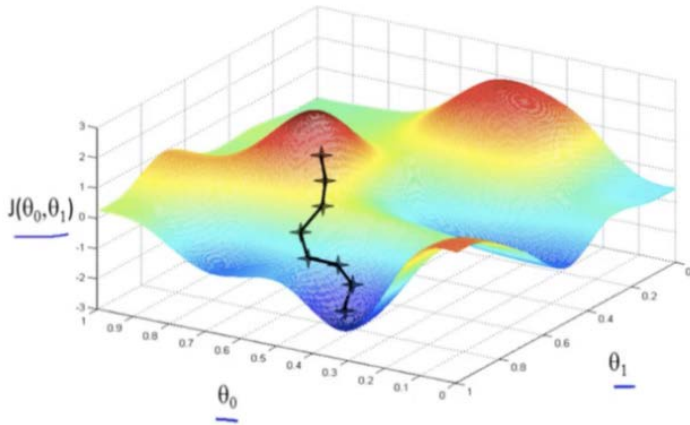
- $J(\theta_0, \theta_1, \dots, \theta_n) = \frac{1}{2m} \sum_{i=1}^m (f(X^{(i)}) - Y^{(i)})^2$
- One of the techniques employed to estimate all the parameter $\theta_0, \theta_1, \dots, \theta_n$ is the **gradient descent optimization technique**.

3. MULTIPLE LINEAR REGRESSION

Gradient descent is an iterative stochastic optimization algorithm used to find the minimum of a function.

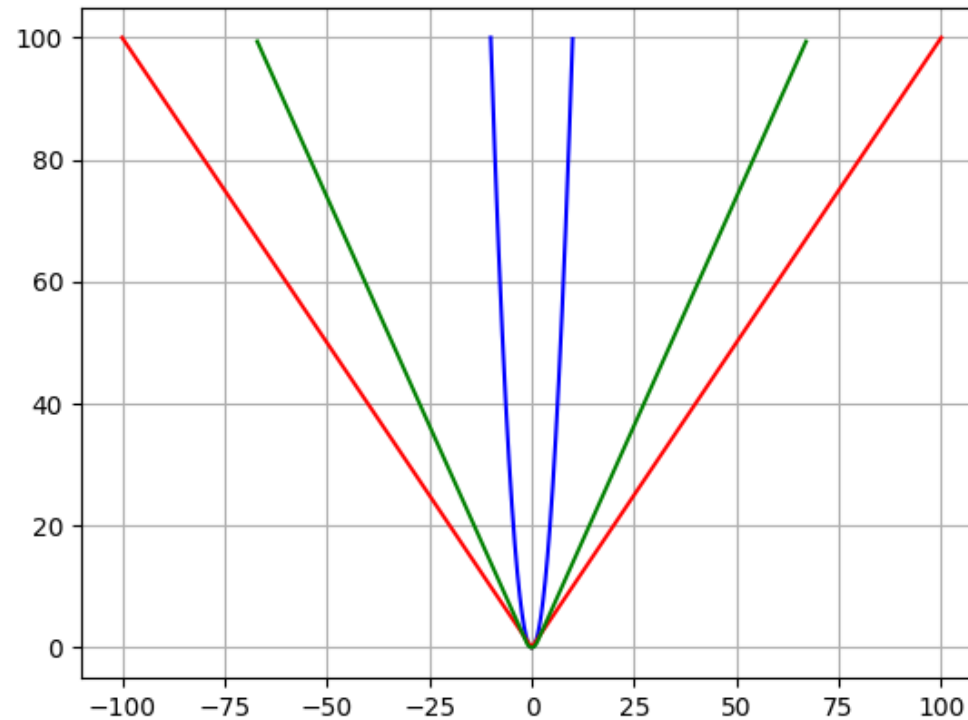


What is the motivation behind using gradient descent? Why is the utilization of partial derivatives important? Can you explain the concepts of local minimum and global minimum?



3. MULTIPLE LINEAR REGRESSION

Loss functions



MAE (red), MSE (blue), and Huber (green) loss functions

3. MULTIPLE LINEAR REGRESSION

In the gradient descent technique, the parameters update is computed as follows:

$$\theta_j = \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta), j \in (1, 2, \dots, n), J(\theta_0, \theta_1, \dots, \theta_n) = \frac{1}{2m} \sum_{i=1}^m (f(X^{(i)}) - Y^{(i)})^2$$

$$\theta_j = \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (F(x^{(i)}) - y^{(i)}) X_j^{(i)}, \theta_0 = \theta_0 - \alpha \frac{1}{m} \sum_{i=1}^m (F(x^{(i)}) - y^{(i)})$$

J is the cost function, and n is the number of features, m is the number of samples, $\frac{\partial}{\partial \theta_j} J(\theta)$ is the partial derivative with respect to θ_j , α is the learning rate.

Algorithm : Gradient descent for multiple linear regression

Initialize randomly θ_j

Repeat {

Update simultaneously θ_j for $j \in (1, 2, \dots, n)$

}

When should we stop?
How to set the learning rate?
What is its use?

3. EVALUATION METRICS IN REGRESSION

- **Mean Absolute Error (MAE):** is the average of the differences between actual values and predicted values. This measure should be minimized.

$$\text{MAE} = \frac{1}{n} \sum_{j=1}^n |y_j - \hat{y}_j|$$

n is the number of observations, y_j is the true value for the observation j and \hat{y}_j is the predicted value.

- **Root Mean Squared Error (RMSE):** is the most commonly used evaluation metric in regression. Unlike MAE, this metric gives more weight to large errors. RMSE has the advantage of penalizing large errors more than MAE.

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{j=1}^n (y_j - \hat{y}_j)^2}$$

- **R-squared:** R-squared (R^2), also known as the coefficient of determination, quantifies the degree to which the regression line fits the actual data, expressed as a percentage between 0 and 100. $R^2 = 1 - \frac{\sum_{j=1}^n (y_j - \hat{y}_j)^2}{\sum_{j=1}^n (y_j - \bar{y})^2}$

3. EVALUATION METRICS IN REGRESSION

The following table presents RMSE and MAE errors of a car braking distance estimation system

	Erreur 1	Erreur 2	MAE	RMSE
Modèle A	10	0	5	7
Modèle B	6	5	5.5	5.52

- In terms of MAE, model A is more efficient, even if it made an error of 10 meters, which can lead to damage in such systems. On the other hand, model B is more efficient in terms of RMSE.

References

- Adrew Ng. Learning: Regression and Classification, Coursera. <https://fr.coursera.org/learn/machine-learning>
- Welch Labs. Self Driving Cars [S1E2: ALVINN]. <https://www.youtube.com/watch?v=H0igiP6Hg1k>
- BOUKRIF Nouara. Polycopié : *Régression Linéaire simple et multiple*.
- Ritchie Ng. Linear Regression with Multiple Variables. <https://www.ritchieng.com/multi-variable-linear-regression/>
- Matthew MacFarquhar. Gradient Descent: The Magic Behind ML. <https://blog.devgenius.io/gradient-descent-the-magic-behind-ml-6dc17668d0af>