

# 1) Formulas - Régression Linéaire(RL)

1-1. Modèle de base :

$$\hat{y} = \beta_0 + \beta_1 x$$

1-2. Erreurs :

$$e_{ii} = y_i - \hat{y}_i$$

1-3. RSS / SSE (erreur totale):

$$SSE = RSS = \sum (y_i - \hat{y}_i)^2$$

1-4. MSE (erreur moyenne):

$$MSE = \frac{1}{n} \sum (y_i - \hat{y}_i)^2$$

1-5. RMSE:

$$RMSE = \sqrt{MSE}$$

1-6. MAE:

$$MAE = \frac{1}{n} \sum |y_i - g_i|$$

~~Scatter plot of residuals~~

1-7. Variance ( $\sigma^2$ ):

$$Var(x) = \frac{1}{n} \sum (x_i - \bar{x})^2$$

1-8. Covariance ( $\sigma_{xy}$ ):

$$\text{Cov}(x, y) = \frac{1}{n} \sum (x_i - \bar{x})(y_i - \bar{y})$$

1-9. Coefficient de corrélation

$$r = \frac{\text{Cov}(x, y)}{\sigma_x \cdot \sigma_y}; \quad \sigma_x = \sqrt{Var(x)}$$

## 1-10 : Coefficient de détermination;

$$R^2 = r^2 = \left( \frac{\text{cov}(x, y)}{s_x \cdot s_y} \right)^2$$

$$R^2 = 1 - \frac{SSE}{SST}$$

## 1-10 : Coefficient de détermination

+ Si  $R^2 = 1$  : Le modèle explique 100 % la variation (Tous les points sont collés à la droite)

+ Si  $R^2 = 0,4$  : Le modèle explique 40 % de la variété de la variable (60 % qui reste c'est pour les autres facteurs)

1-12. SST (variation totale).

$$SST = \sum (y_i - \bar{y})^2$$

1-12. SSR (variable expliquée)

$$SSR = SST - SSE$$

c) Regression Multiple (plusieurs x).

Q. 1. Modèle :

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 - \beta_3 x_3$$

en matrice :

$$\hat{y} = X \beta$$

$$\beta = (X^T X)^{-1} X^T y$$

### 3) Régression Polynomiale:

$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_k x^k$$

↳ C'est juste une RL avec nouvelles variables.

$$x, x^2, x^3$$

### 4) Résente de gradient:

#### 4-1 Mise à jour d'un paramètre.

$$\beta_j := \beta_j - \alpha \frac{\partial J}{\partial \beta_j}$$

#### 4-2 Cost Function ~~Loss Function~~:

$$J(\beta) = \frac{1}{2m} \sum (y_i - \hat{y}_i)^2$$

5) Régression logistique (classification  
Binnaire).

### 5.1. Modèle

$$\hat{y} = \sigma(z) \text{ où } z = \beta_0 + \beta_1 x$$

### 5.2. Sigmoid:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$

### 5.3. seuil:

$$g = \begin{cases} 1 & \text{si } \sigma(z) > 0,5 \\ 0 & \text{sinon} \end{cases}$$

## 6) KNN:

6-1) Distance euclidienne:

$$d(A, B) = \sqrt{(x_a - x_b)^2 + (y_a - y_b)^2}$$

6-2) Vote majoritaire (classification)

Classe = classe la plus fréquente  
parmi les  $K$  plus proches

6-3) Moyenne (régression)

$$\bar{y} = \frac{1}{K} \sum y_i$$

## 2.0) Métriques de classification

		Valeurs réelles	
		Positives	Négatives
Prédites	Positives	TP	FP
	Négatives	FN	TN

Valeurs précises

1) Accuracy (taux de données prédictives)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

→ Sur tous ce que le modèle prédit,  
combien sont corrects ?

### 2) Precision :

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

→ Parmi tous ce que le modèle a prédit positif, combien étaient vraiment positifs?

↳ Sensible aux faux positifs

### 3) Recall

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

↳ Parmi tous les vrais positifs, combien le modèle a détectés?

Sensible aux faux négatifs

k) F1-score:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$