Tout savoir sur le k-means

Partie 1 : La théorie

Présenté par **Morgan Gautherot**

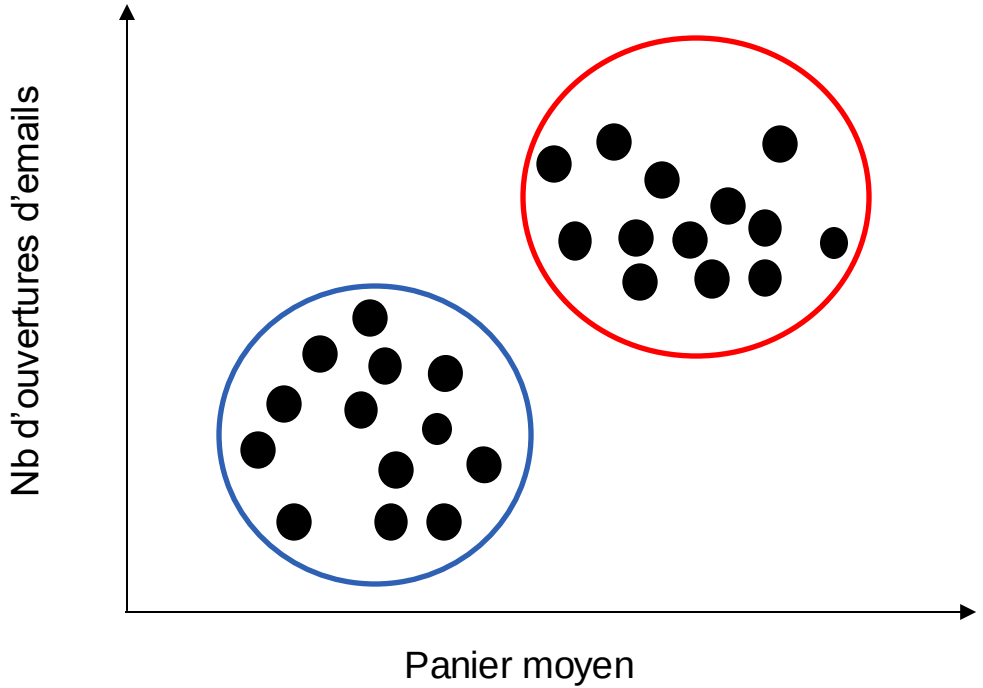
Problème de clustering

X → k

	Nb d'e-mails ouverts (x_1)	Nb de produits achetés (x_2)	Panier moyen (x_3)
1	12	3	120
2	0	1	40
3	30	10	1800
4	14	5	799
…	…	…	…
m	25	2	260

Jeu d'entraînement pour mieux comprendre nos utilisateurs

Visualisation



Qu’est ce qu’une distance ?

Distance euclidenne $$d(x,y) = \sqrt{\sum_{i=1}^{n}(y_i - x_i)^2}$$

Distance de Manhattan $$d(x,y) = (\sum_{i=1}^{m}|x_i - y_i|)$$

Distance de Minkowski $$d(x,y) = (\sum_{i=1}^{n}|x_i - y_i|)^{1/p}$$

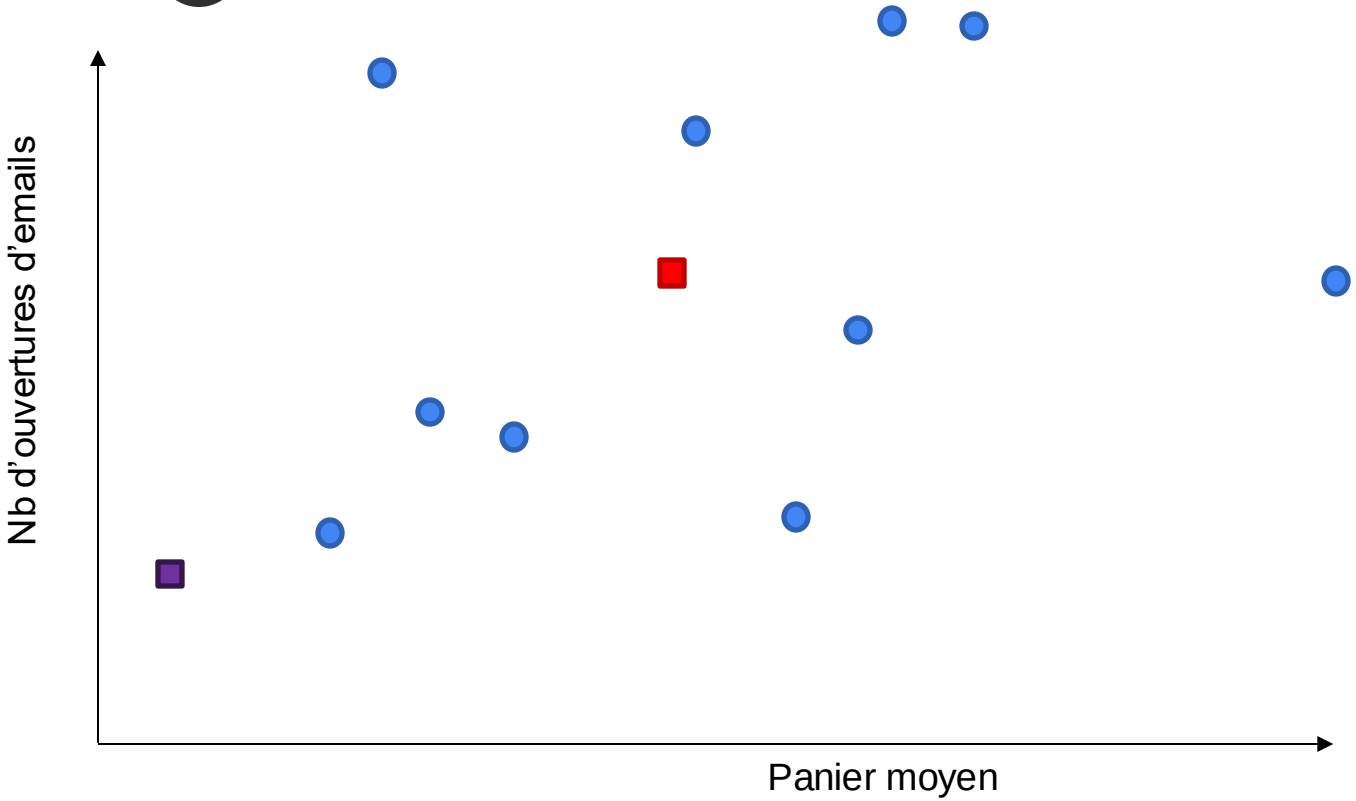
⋮ ⋮

Standardisation des données

$$x_{std} = \frac{x - mean(x)}{std(x)}$$

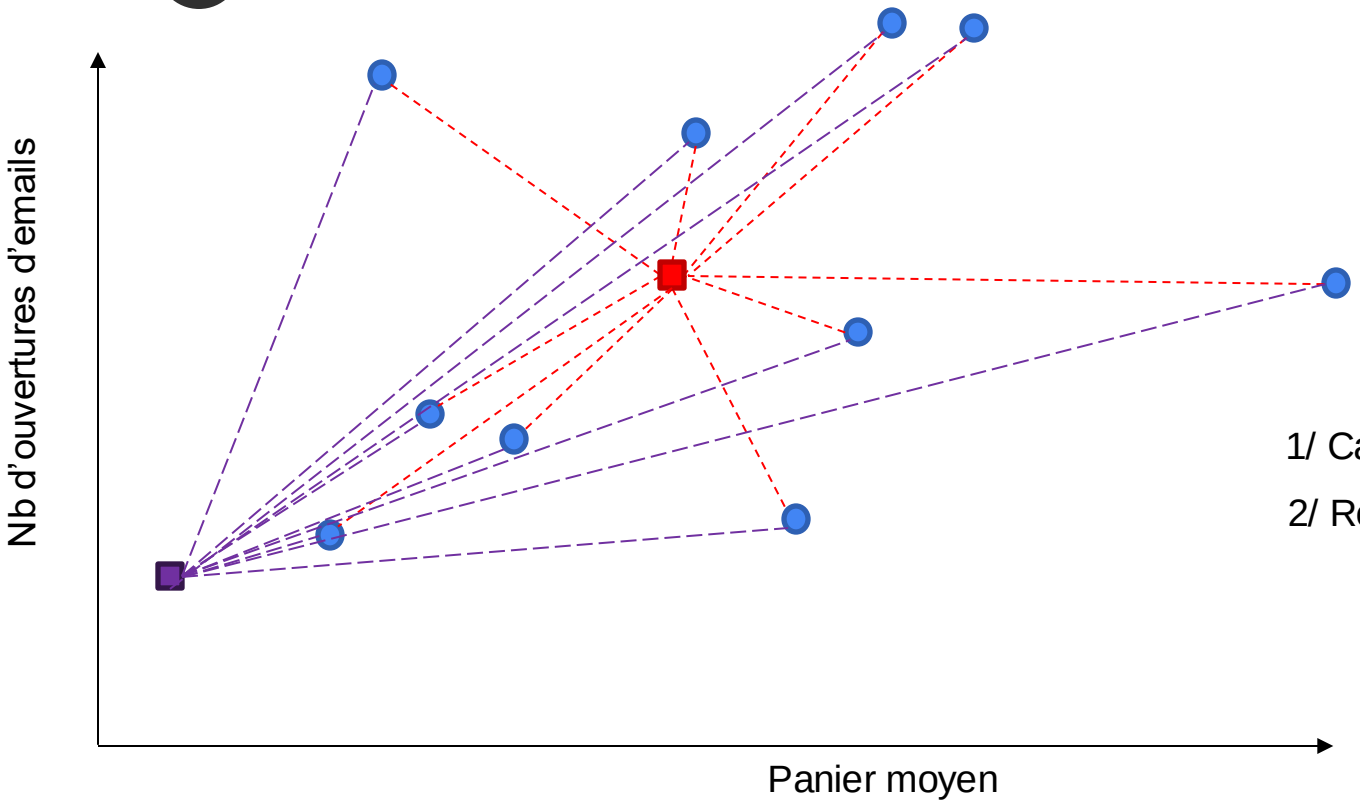
K=2

Initialisation



K=2

Etape 1

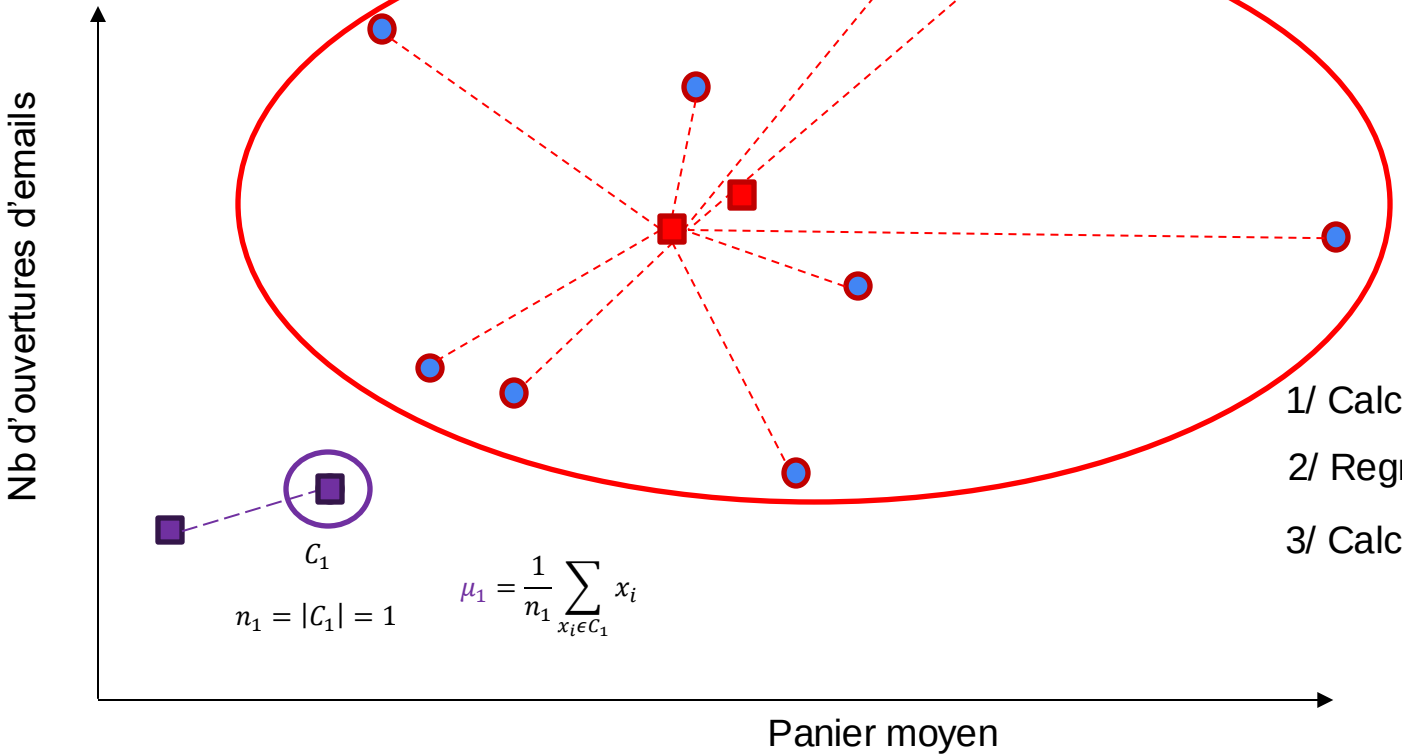


1/ Calcul des distances

2/ Regroupement en classe

K=2

Etape 1



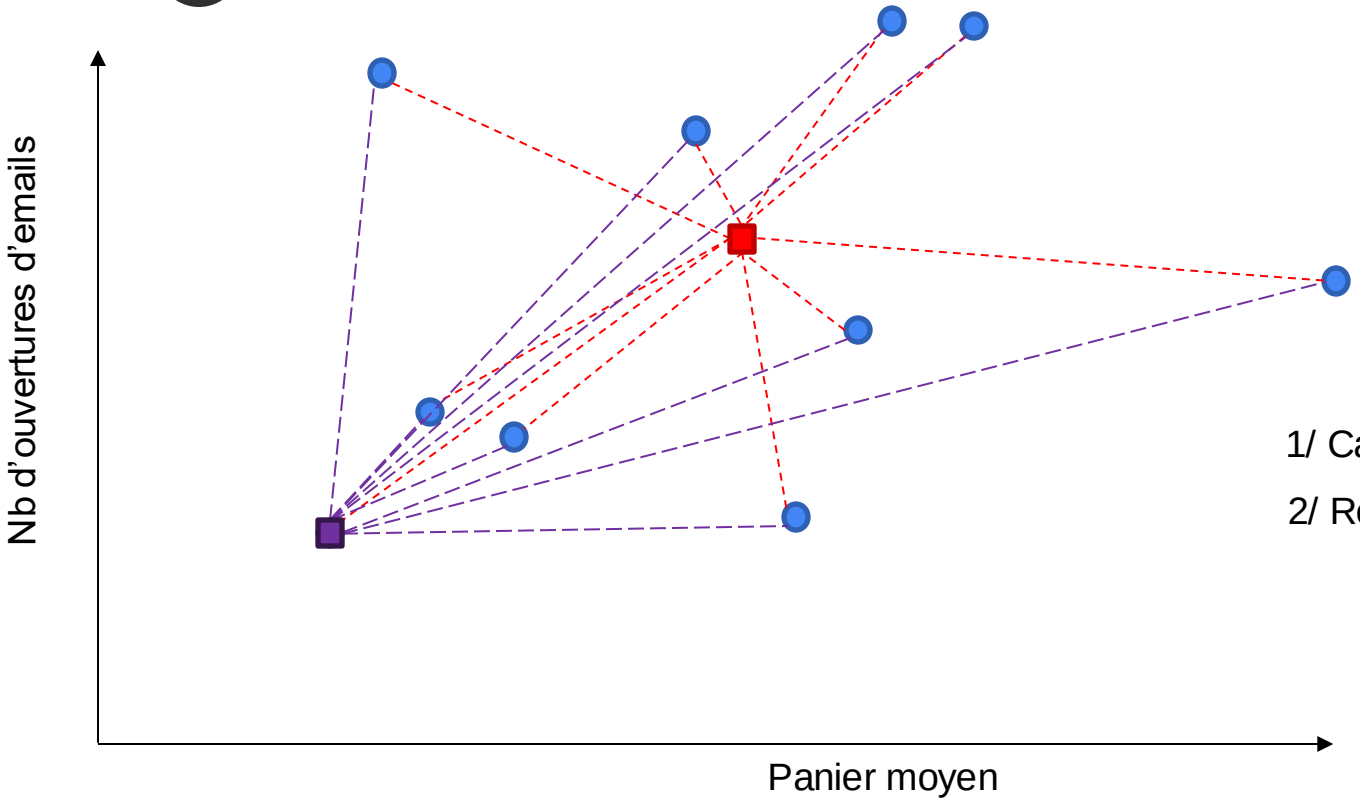
1/ Calcul des distances

2/ Regroupement en classe

3/ Calcul des nouveaux centroïdes

K=2

Etape 2

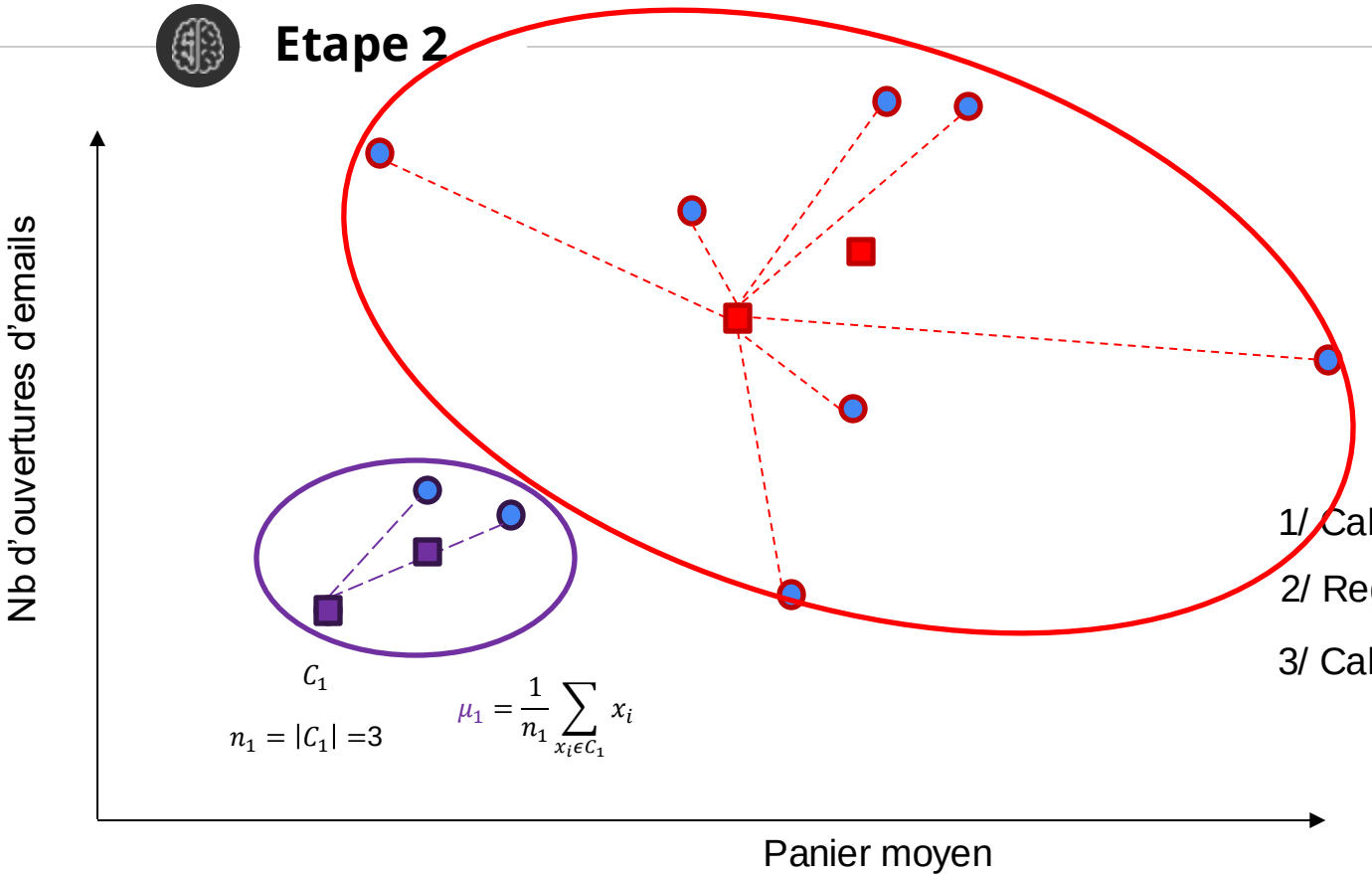


1/ Calcul des distances

2/ Regroupement en classe

K=2

Etape 2



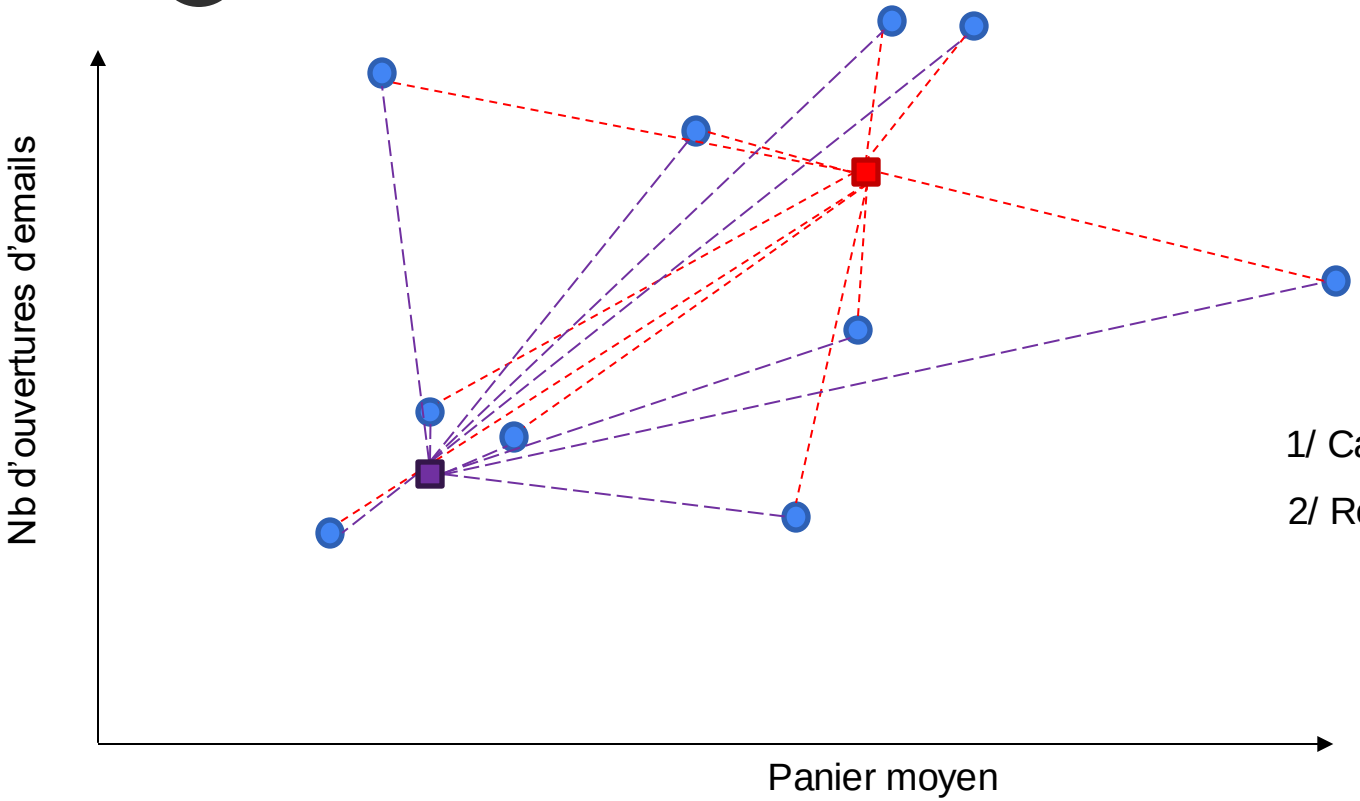
1/ Calcul des distances

2/ Regroupement en classe

3/ Calcul des nouveaux centroïdes

K=2

Etape 3

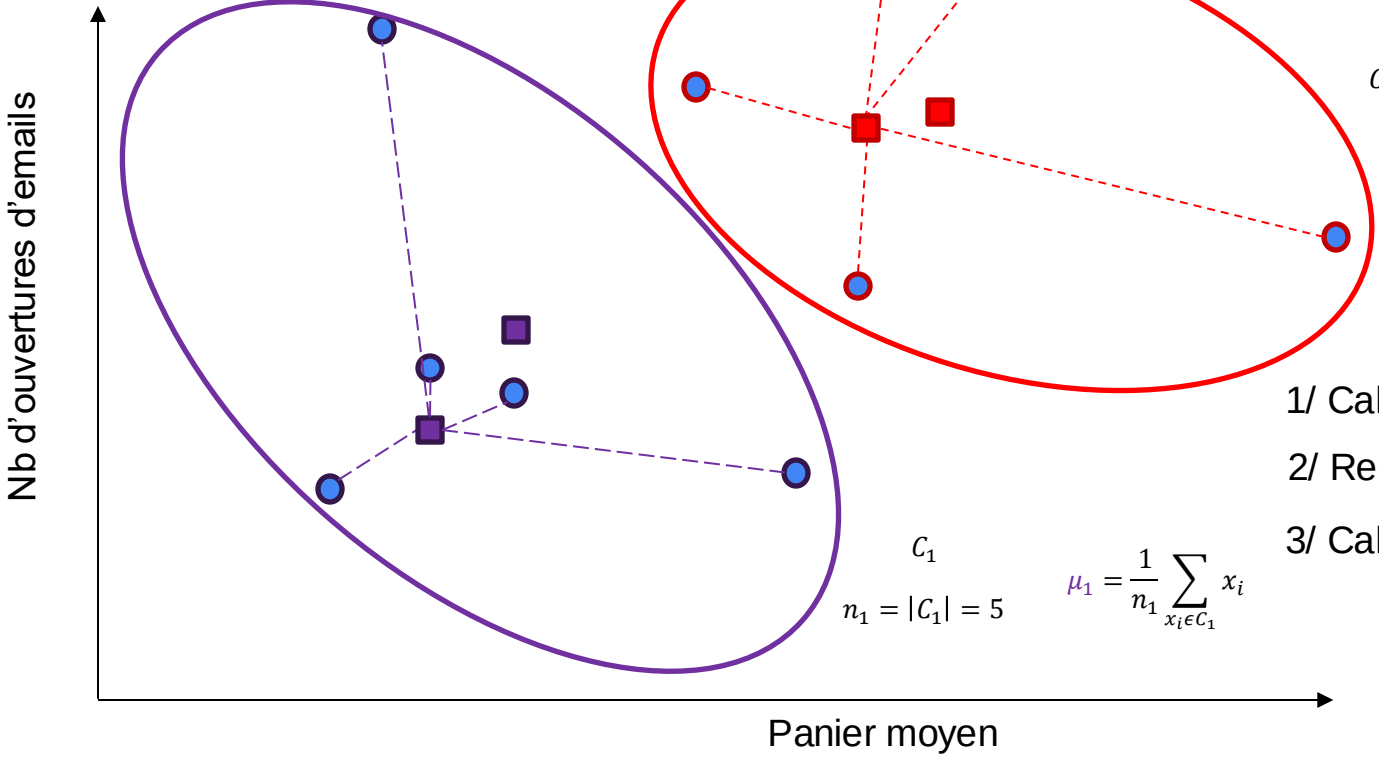


1/ Calcul des distances

2/ Regroupement en classe

K=2

Etape 3



C_2 $n_2 = |C_2| = 5$

$$\mu_2 = \frac{1}{n_2} \sum_{x_i \epsilon C_2} x_i$$

C_1

$n_1 = |C_1| = 5$

$$\mu_1 = \frac{1}{n_1} \sum_{x_i \epsilon C_1} x_i$$

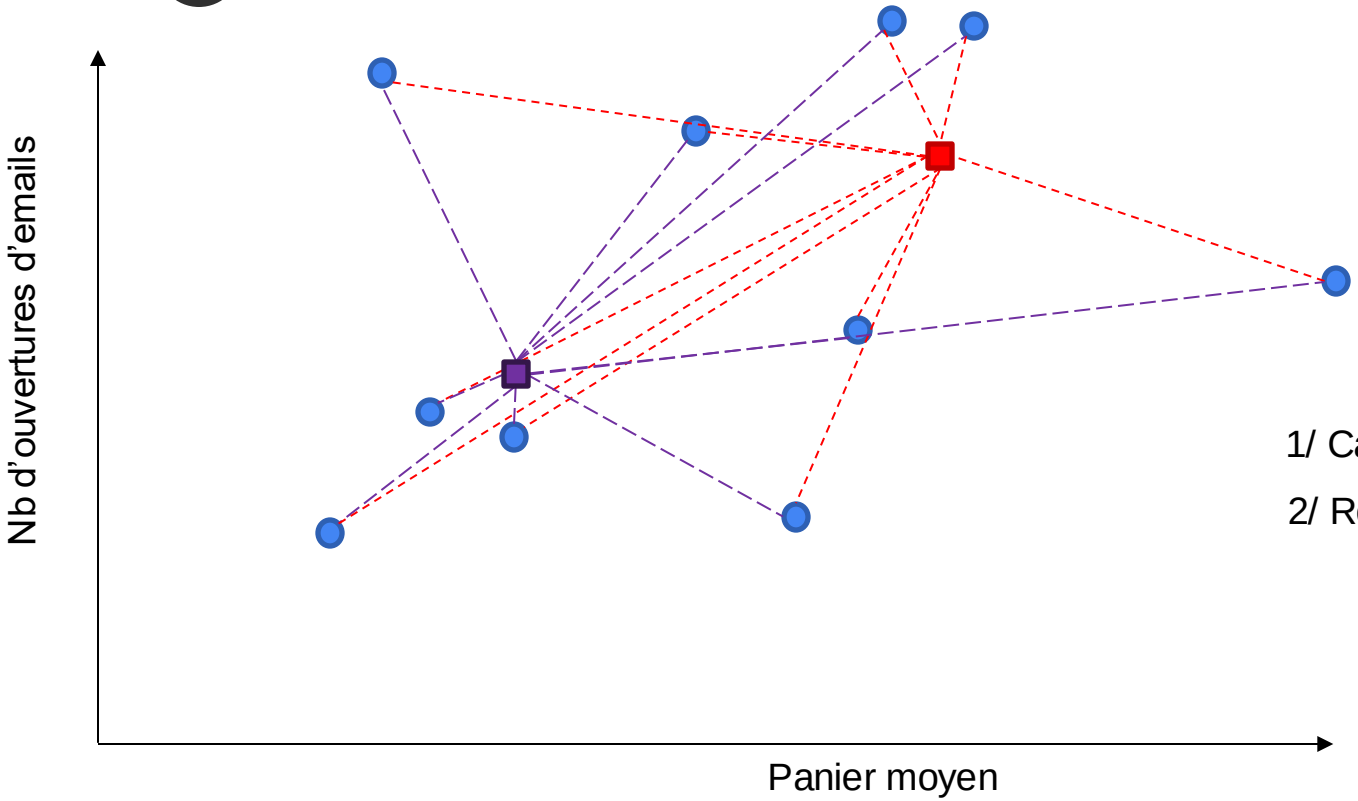
1/ Calcul des distances

2/ Regroupement en classe

3/ Calcul des nouveaux centroïdes

K=2

Etape 4



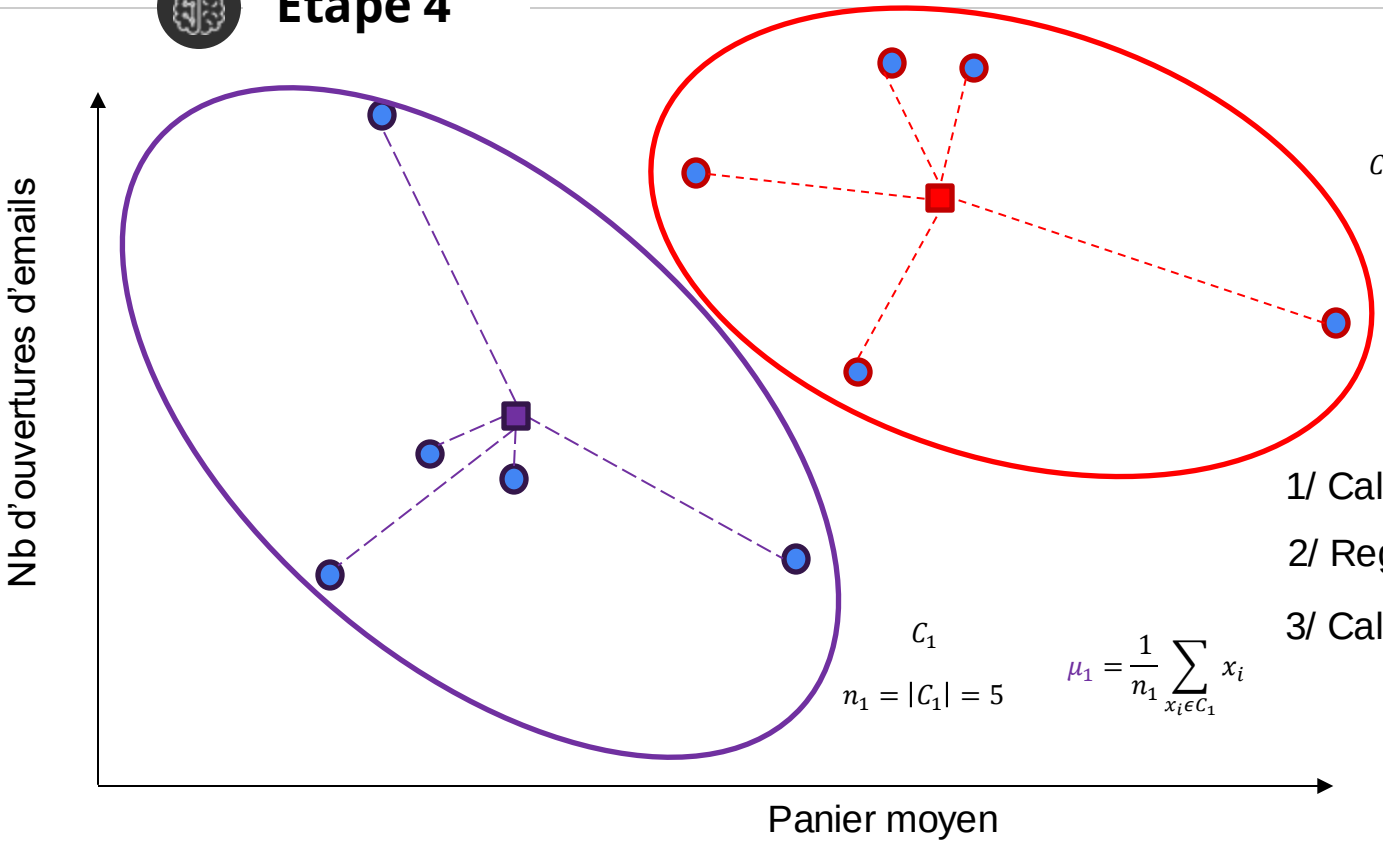
1/ Calcul des distances

2/ Regroupement en classe

Fin de l'entraînement

K=2

Etape 4



C_2 $n_2 = |C_2| = 5$

$$\mu_2 = \frac{1}{n_2} \sum_{x_i \epsilon C_2} x_i$$

C_1

$n_1 = |C_1| = 5$

$$\mu_1 = \frac{1}{n_1} \sum_{x_i \epsilon C_1} x_i$$

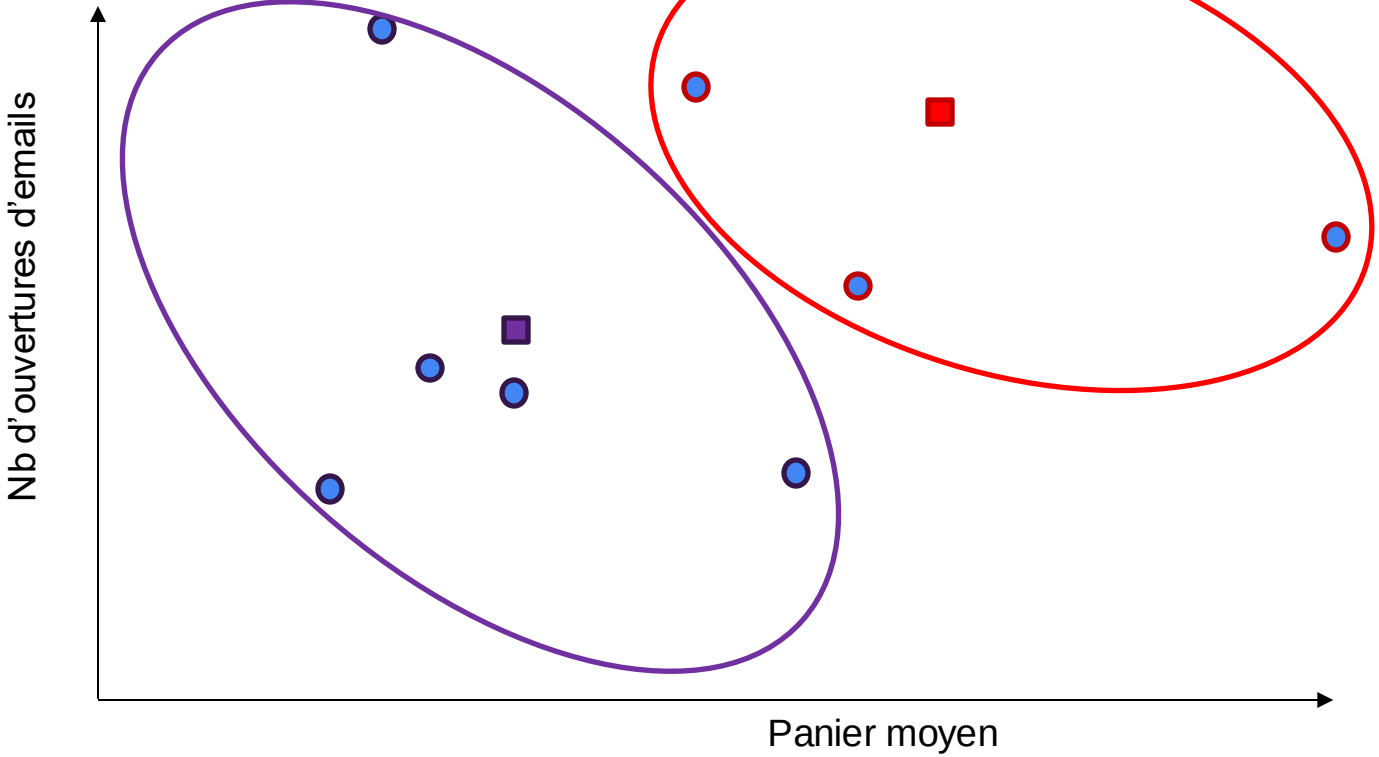
1/ Calcul des distances

2/ Regroupement en classe

3/ Calcul des nouveaux centroïdes

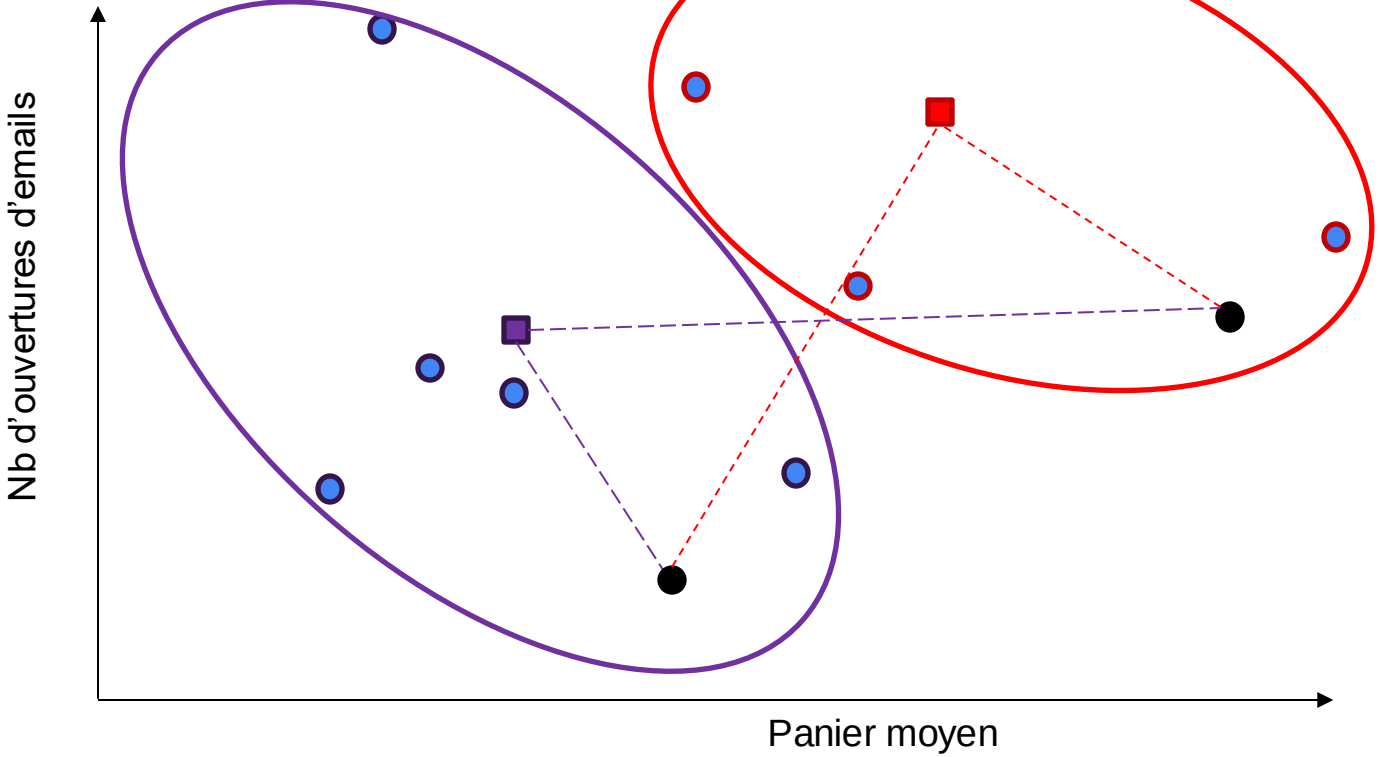
K=2

Clusters finaux



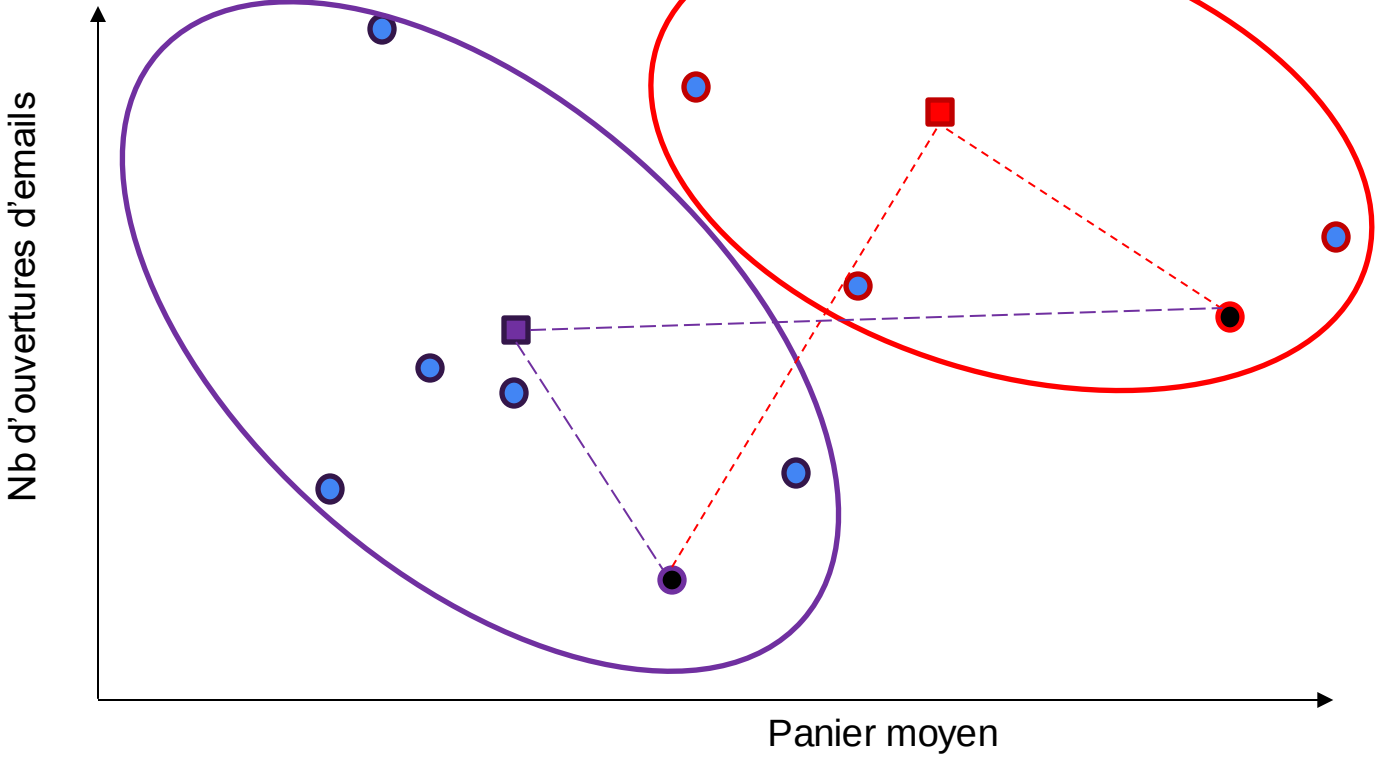
K=2

Nouvelles données



K=2

Nouvelles données



Avantages et inconvénients

- Simple à comprendre et à utiliser
- Peut-être appliquez à de nouvelles données

- Valeur fixe de K
- Sensible à l'initialisation
- Sensible aux valeurs aberrantes
- Distribution sphérique uniquement