Données web et multimédia

Cours 1 : Introduction à l'apprentissage

Alexis Lechervy

Sommaire

L'informatique

Qu'est ce que l'informatique ?

C'est le développement de techniques permettant **le traitement automatique de données** par des **machines** (ordinateur, robots, automates, systèmes embarqués...).

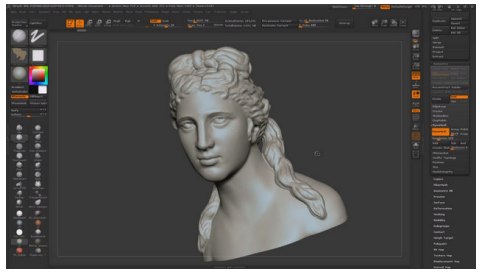
Les enjeux de l'informatique

- Permettre l'exécution de tâches toujours plus précises, en toujours moins de temps.
- Assister les humains dans des tâches de plus plus complexe.
- Pouvoir gérer une quantité d'information toujours plus croissante. (Youtube : >400h de nouvelles vidéos par minute, Facebook : >2 000 de nouvelles photos par seconde...)
- Traiter des données toujours plus varié.

Les données en informatique

Acquisition et synthèse de données

- l'acquisition de nouvelles données par l'utilisation de différentes capteurs analysant le monde réel (image, son, vidéo...)
- Création de nouveau contenu et synthèse de contenu..

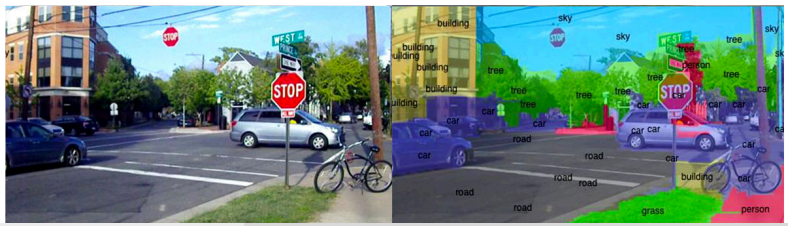


Amélioration de données existantes

- Correction,retouches de données,
- Amélioration/manipulation dimages...

Analyse de données et extraction de connaissance

- Analyse d'image (camera de surveillance...),
- Analyse des évolutions boursières,
- Aide au diagnostic, détection de pathologie, proposition de traitement...



Apprendre pour mieux comprendre et analyser

Apprendre ? Qu'est ce que c'est ?

- **Apprendre c'est s'adapter** à des situations **nouvelles et inconnues** en prenant en compte l'expérience passée.
- Apprendre est une **propriété** humaine **essentielle**.
- Apprendre signifie **s'améliorer afin dêtre meilleur**.

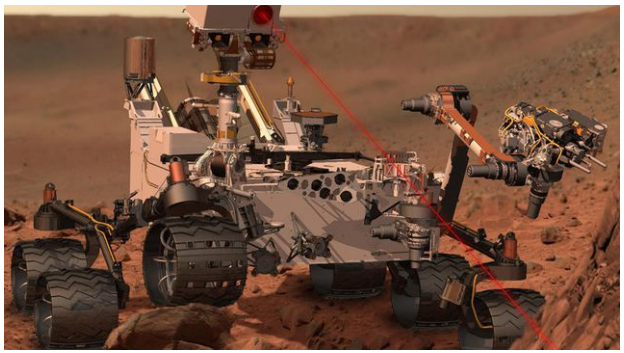
Mais pourquoi apprendre à un ordinateur ?

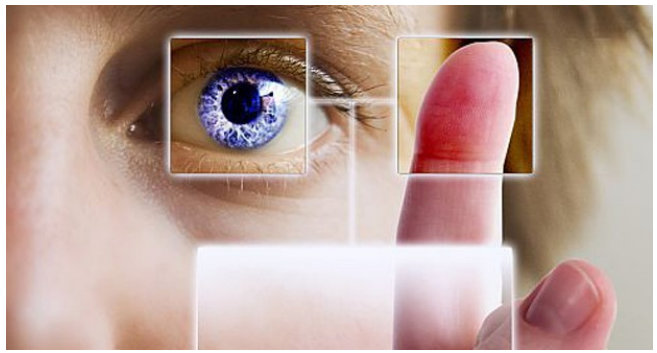
- Pouvoir gérer une quantité de données très importante de manière automatique ou de manière temps réels.
- Pouvoir effectuer une action dans un contexte non prévu préalablement sans l'intervention d'un humain.
- Pouvoir prévoir des comportements ou des évolutions pour aider un humain dans sa prise de décision.

L'Apprentissage en informatique, quand faut-il l'utiliser ?

Quand ?

- L'expertise humaine n'est pas possible (ex : navigation sur Mars)
- La quantité d'information est trop grande pour être traitée par un humain (ex : recherche d'une personne dans une base d'image)
- Besoin de traitement temps réel (ex : mise au point d'un appareil photo sur un visage)
- Automatisation d'une chaine de traitement (ex : détection d'anomalie sur une chaine de montage)
- Les êtres humains ne savent pas expliquer leurs expertises (ex : reconnaissance de la parole)
- Trouver une solution optimale à un problème (ex : trouver le meilleur modèle)...





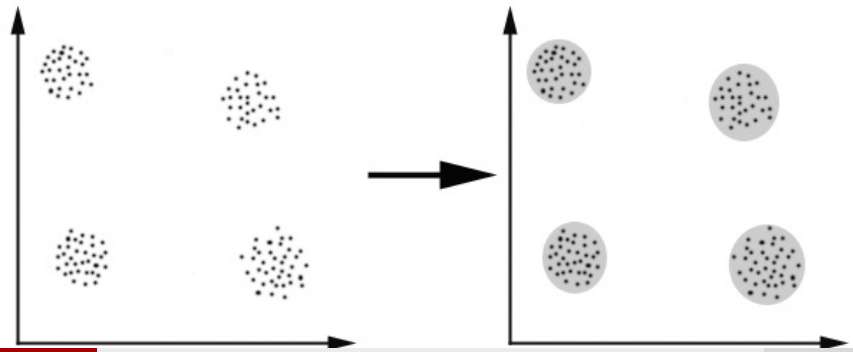
Sommaire

L'apprentissage non supervisé (Clustering)

Principes

- Diviser les données en plusieurs groupes séparés,
- Extraire une connaissance organisée sans intervention humaine,
- les données les plus similaires sont associées au sein d'un groupe homogène
- les données considérées comme différentes se retrouvent dans d'autres groupes distincts
- Pas d'apriori sur les données
- Il y a une seule entrée, les données collectées

Exemple de clustering

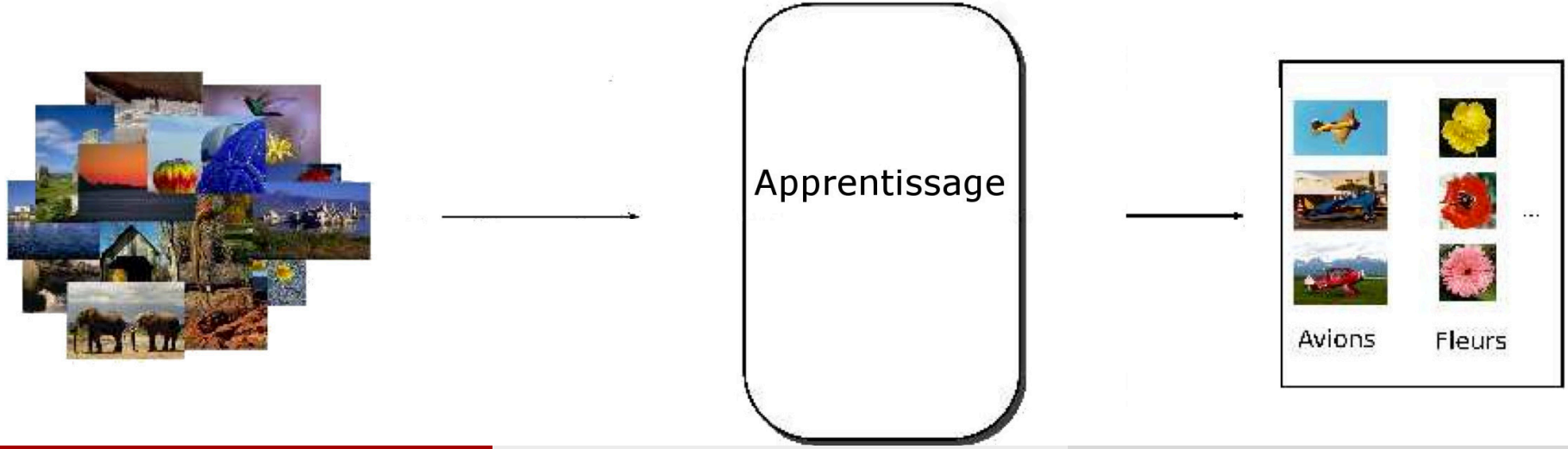


L'apprentissage supervisé

Principes

- On détermine automatiquement une règle à partir de données d'apprentissage annotées par un expert,
- Un expert a défini un ensemble de couples (donnée,label),
- Il y a un apriori sur les données,
- Les données entrées sont des couples (données collectées, observations).

Exemple : Catégorisation d'image



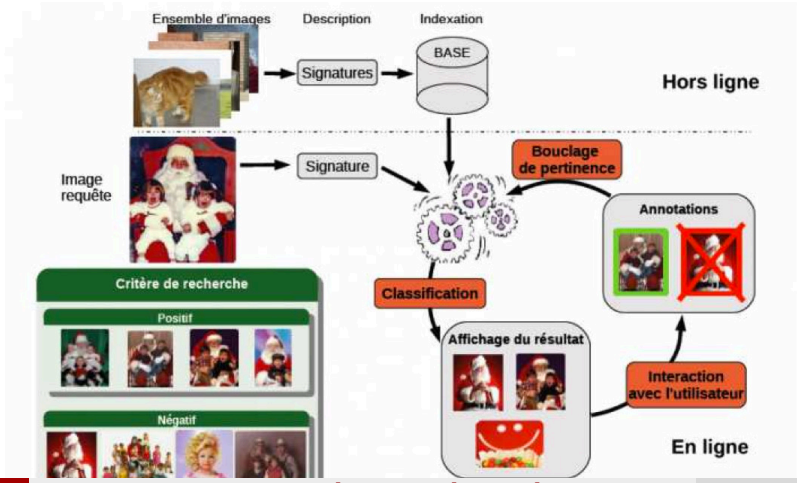
L'apprentissage semi-supervisé

Principes

- On dispose de quelques exemples labellisés.
- Les autres ne le sont pas.
- Permet de travailler avec moins de labels.

Exemples d'apprentissage non supervisé

- L'apprentissage interactif.
- L'apprentissage sur des ensembles de données trop important.

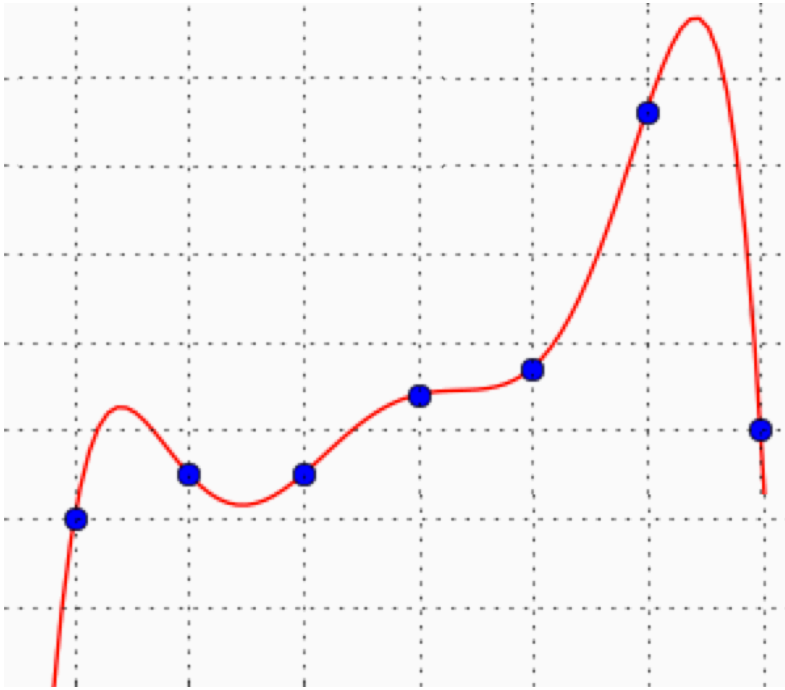


La régression

Principes

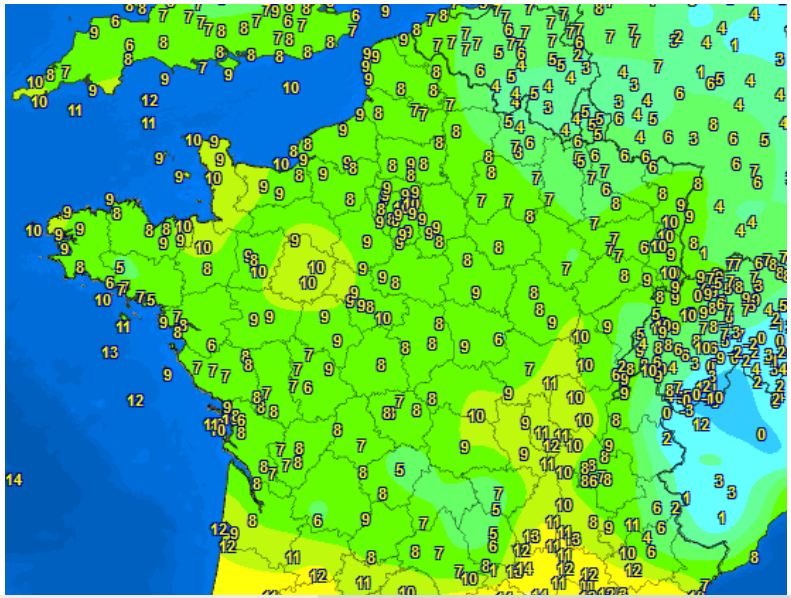
- Trouver la relation entre une variable et une ou plusieurs autres variables.
- Les valeurs de sorties de la fonction recherchée sont des valeurs réelles, non discrètes.

Exemple



Exemple d'application

Faire une carte de température en fonction de points de prise de mesures :

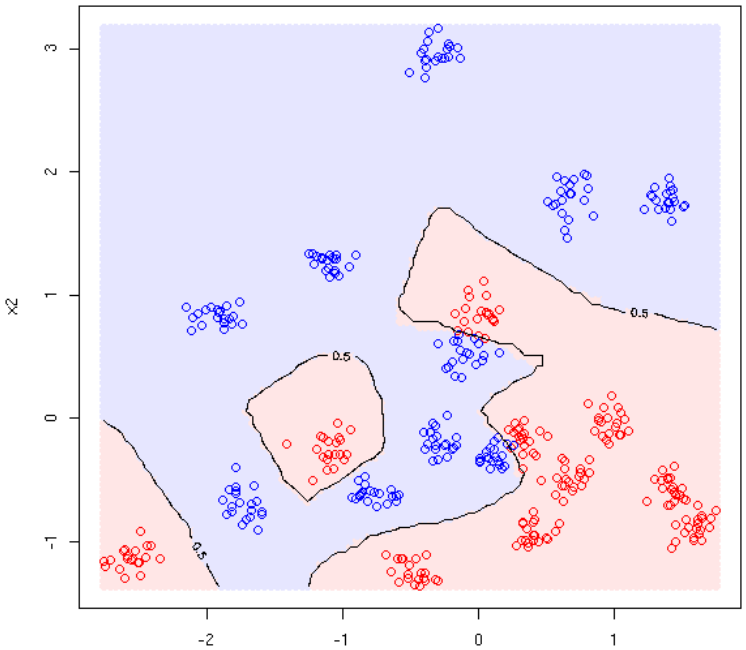


La classification

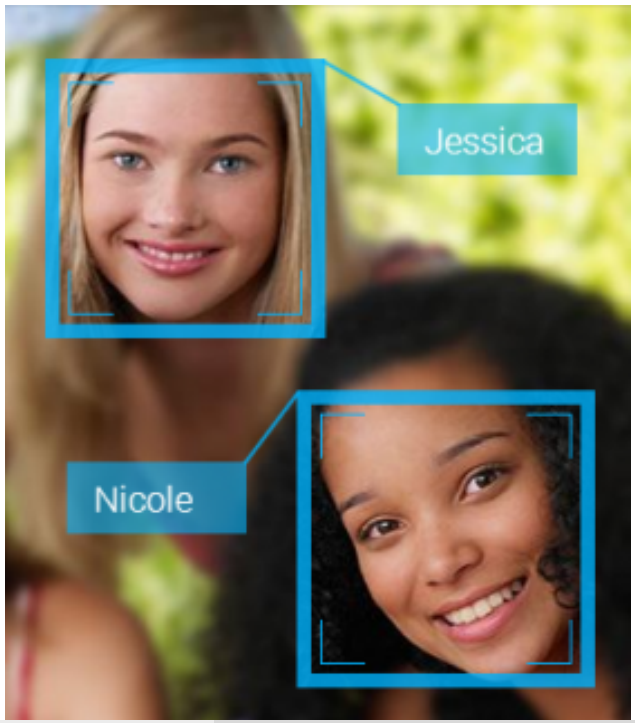
Principes

- Attribuer une classe à chaque objet.
- Les valeurs de sorties sont des valeurs discrètes correspondant à un numéro de classe.

Exemple



Exemple d'application : la reconnaissance faciale

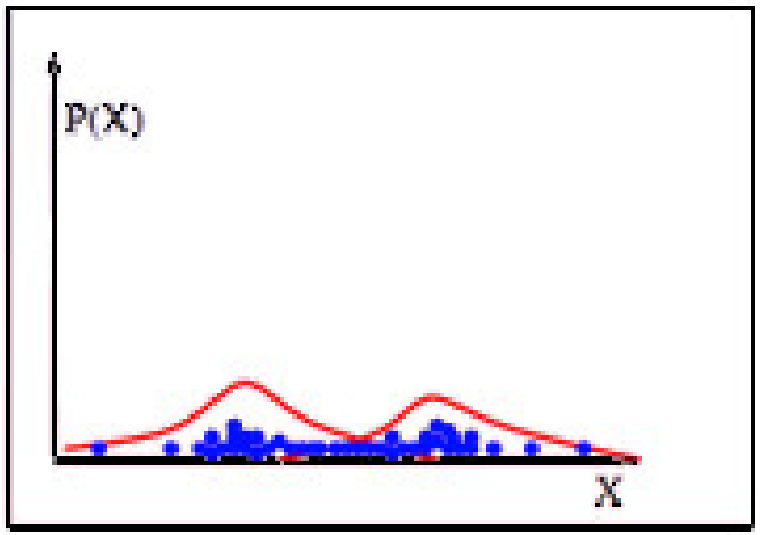


Estimation de densité

Principes

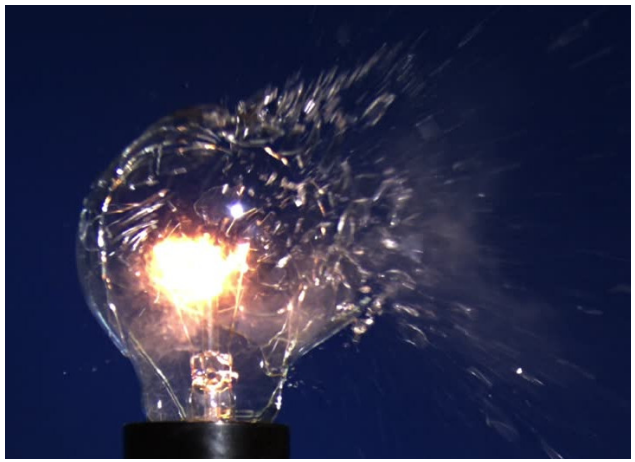
- Trouver les paramètres d'une loi de probabilité permettant d'estimer au mieux une distribution de points.

Exemple



Exemple d'application

Quelle est la probabilité qu'une gamme de produit soit défaillant au bout de x temps ?



Sommaire

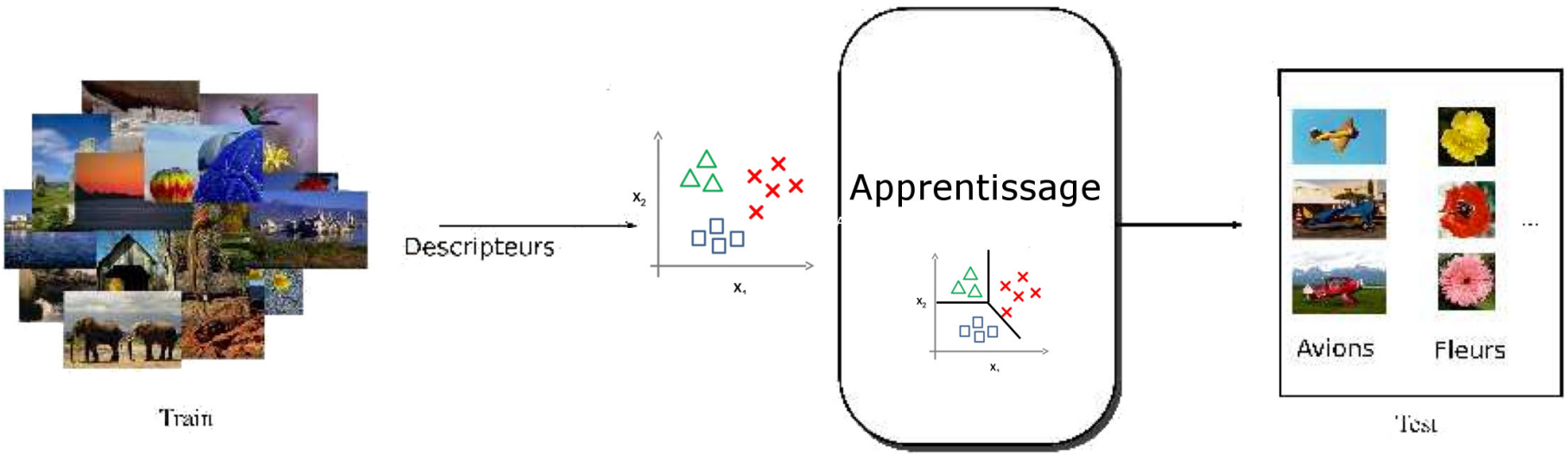
Données et apprentissage

Les données dans les problèmes d'apprentissage

Les données d'un problème d'apprentissage sont généralement vu comme des points/vecteurs dans des espaces possiblement de grandes dimensions.

Apprendre : un problème de recherche de frontière

Les problèmes d'apprentissage consiste à partitionner l'espace en zone associé à une classe particulière. L'algorithme d'apprentissage à pour but la recherche de frontière entre les différentes zones de l'espace.



Retour sur la notion de distance euclidienne

En 1D

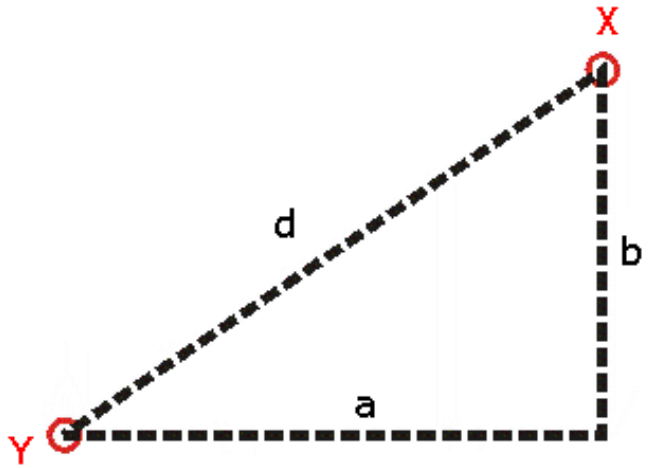
La distance entre les points X et Y correspond à la valeur d :

$$d(X, Y) = a = |X_1 - Y_1| = \sqrt{(X_1 - Y_1)^2}$$

En 2D

La distance entre les poins X et Y correspond à la valeur d. En utilisant le théorème de Pythagore, on a :

$$d = \sqrt{a^2 + b^2} = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2}$$

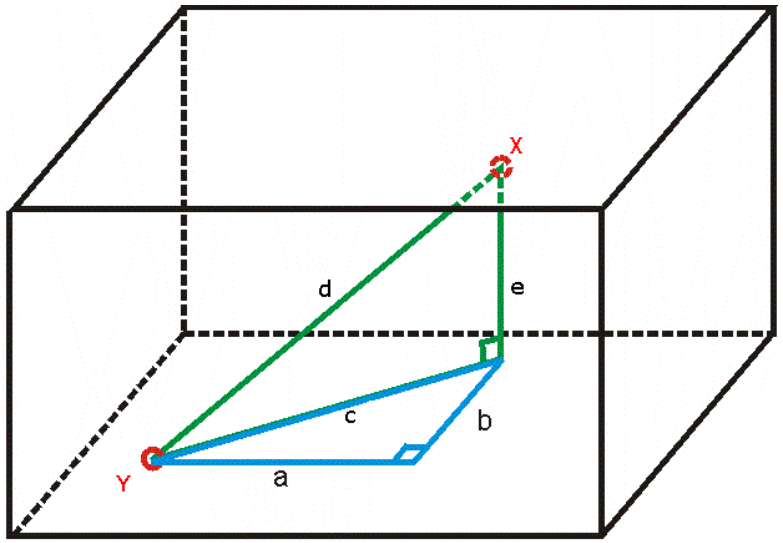


Retour sur la notion de distance euclidienne

En 3D

La distance entre les poins X et Y correspond à la valeur d. En utilisant le théorème de Pythagore, on a :

$$\begin{aligned} d &= \sqrt{e^2 + c^2} \\ &= \sqrt{e^2 + a^2 + b^2} \\ &= \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + (X_3 - Y_3)^2} \end{aligned}$$



En nD

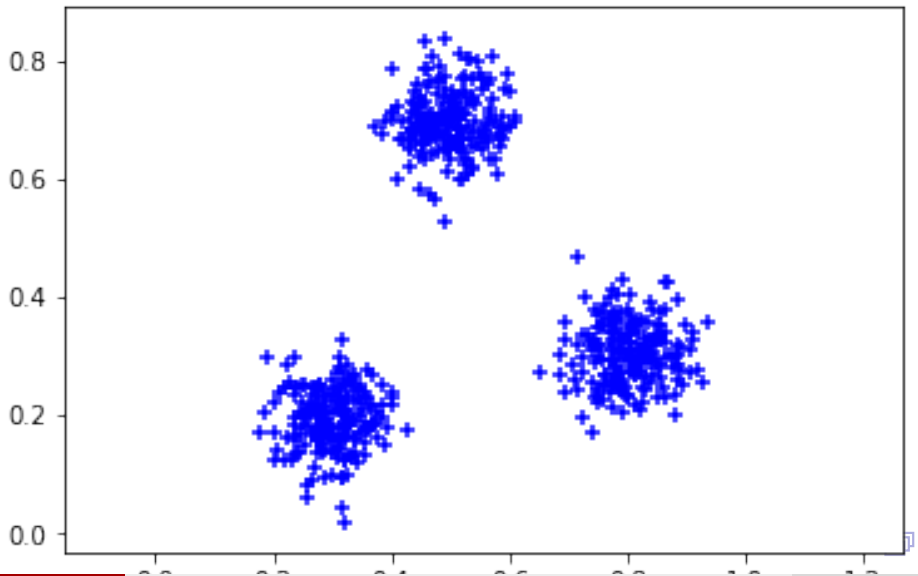
La distance euclidienne se généralise à n'importe quelle dimension. Ainsi la distance euclidienne entre X et Y des vecteurs de taille n est :

$$\sqrt{\sum_{i=1}^{n}(X_i - Y_i)^2}$$

L'algorithme des K-Moyennes (Kmeans)

Objectif

- L'objectif est de partitionner l'espace en k sous-espace. k étant fixé par l'utilisateur au préalable.
- On souhaite minimiser la distance entre les points au sein de chaque partition.
- Chaque partition sera défini par son centre. Les points sont associés à leur centre le plus proche.
- Les K-moyennes font partie des problèmes non-supervisé.



L'algorithme des K-Moyennes (Kmeans)

Initialisation de l'algorithme

Entrées de l'algorithme :

- le nombre k de cluster à rechercher,
- n exemples x d'apprentissage,
- le nombre T d'itération de l'algorithme

Initialisation :

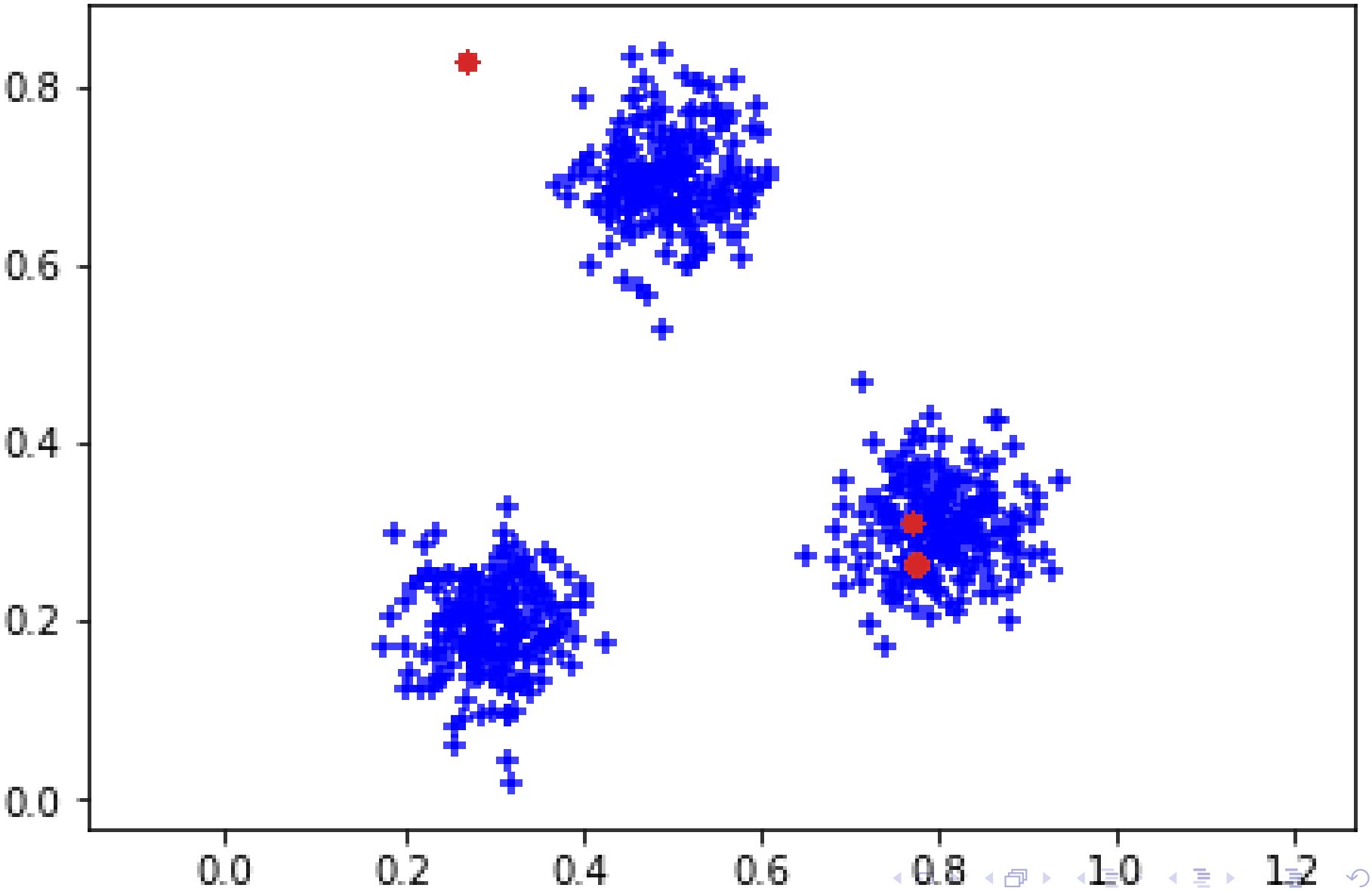
- On initialise k centres de clusters au hasard.

Déroulement de l'algorithme

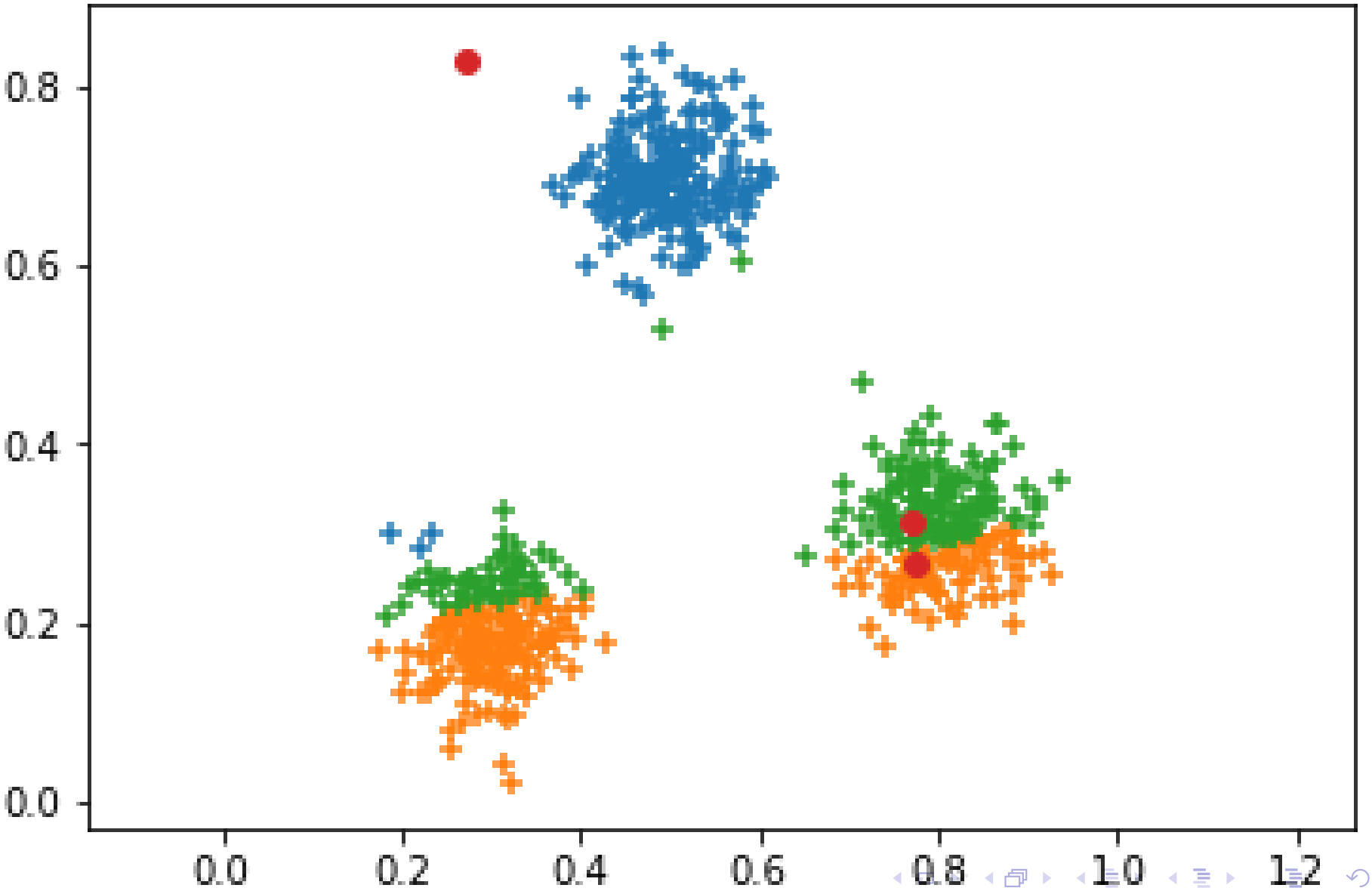
Pour chaque itération t de 0 à T faire :

1. Assigner chaque point d'apprentissage au centre de cluster le plus proche.
2. On met à jours les centres de cluster en calculant la moyenne des points à l'intérieur. Si aucun point n'a été attribué au cluster, on tire un autre centre au hasard.

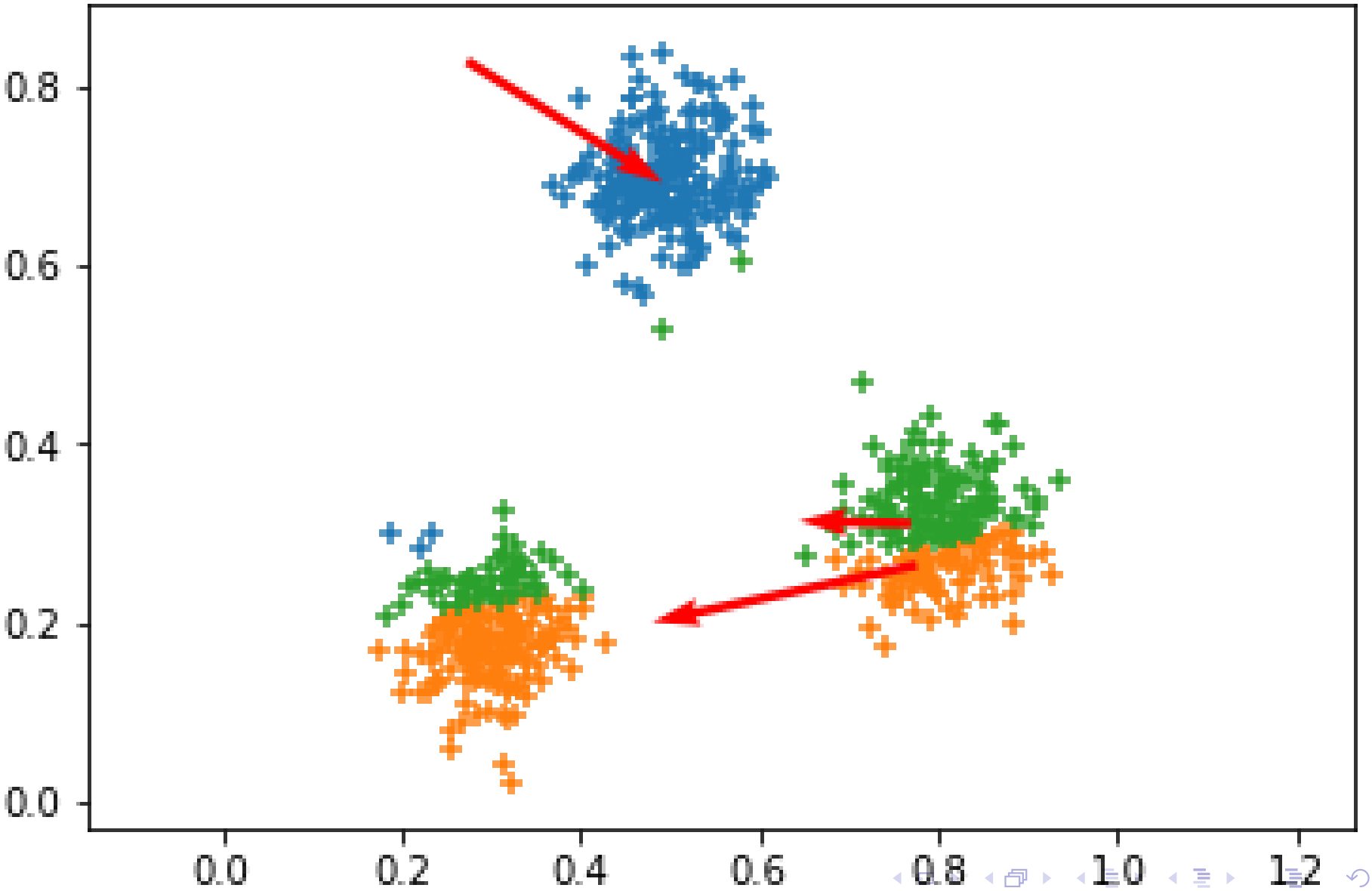
Exemple de déroulement de l'algorithme des K-Moyennes



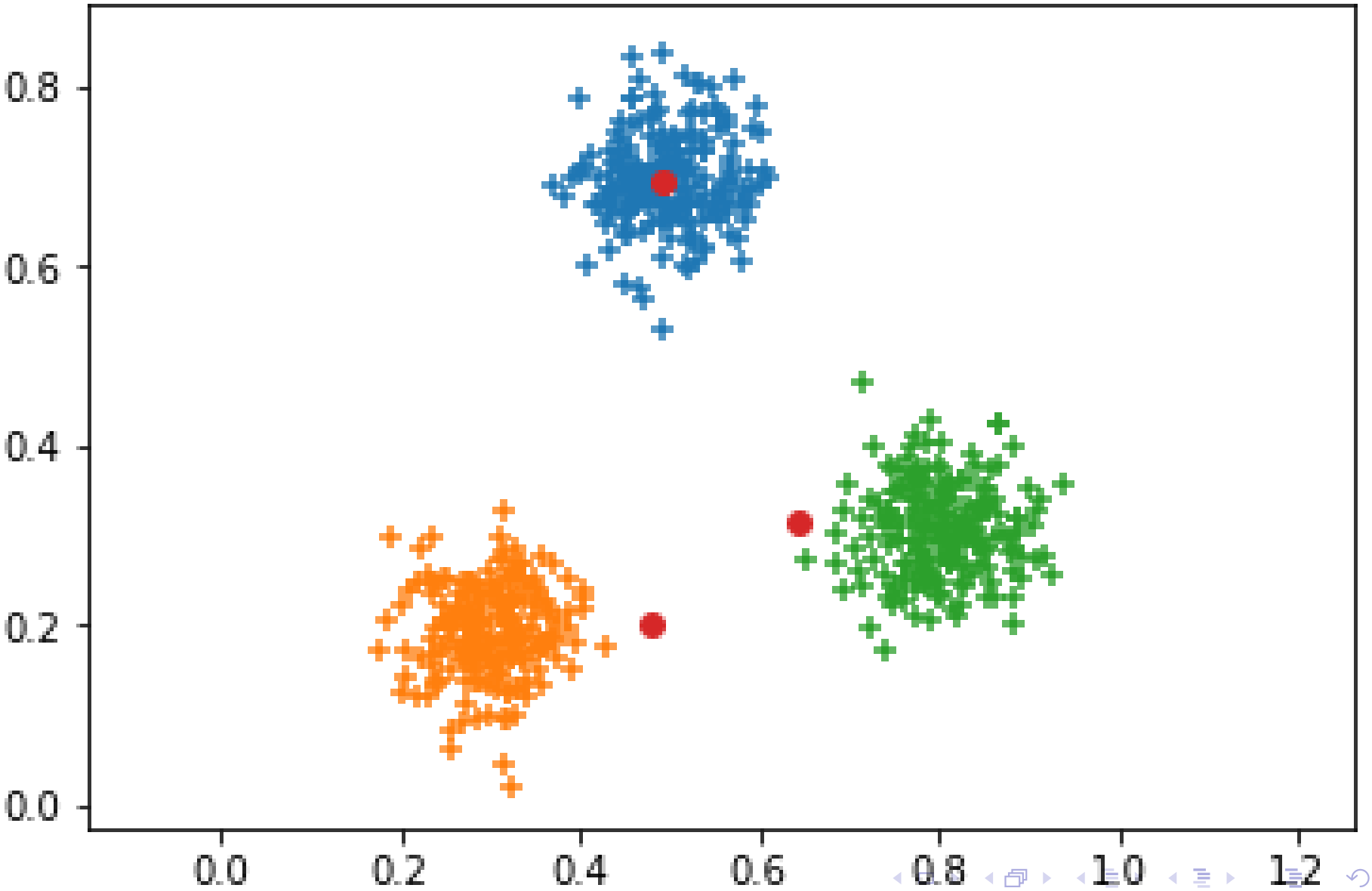
Exemple de déroulement de l'algorithme des K-Moyennes



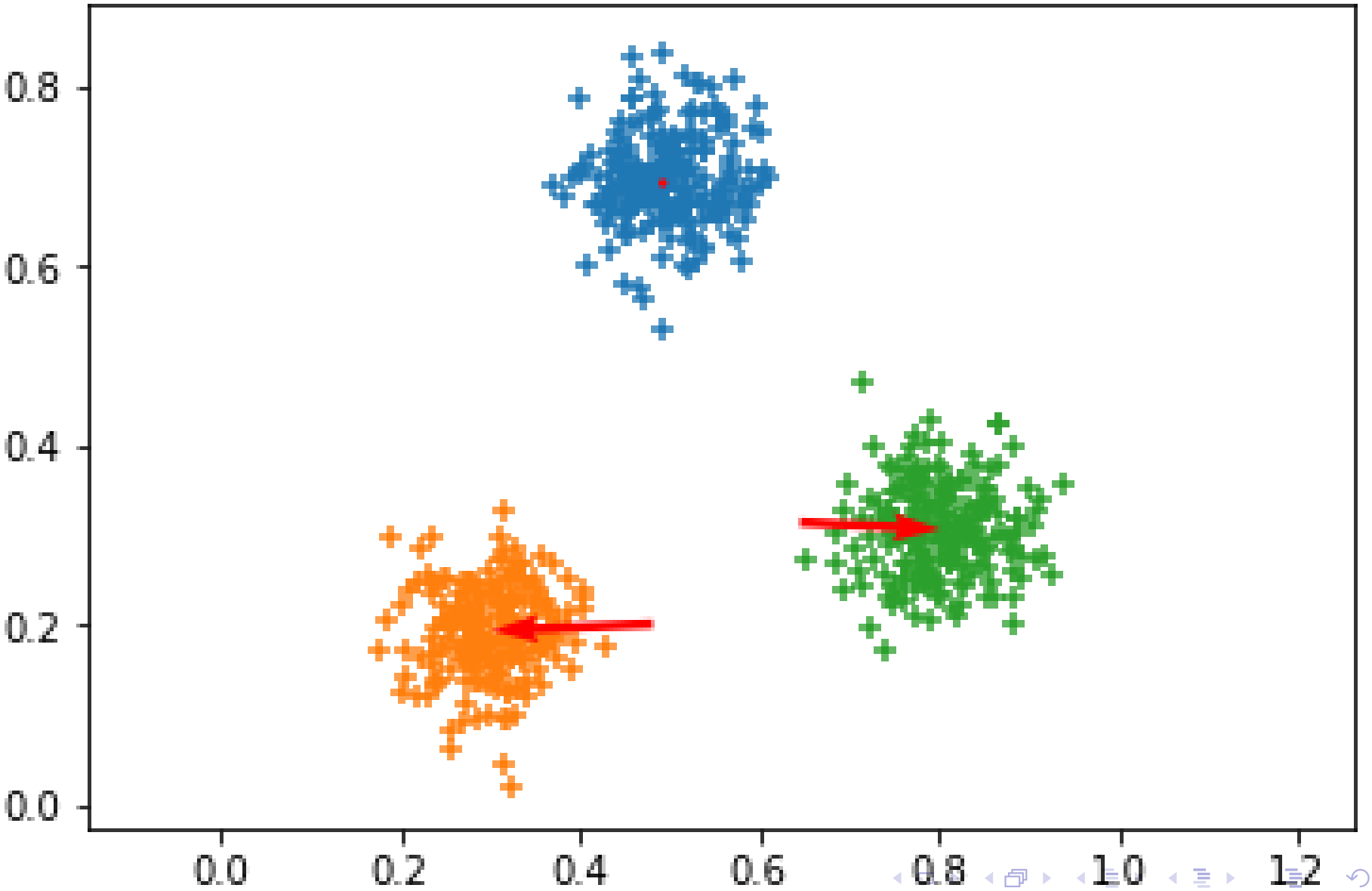
Exemple de déroulement de l'algorithme des K-Moyennes



Exemple de déroulement de l'algorithme des K-Moyennes



Exemple de déroulement de l'algorithme des K-Moyennes



Exemple de déroulement de l'algorithme des K-Moyennes