Introdução a Aprendizagem De Máquina

Pós-graduação em Ciência de Dados e Machine Learning
Módulo 3 - Data Mining e Machine Learning

Professor Msc. Ricardo José Menezes Maia

K Means Clustering

K Means Clustering

Seção 10 do Introduction to Statical Learning de Gareth James

K Means Clustering

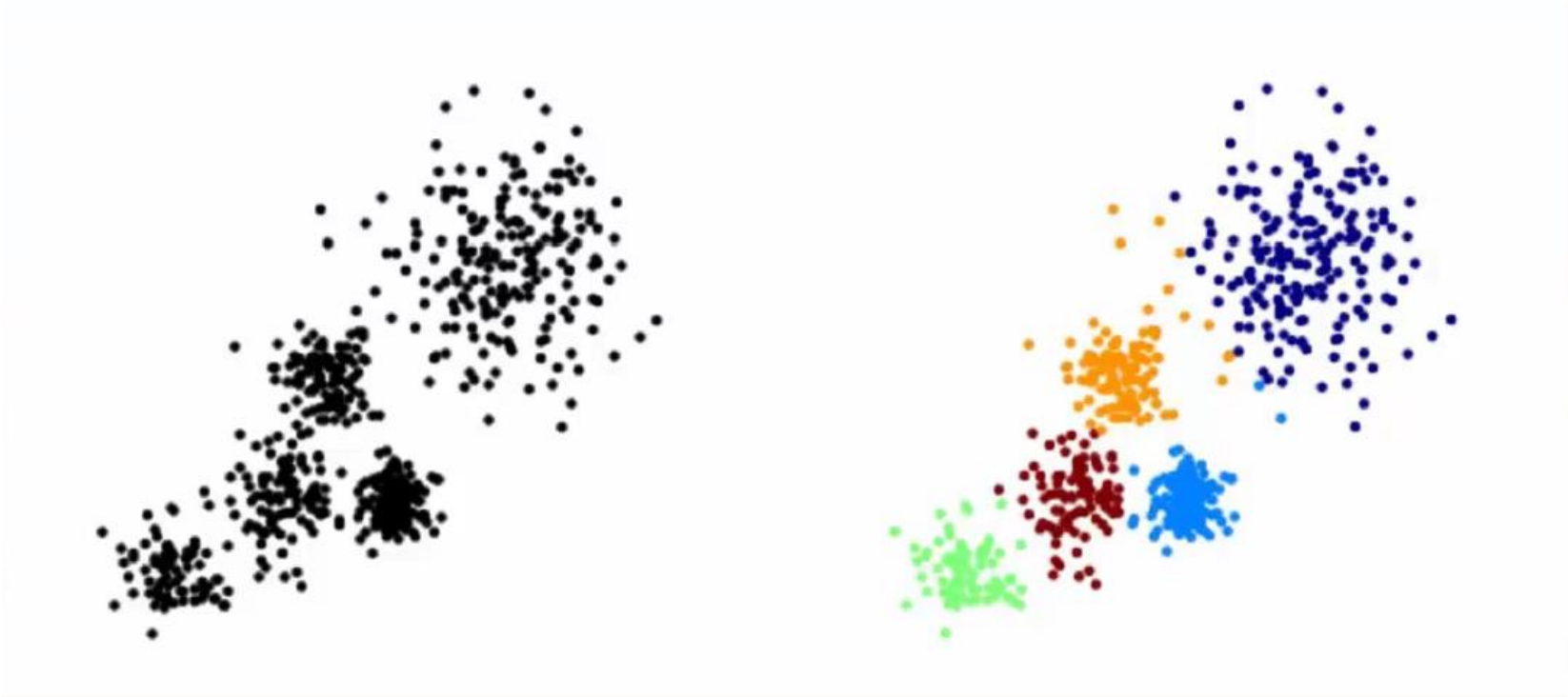
K Means Clustering é um método de Machine Learning baseado em aprendizado não supervisionado que tentará agrupar seus dados em grupos em características similares.

São usados para:

- Agrupamento automático de documentos
- Agrupamento de clientes
- Segmentação de mercado
- Geoestatística

K Means Clustering

O objetivo é dividir os dados em K grupos distintos baseados nos parâmetros

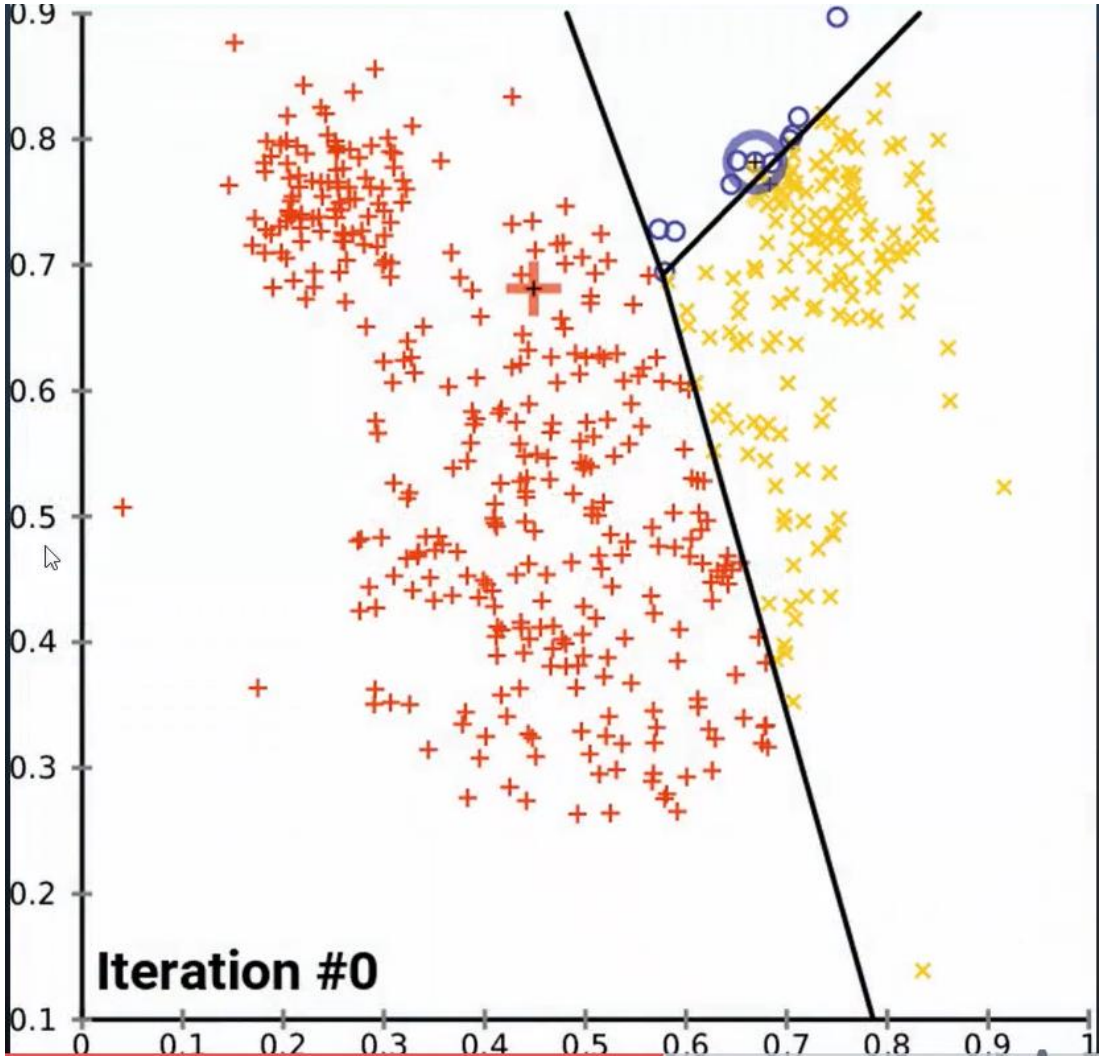


K Means Clustering - Algoritmo

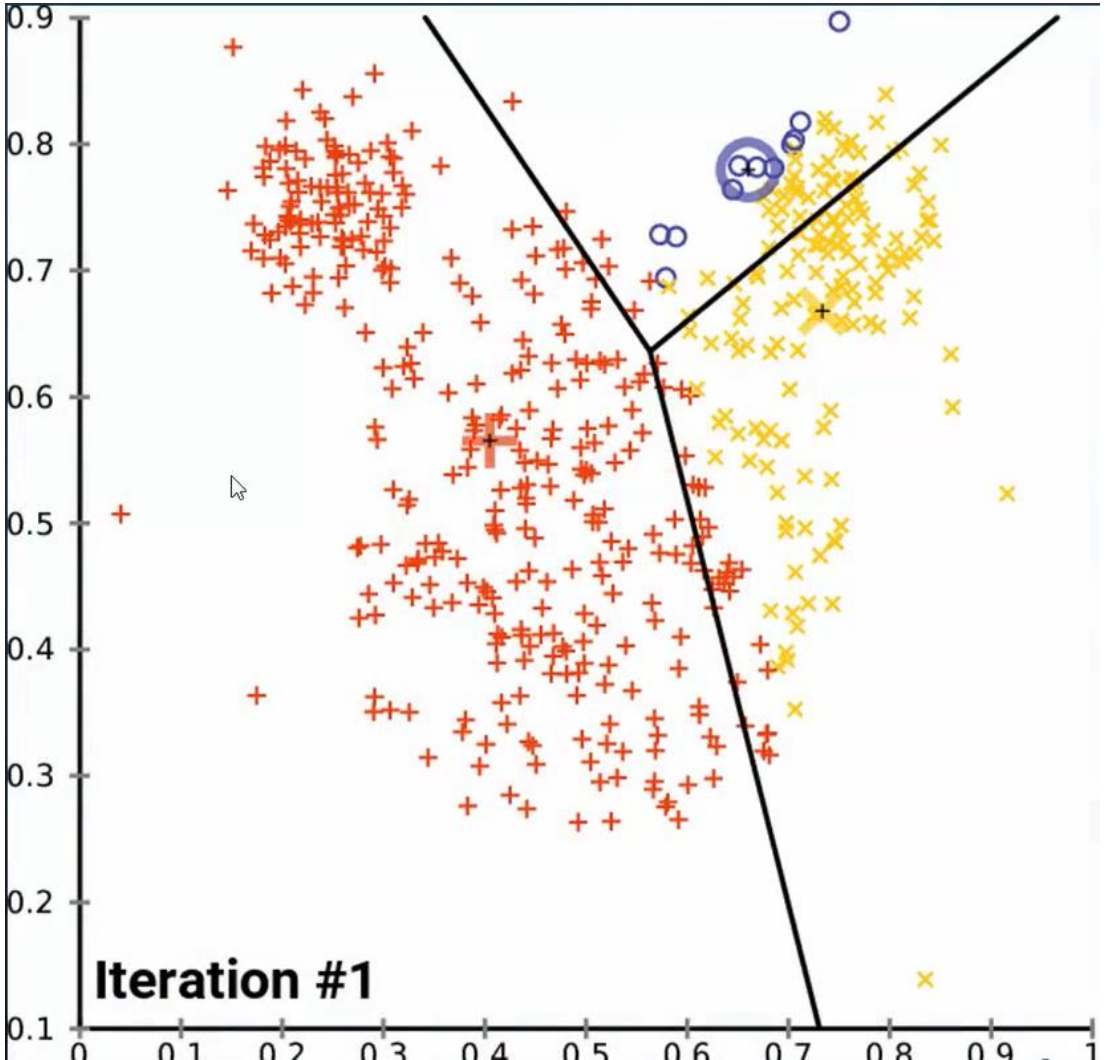
O algoritmo:

- Escolher um número K de grupos (clusters)
- Aleatoriamente definir uma classe para todos os pontos
- Até os clusters pararem de mudar faça:
  - Para cada cluster, obtenha o centroide do mesmo calculando a média dos vetores dos pontos do cluster
  - Defina cada ponto ao cluster na qual o centroide é o mais próximo

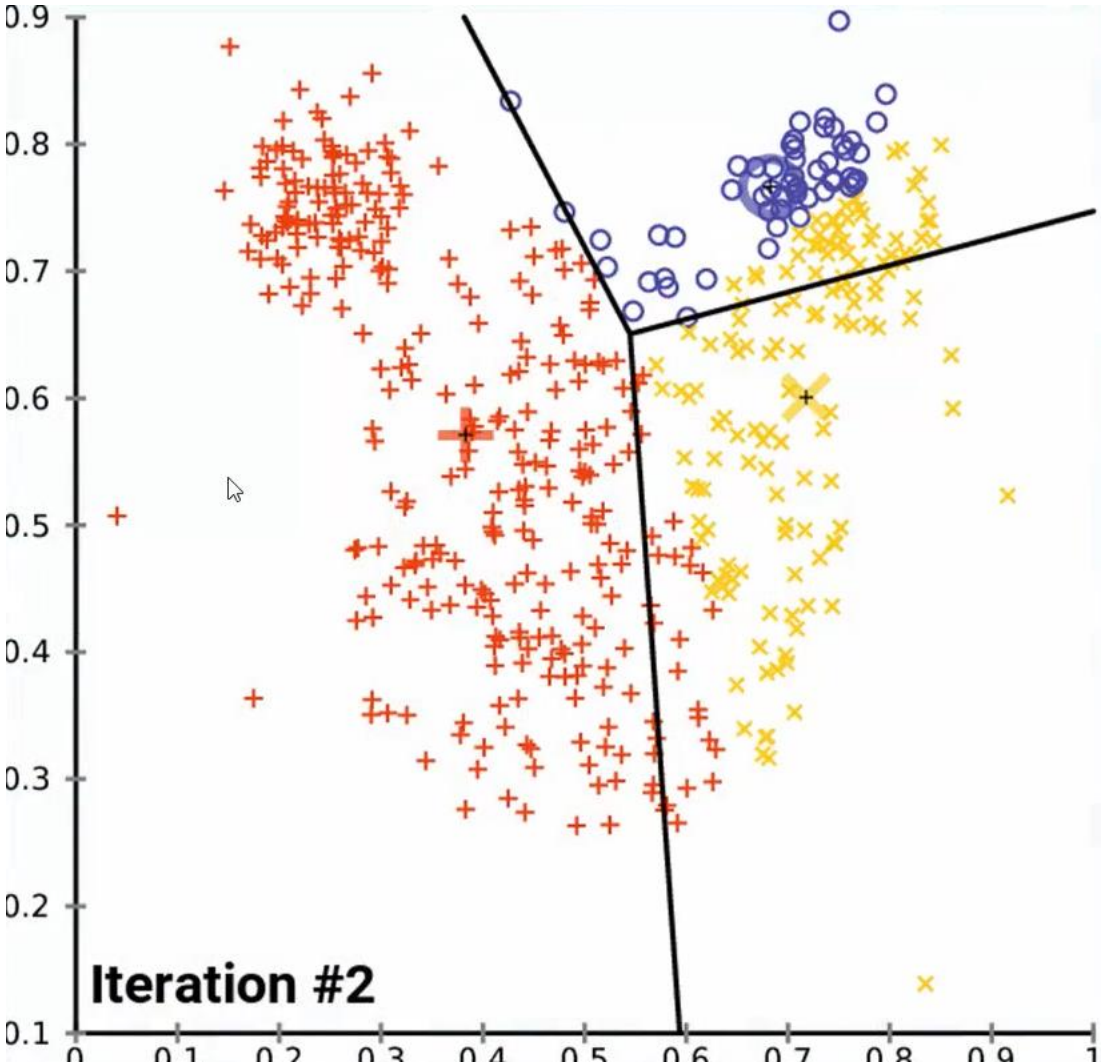
K Means Clustering - Algoritmo



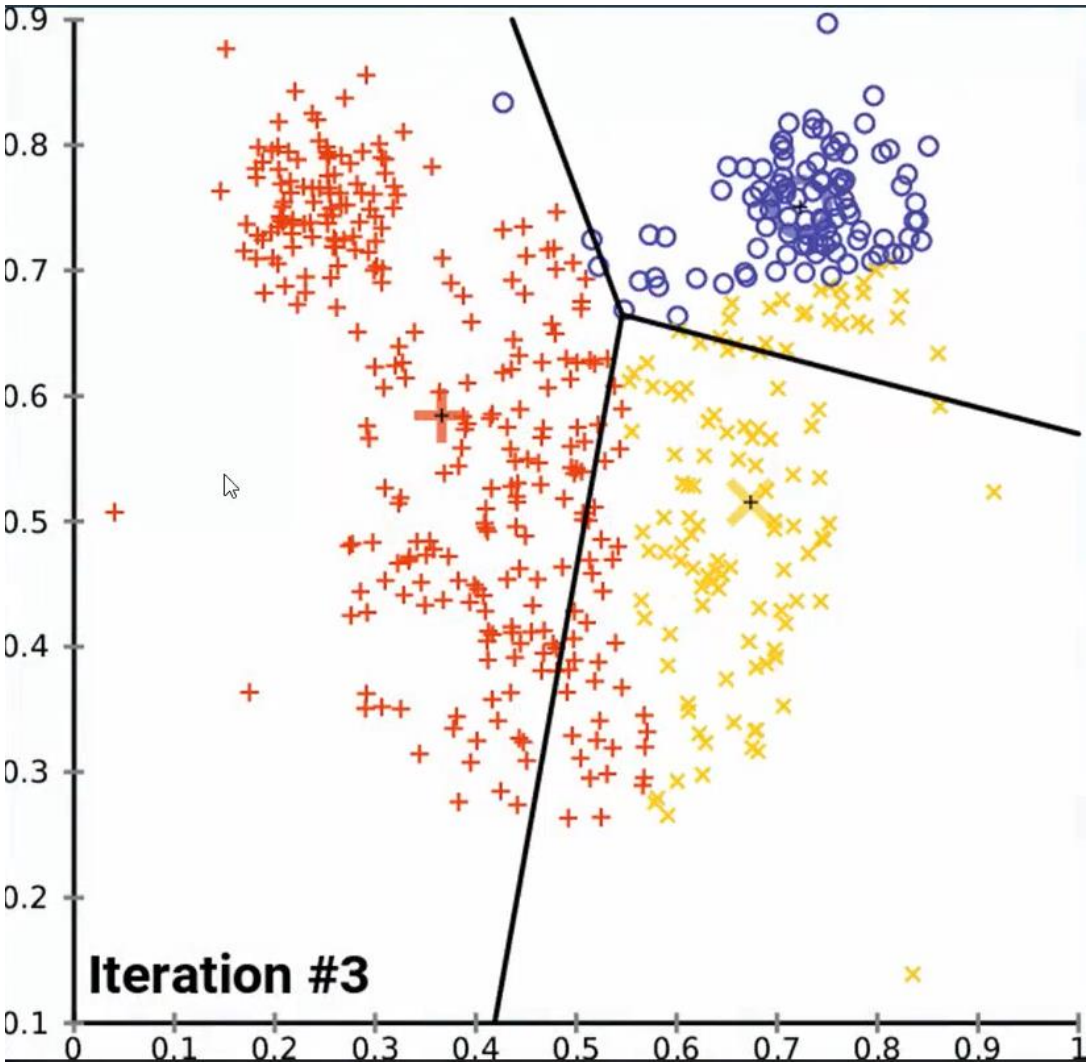
K Means Clustering - Algoritmo



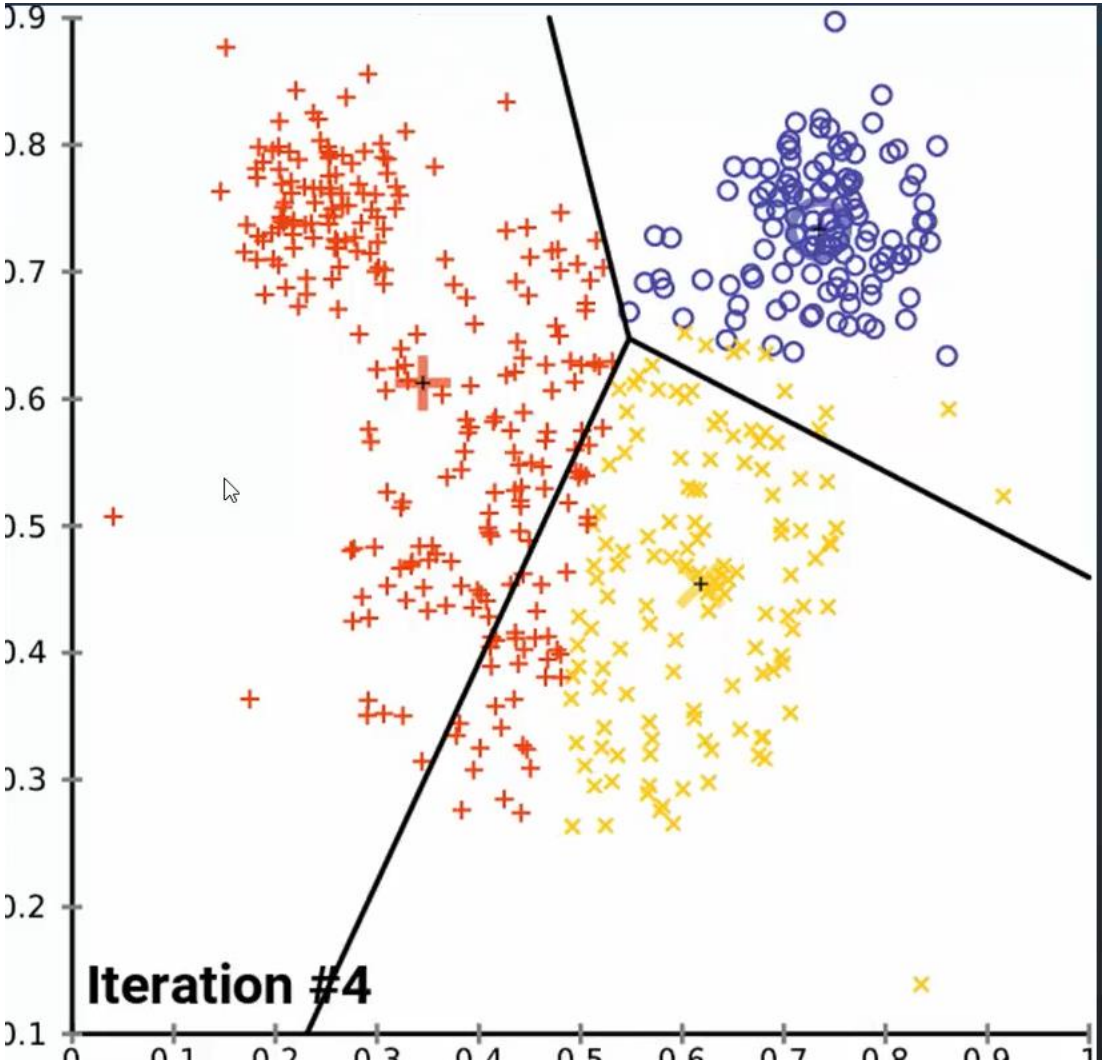
K Means Clustering - Algoritmo



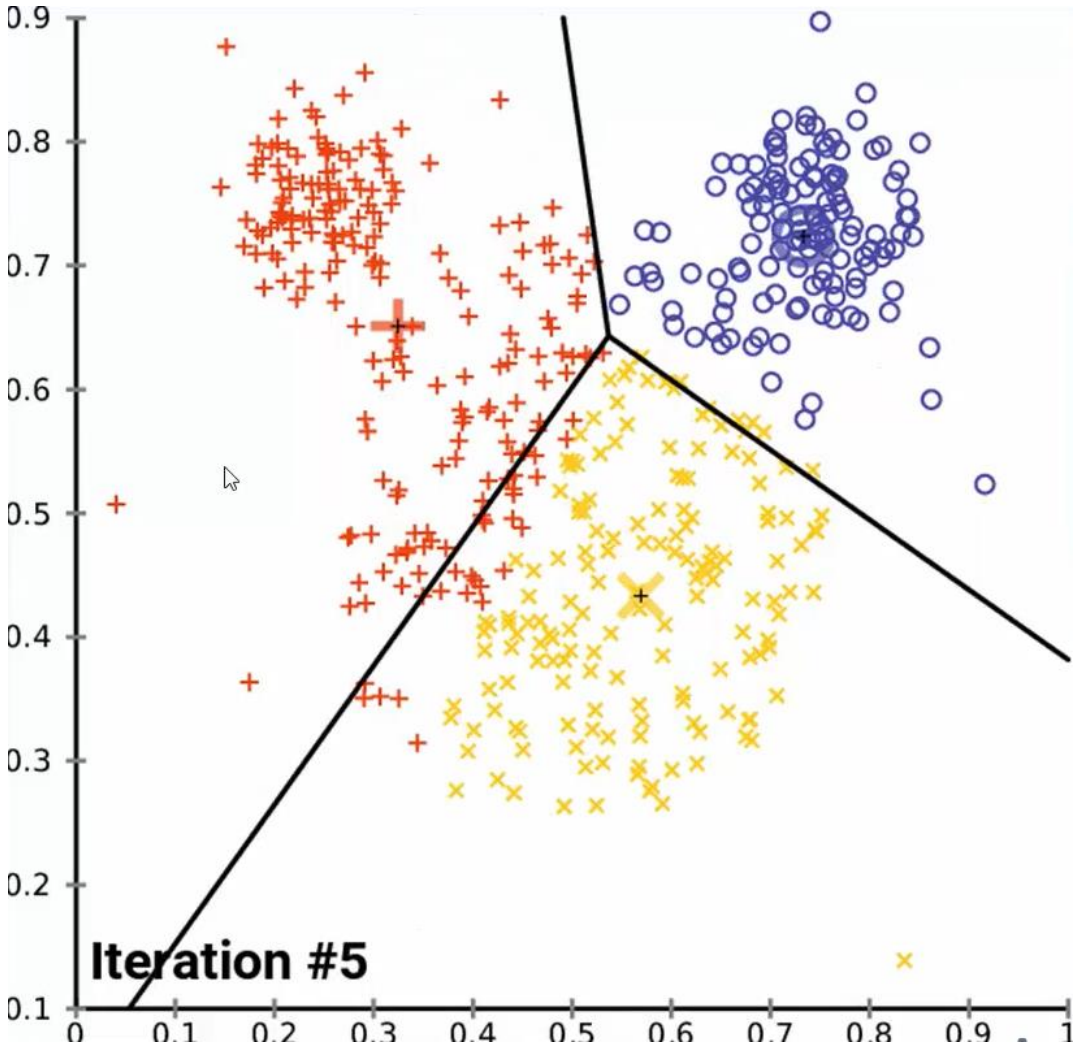
K Means Clustering - Algoritmo



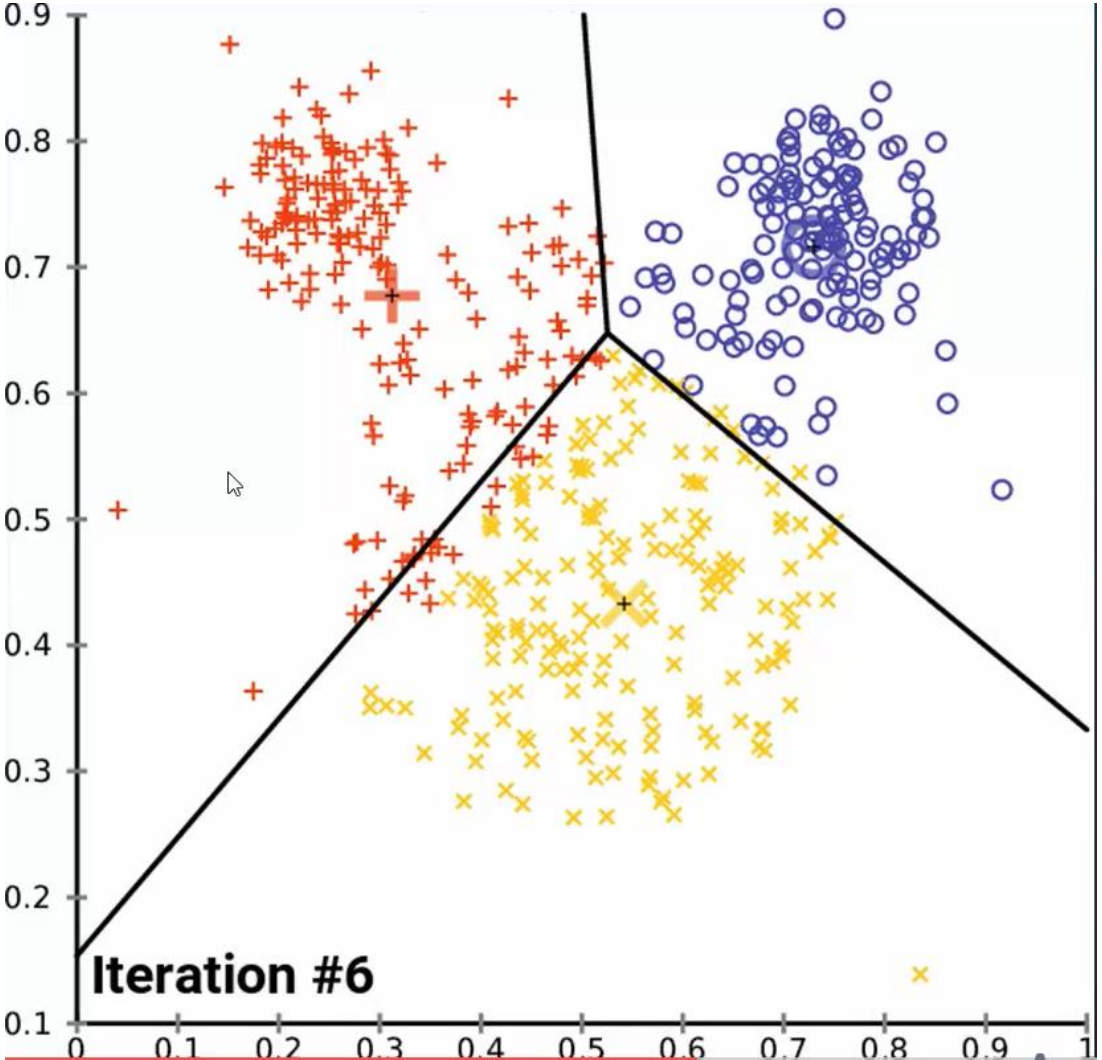
K Means Clustering - Algoritmo



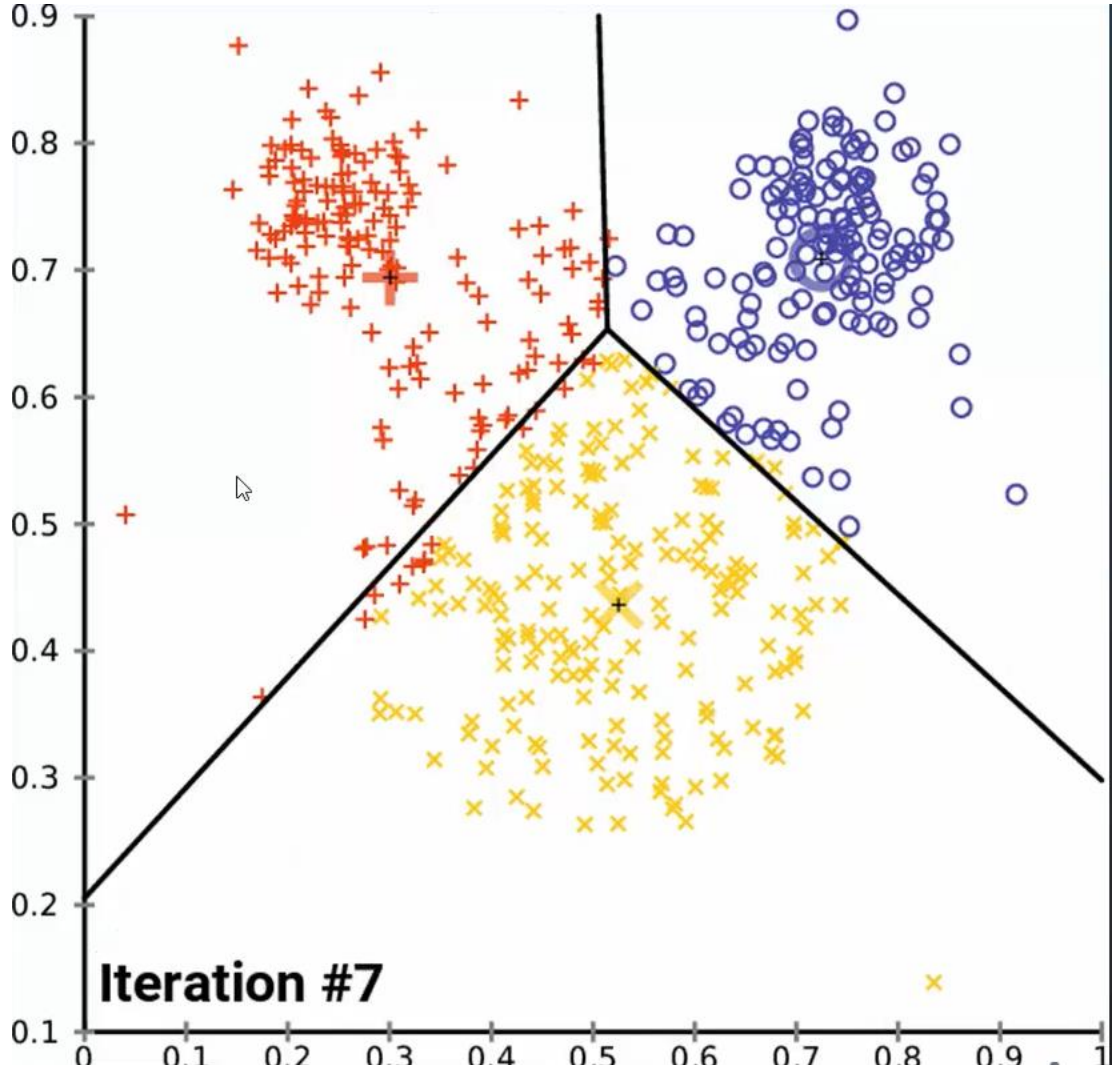
K Means Clustering - Algoritmo



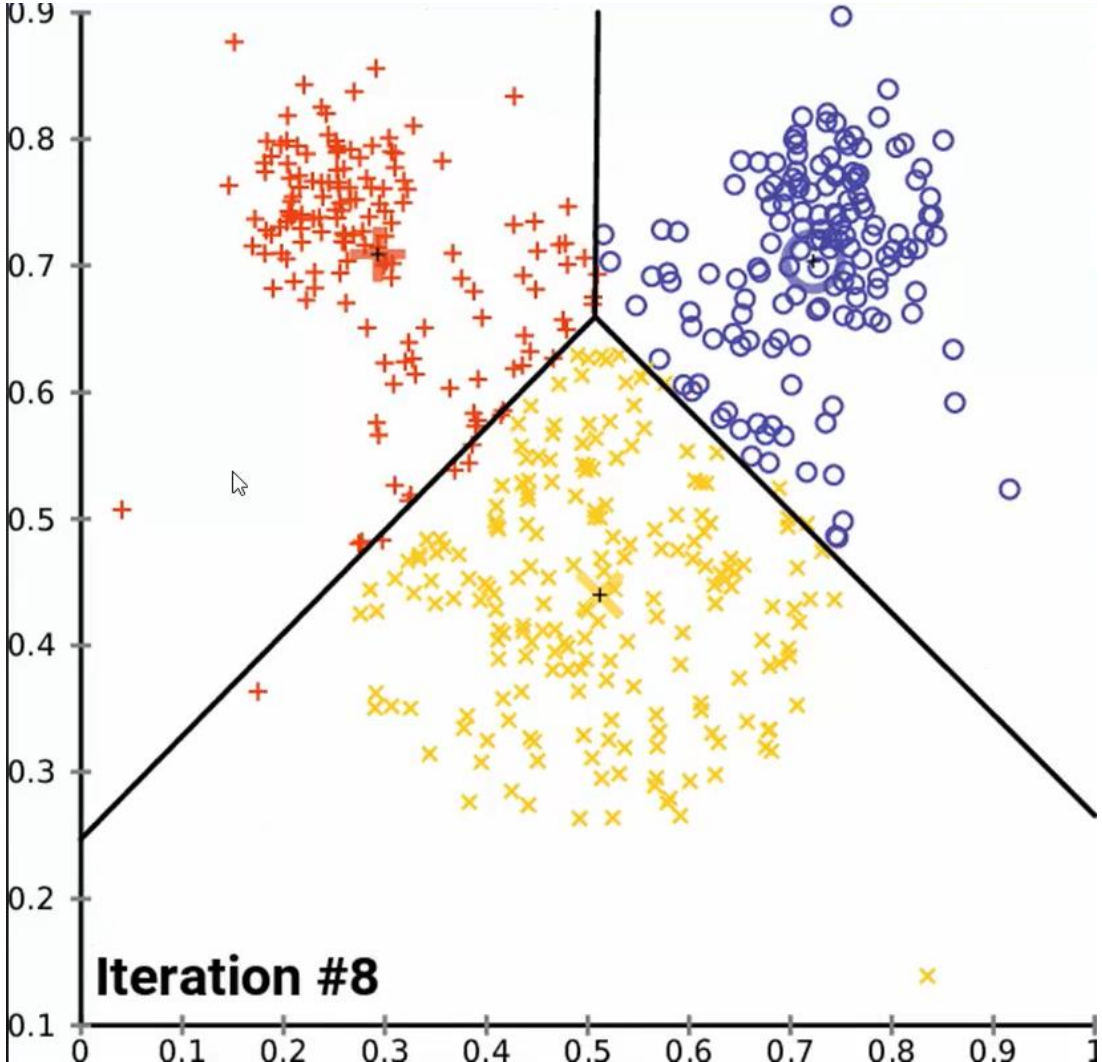
K Means Clustering - Algoritmo



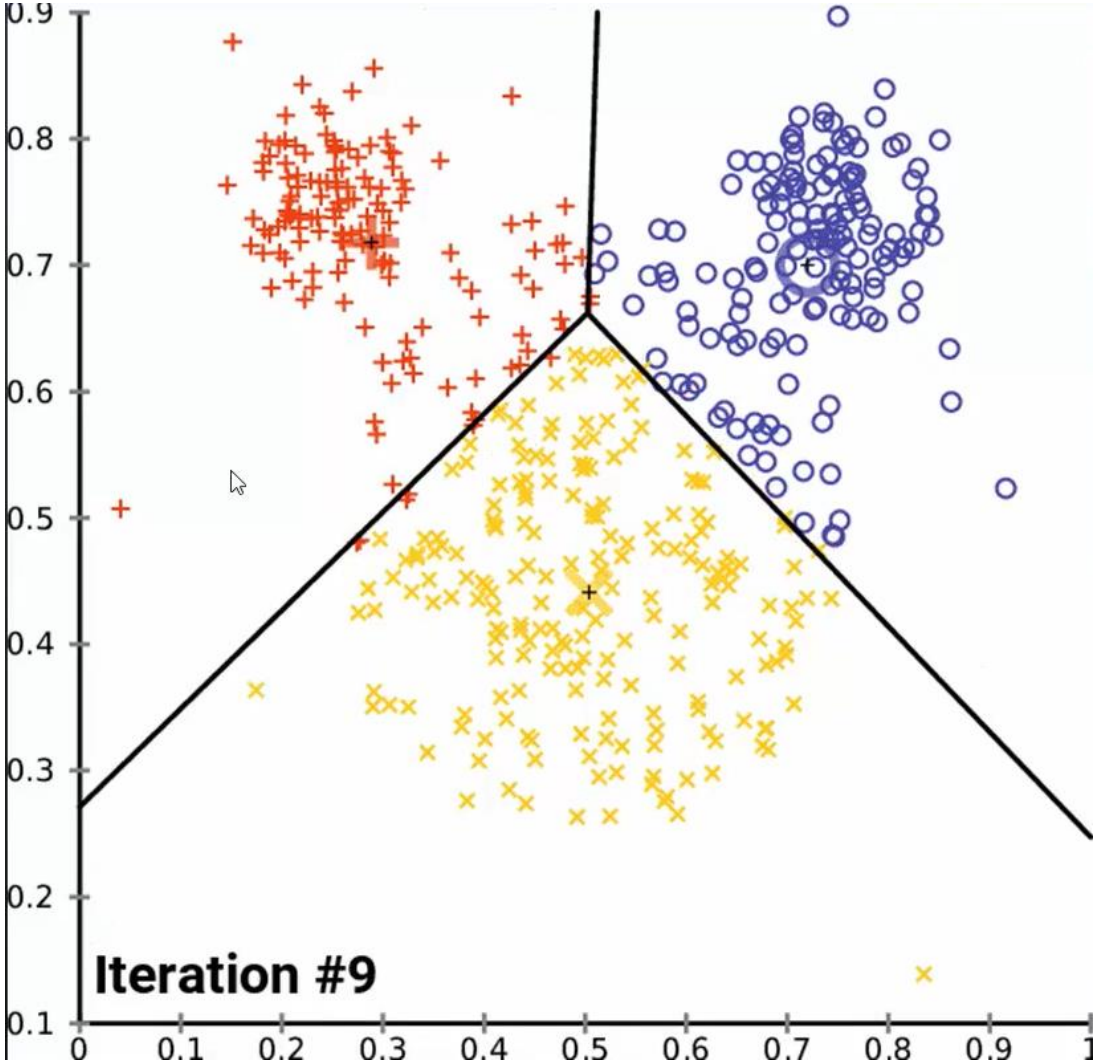
K Means Clustering - Algoritmo



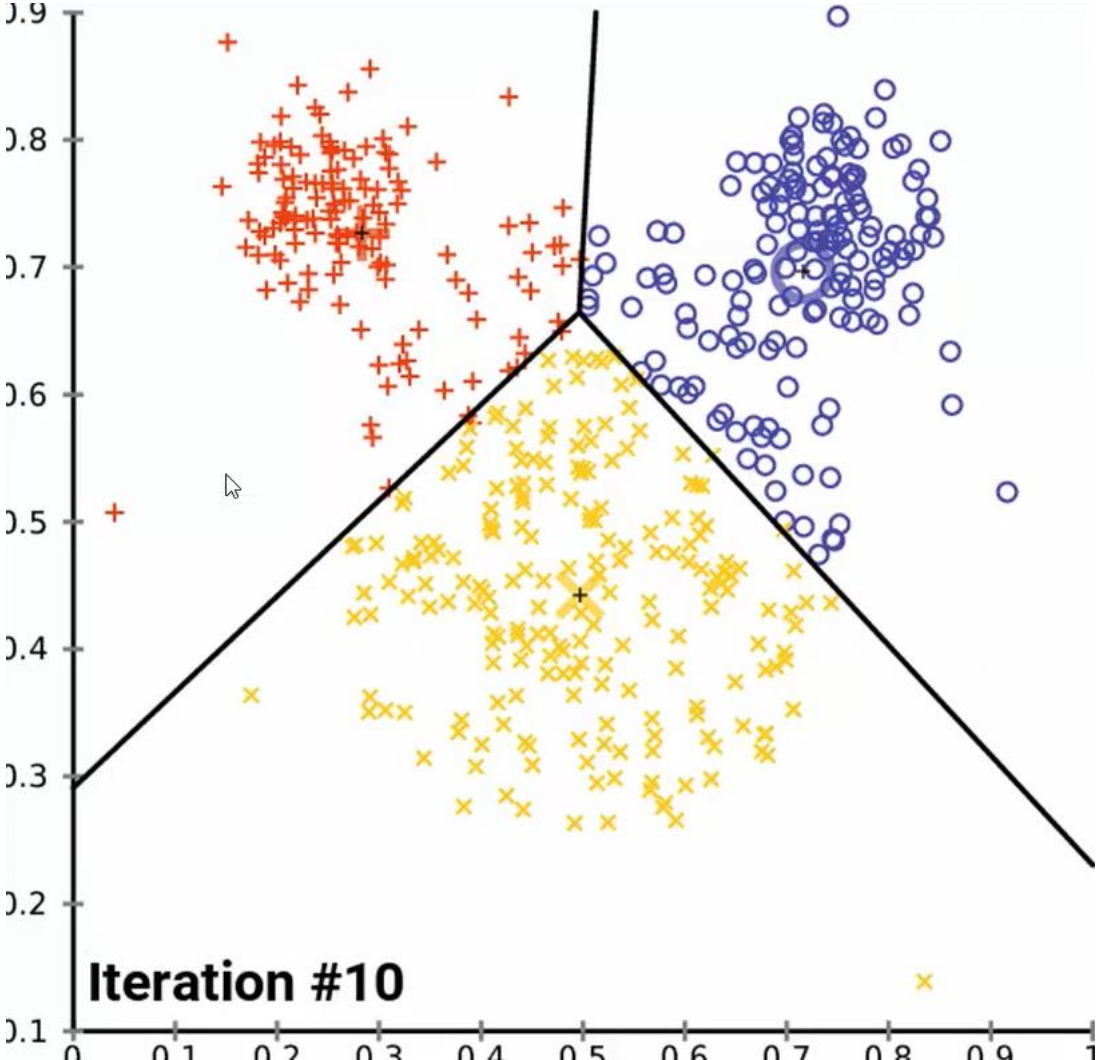
K Means Clustering - Algoritmo



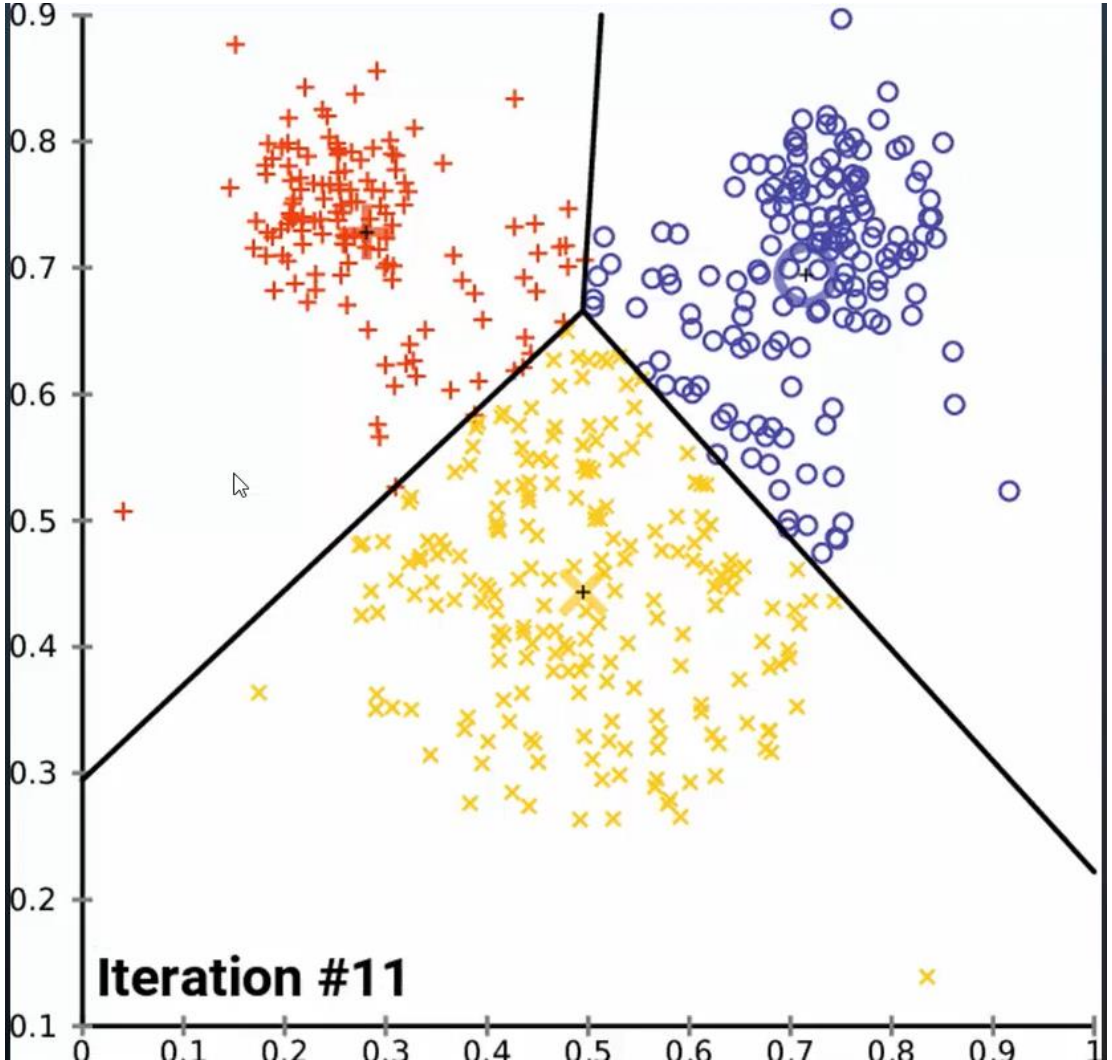
K Means Clustering - Algoritmo



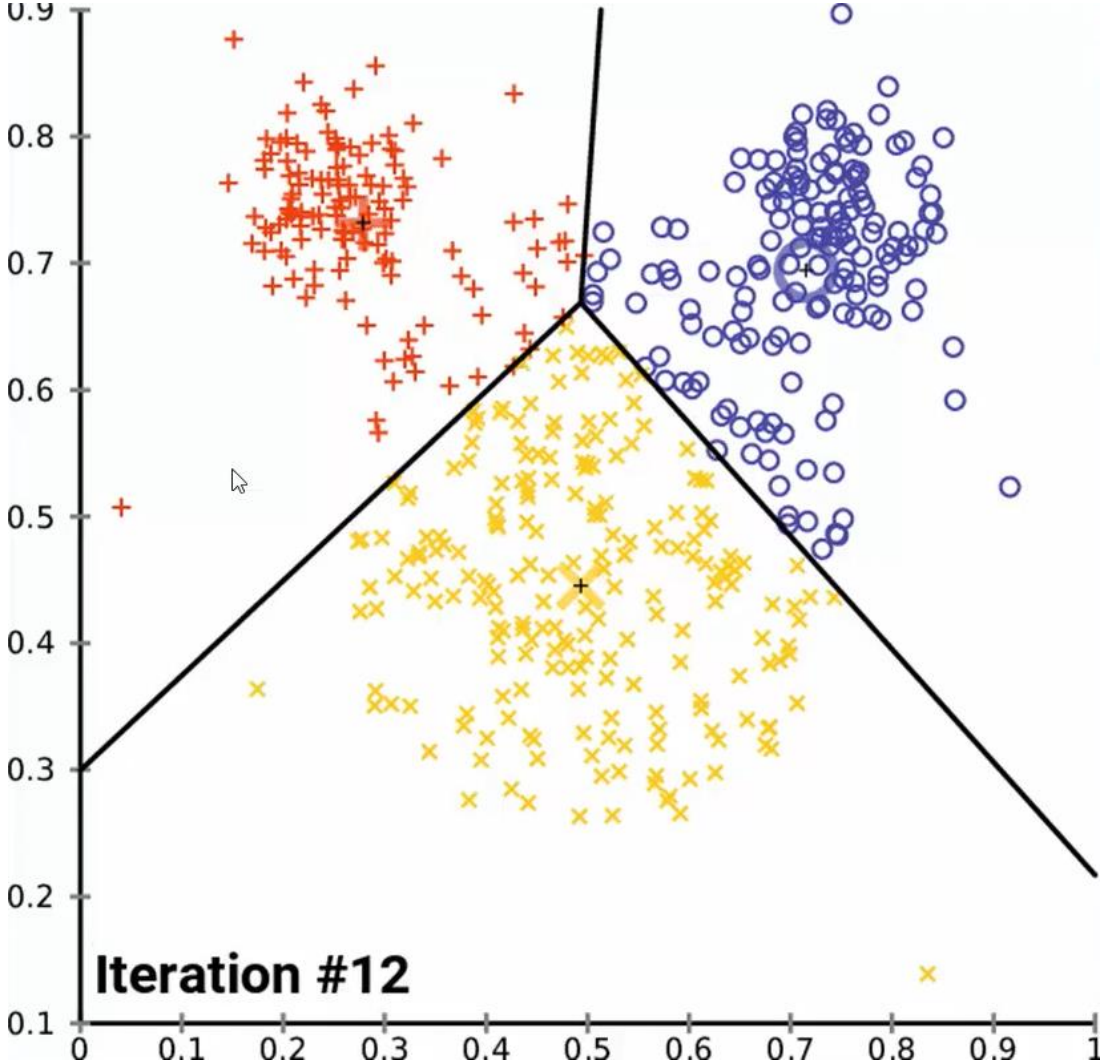
K Means Clustering - Algoritmo



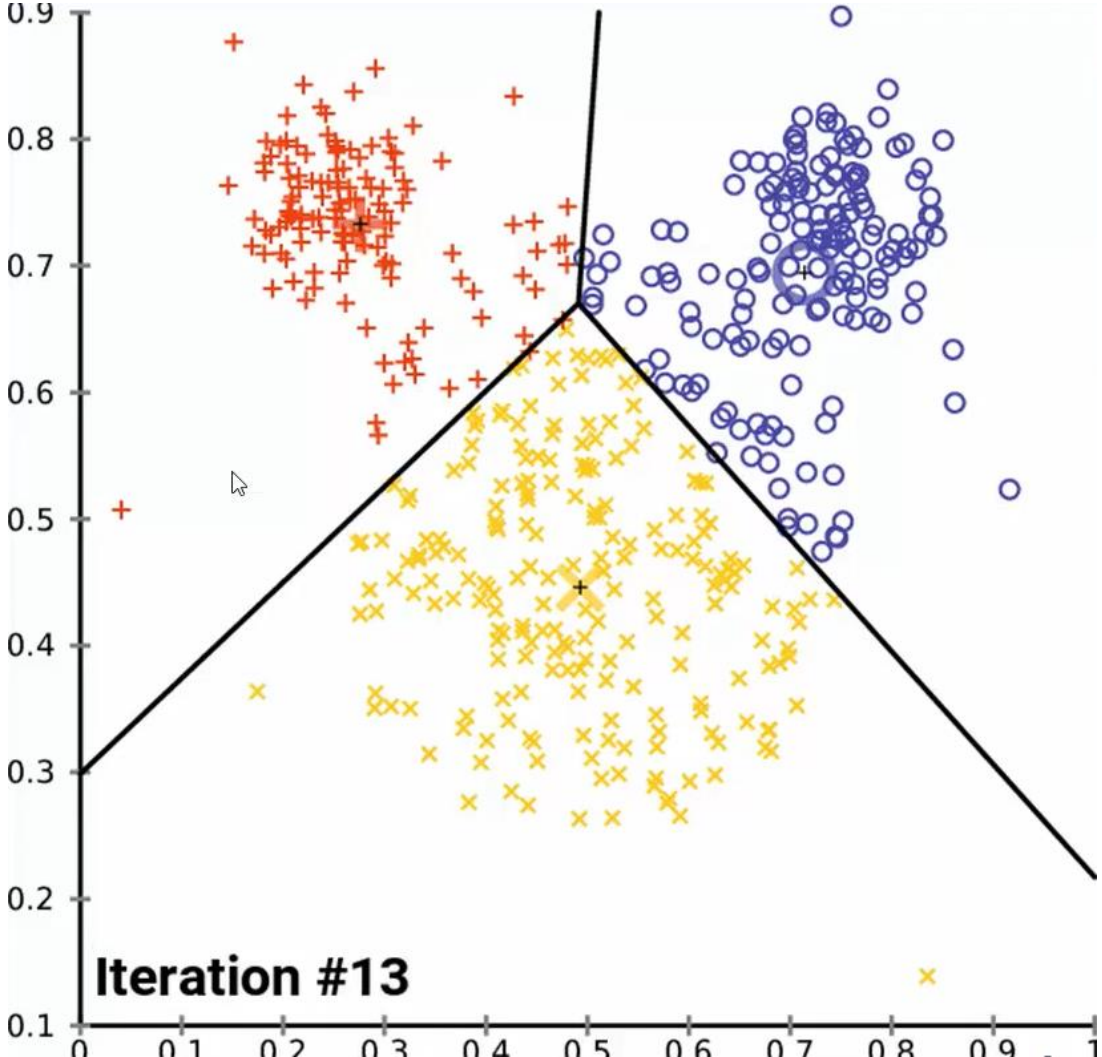
K Means Clustering - Algoritmo



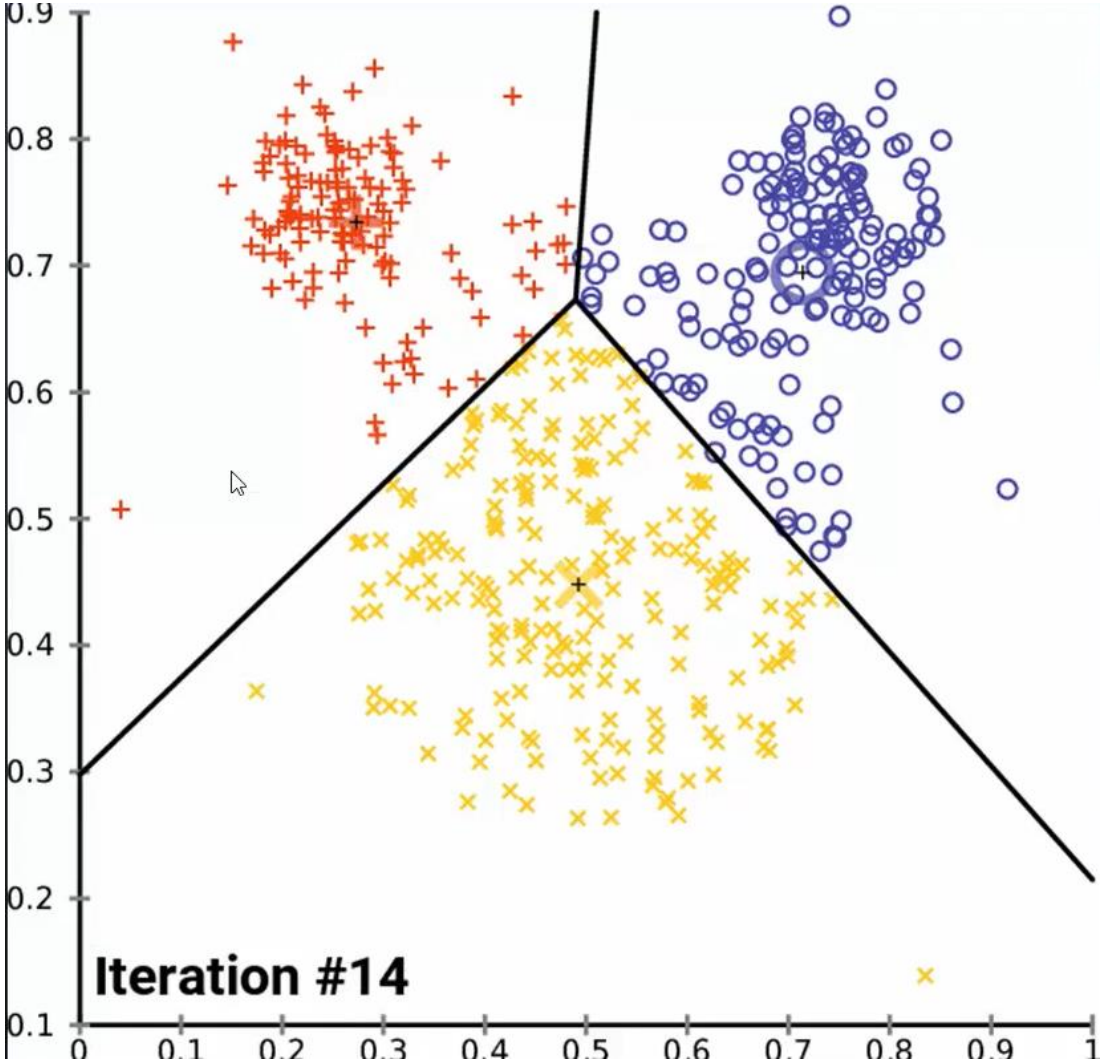
K Means Clustering - Algoritmo



K Means Clustering - Algoritmo

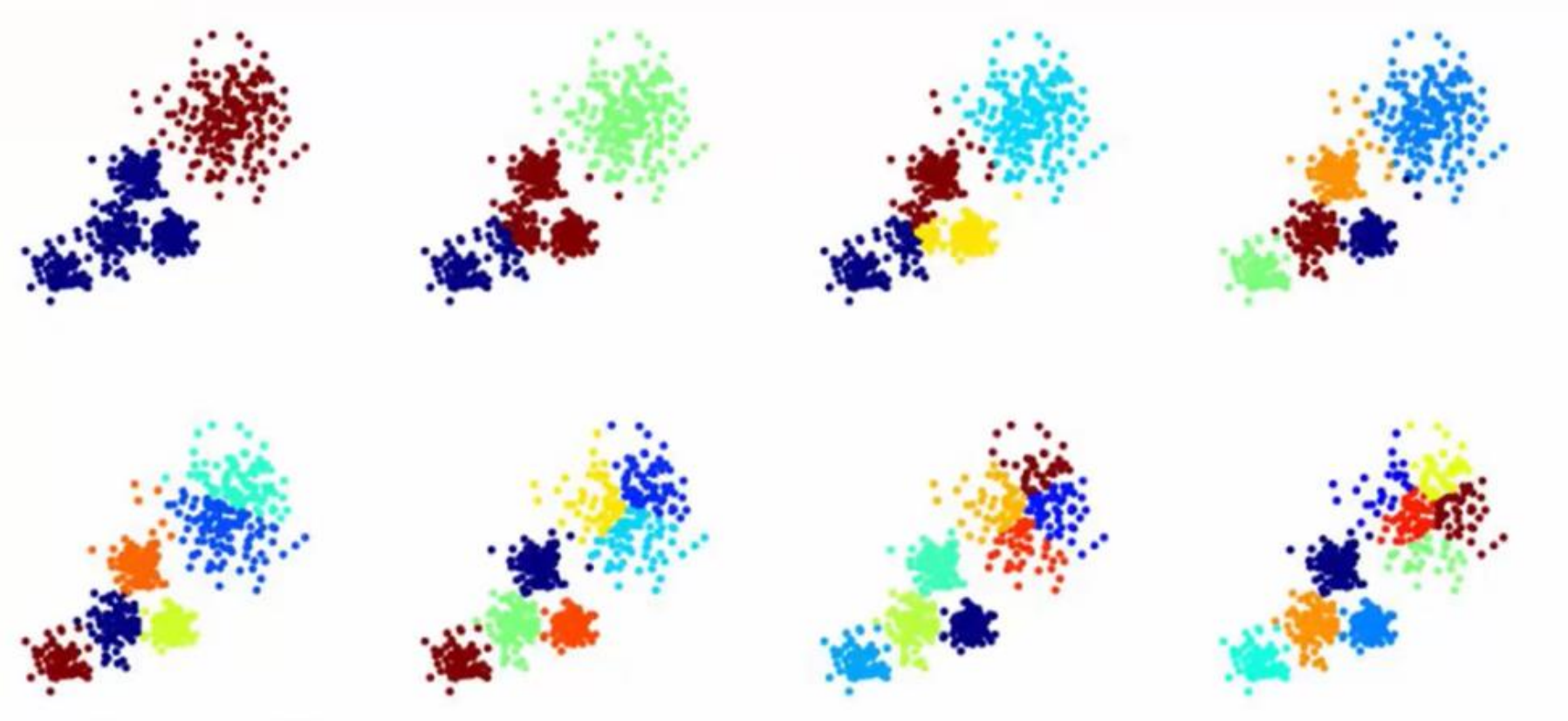


K Means Clustering - Algoritmo



K Means Clustering - Algoritmo

Escolhendo um valor para K



K Means Clustering - Algoritmo

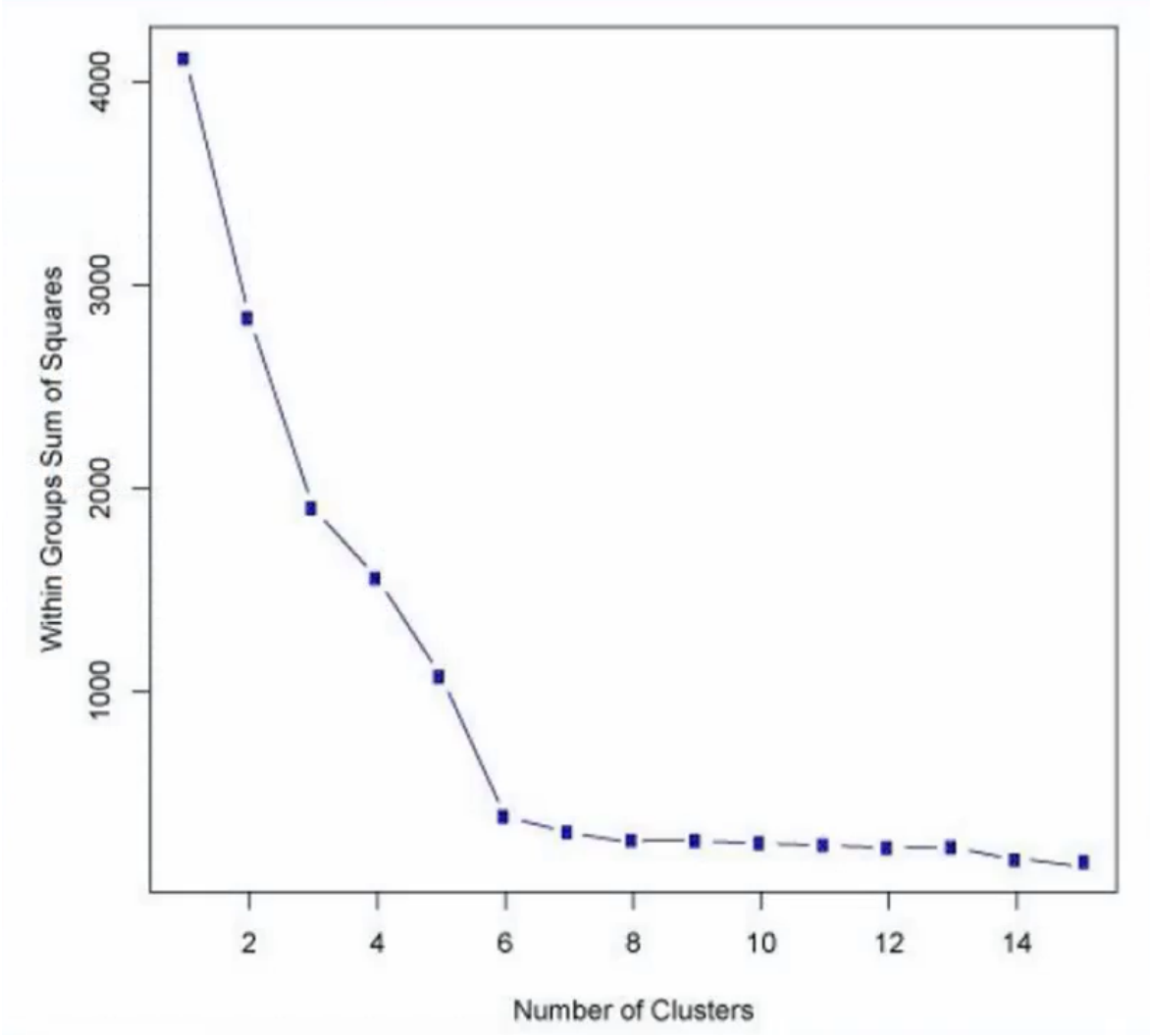
- Não existe uma resposta fácil para escolher o melhor valor para K.
- Uma forma é o método do cotovelo.

- Primeiro calcule a soma dos erros quadrados (SEQ) para alguns valores de K (por exemplo 2,4,6,...)
- A soma dos quadrados dos erros é definido como o quadrado das distâncias entre cada membro e seu centroide.

K Means Clustering - Algoritmo

- Se você plotar K versus SEQ você verá o erro diminuir a medida em que K aumenta
- A idéia do método do cotovelo é escolher um valor de K na qual o SEQ caia abruptamente.
- Isso produz um “efeito cotovelo” no gráfico, como você pode ver a seguir.

K Means Clustering - Algoritmo



K Means Clustering - Algoritmo

- Usarmos o scikit-learn para criar alguns clusters e testá-los usando o K-Means.