

Aprendizaje no supervisado K-Medias

Pontificia Universidad Javeriana

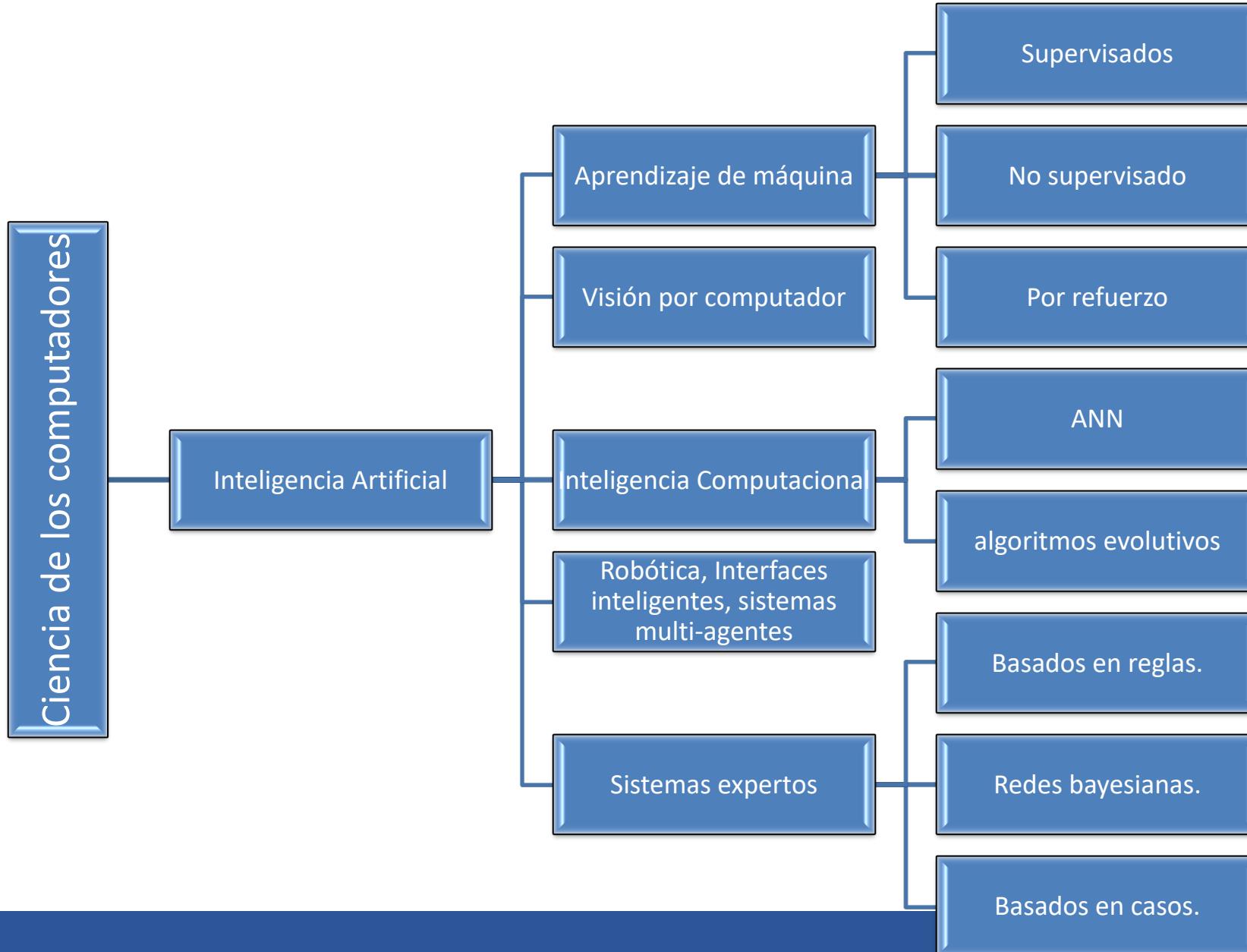
Francisco Carlos Calderon Ph.D

2020

Objetivos

Reconocer un problema de clasificación no supervisado.

Usar la K-means como un método de clasificación no supervisado



Aprendizaje no supervisado K-Medias

Facultad de ingeniería
Deptº de electrónica

Aprendizaje de máquina, clasificación general

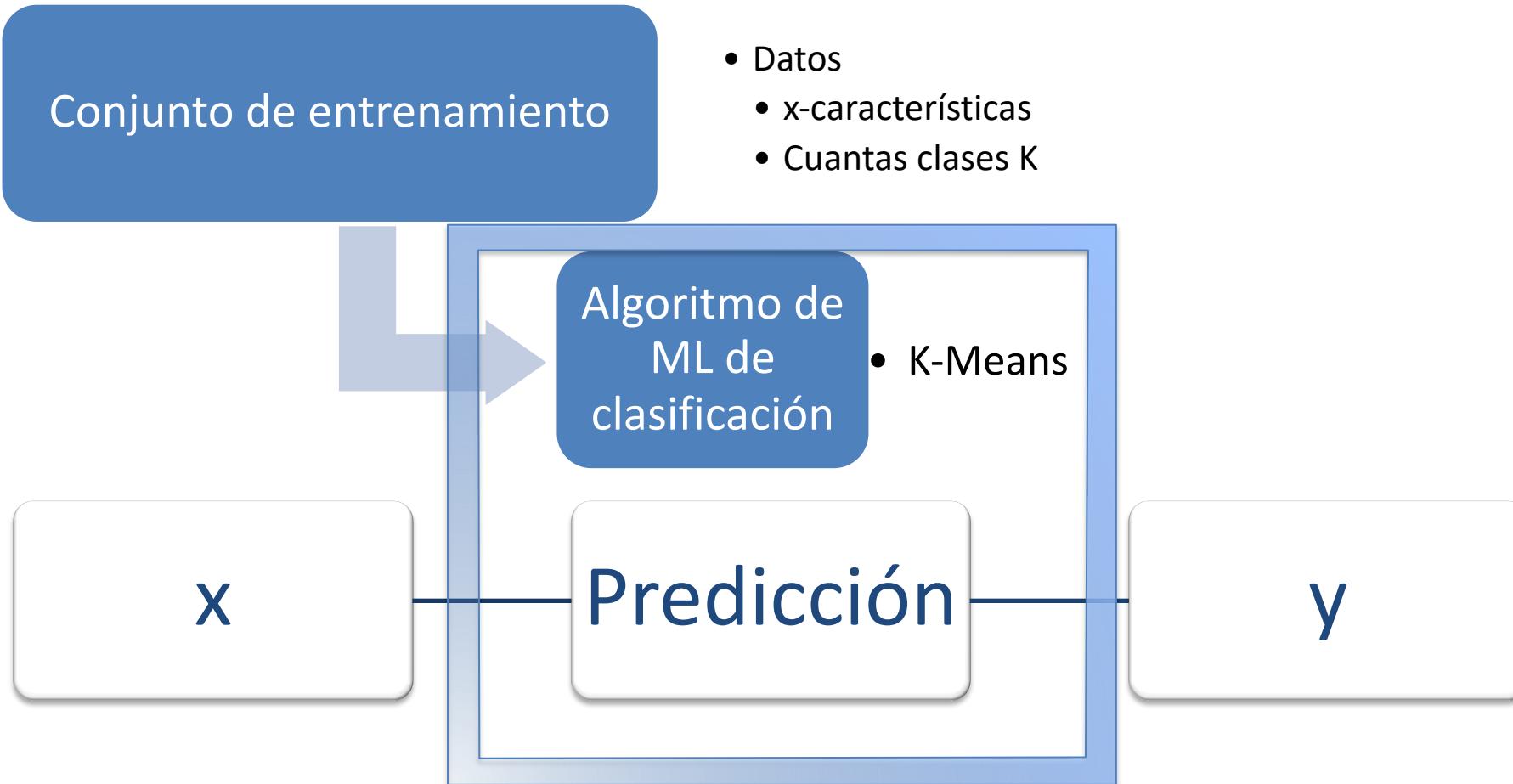
- Supervisados

- Crean un modelo matemático que busca explicar unas “**etiquetas**” de entrada/salida a partir de un conjunto de “**características**” de entrada.
- Se pueden dividir principalmente en:
 - Clasificación
 - Regresión
- Existen otros sub-métodos como:
 - Aprendizaje activo.
 - “Similarity learning”
 - Recommender systems

- No Supervisados

- Crean un modelo que busca explicar las **características** de entrada sin contar con etiquetas.
- Se pueden dividir en
 - Agrupamiento. “clustering”
 - Estimación de densidad (pdf).
 - Reducción dimensional

Idea de Clasificación no supervisada

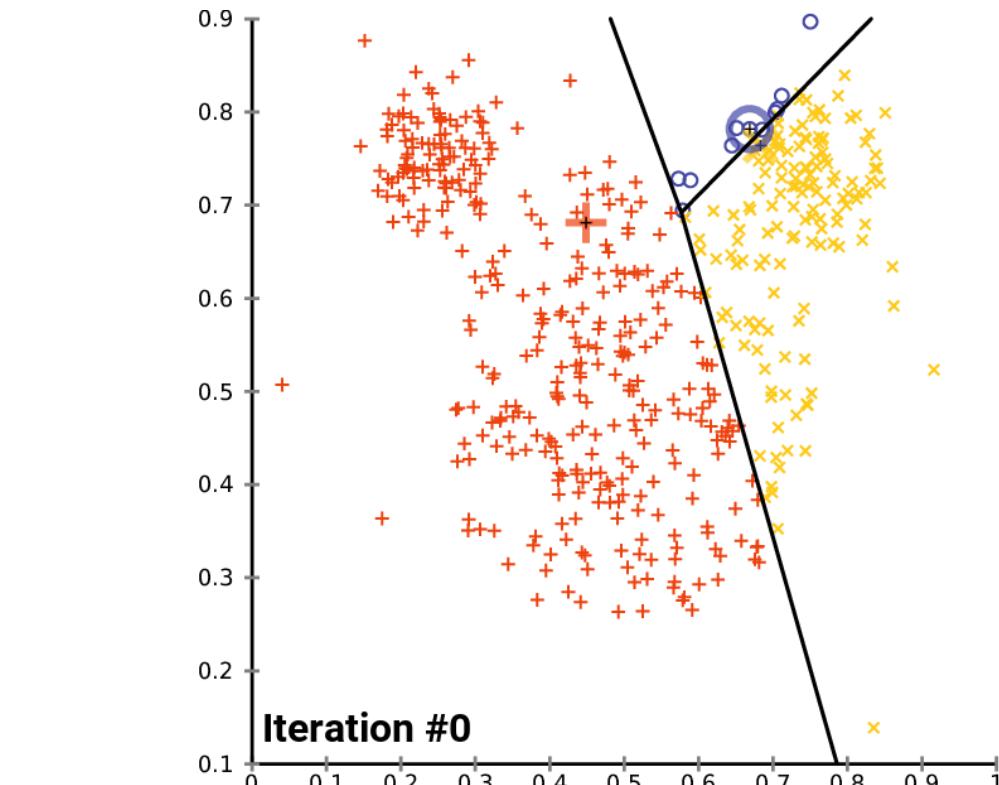


Clasificación no supervisada

Es un método de Aprendizaje Automático donde un modelo se ajusta a las observaciones.

No hay etiquetas “y” dadas por un experto.

Se van a hallar del problema



Fuente: https://en.wikipedia.org/wiki/File:K-means_convergence.gif

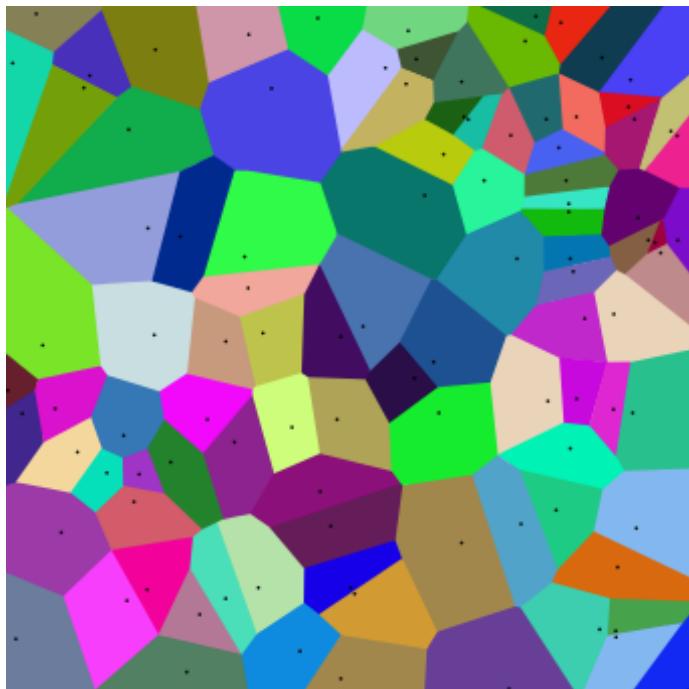
Teselación

Teselación no forma parte del diccionario de la Real Academia Española El término que sí aparece es teselado: Formado con teselas: Cada una de las piezas con que se forma un mosaico



https://www.reddit.com/r/wallpapers/comments/ehmvad/12880_meme_mosaic_wallpaper_version_14000_9200/

Teselación

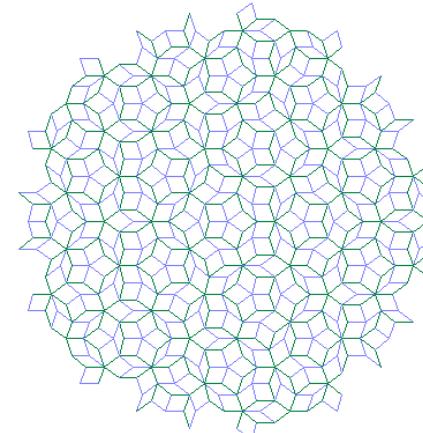


Teselación de Voronoi,
Polígono de Thiessen.

*Todas las figuras son tomadas de wikipedia

Los términos **teselaciones** y **teselado** hacen referencia a una regularidad o patrón de figuras que recubren o pavimentan completamente una superficie plana que cumple con dos requisitos:

- Que no queden espacios.
- Que no se superpongan las figuras.



Teselación de Penrose
https://en.wikipedia.org/wiki/Roger_Penrose

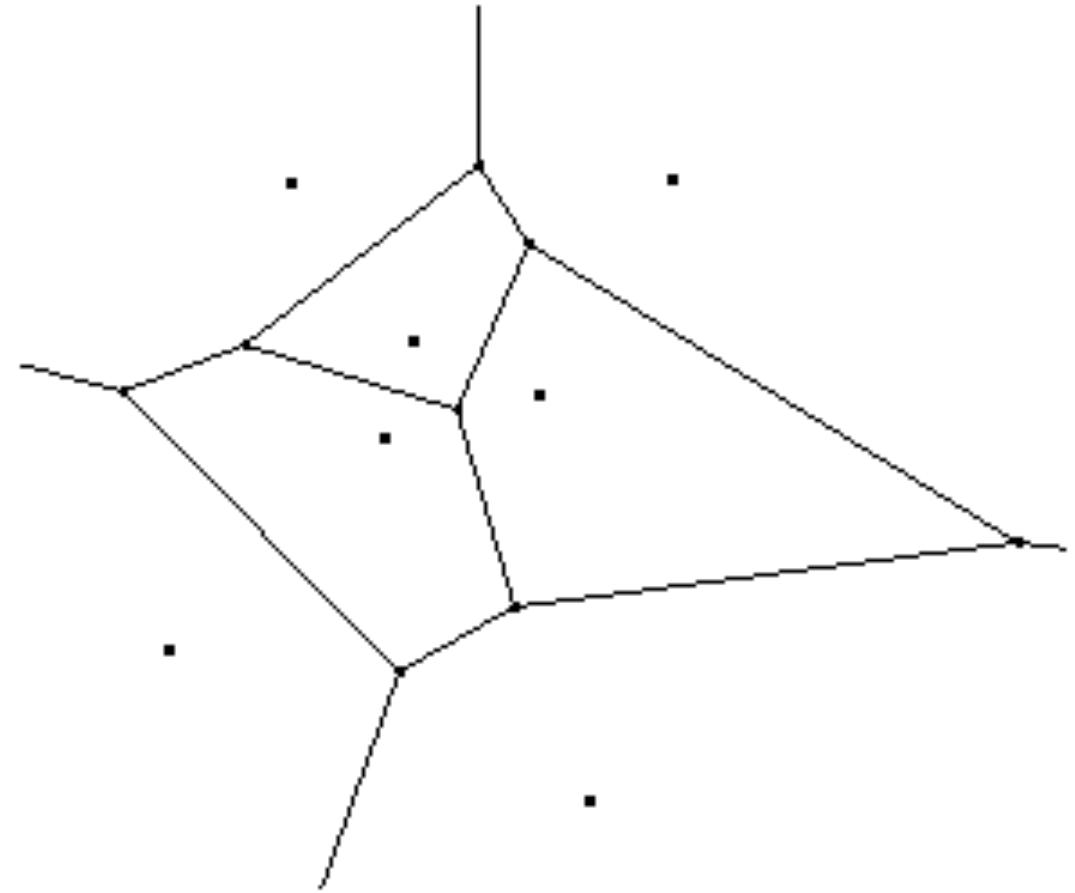


Patrón Teselado

Diagramas de Voronoi

Los puntos se denominan “Sitios” serán nuestros puntos de interés.

La superficie que delimita el diagrama de Voronoi se define como las líneas equidistantes entre pares de sitios



K-medias o K-Means

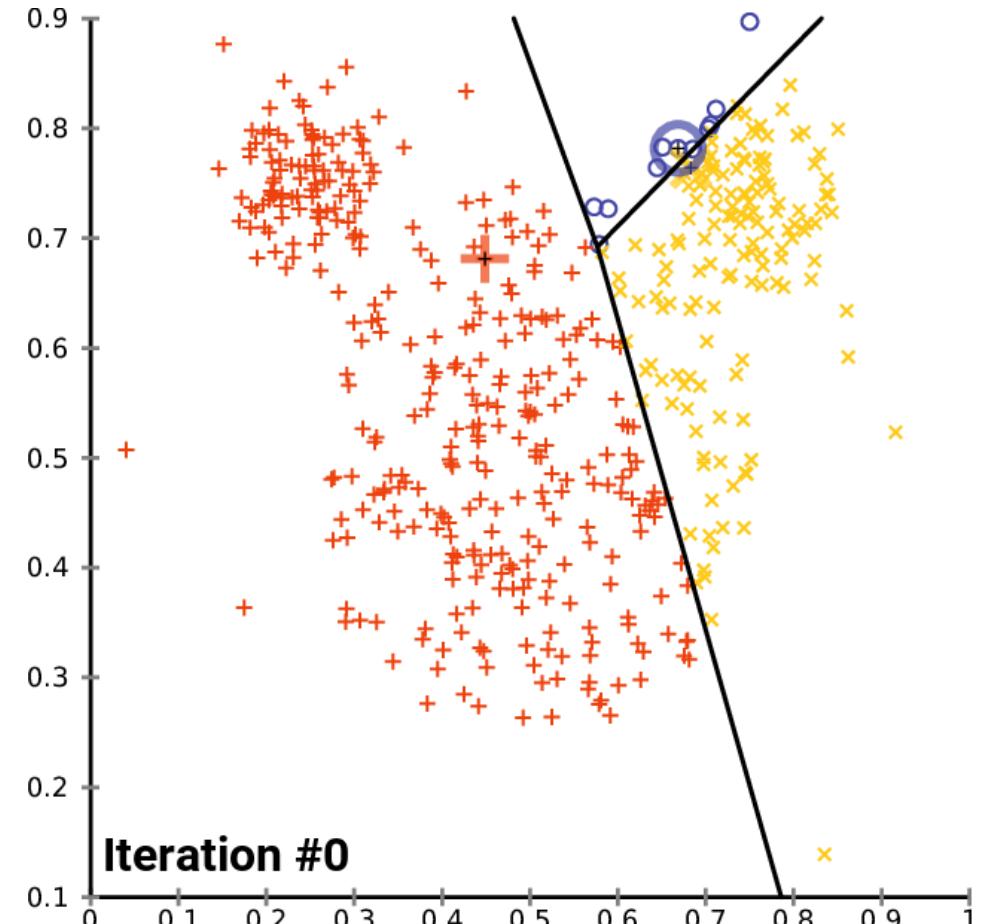
Es un método heurístico de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en K grupos.

- Cada observación pertenece al grupo cuyo valor medio es más cercano.
- El término "k-medias" fue utilizado por primera vez por James MacQueen en 1967



<http://projecteuclid.org/euclid.bsmsp/1200512992>

<https://es.wikipedia.org/wiki/K-medias#Historia>

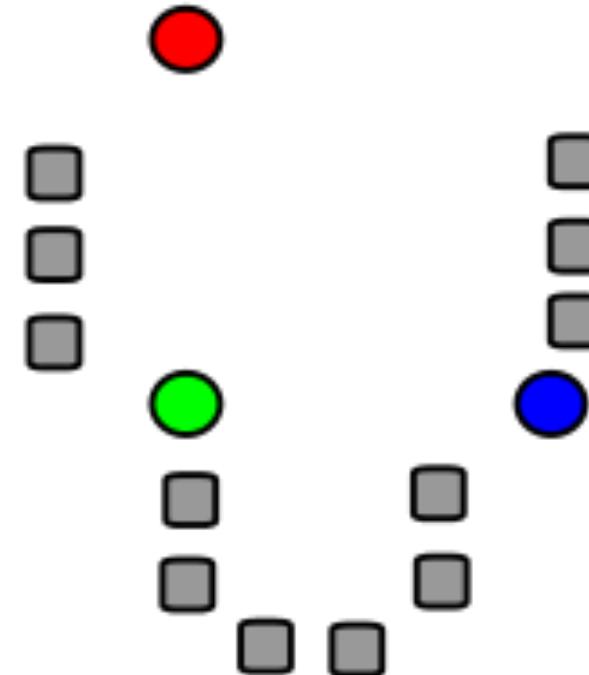


Fuente: https://en.wikipedia.org/wiki/File:K-means_convergence.gif

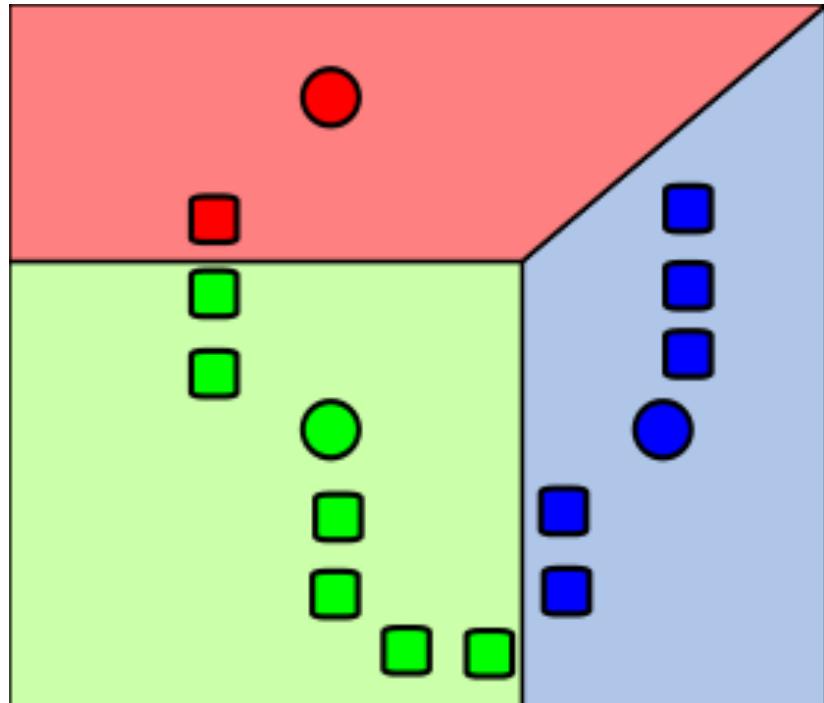
K-medias o K-Means PASO1

Se Asignan K centroides iniciales:

- Pueden ser o no parte de los x del conjunto de entrenamiento.
- Para nuestro caso se marcan en 3 colores diferentes.



K-medias o K-Means PASO 2

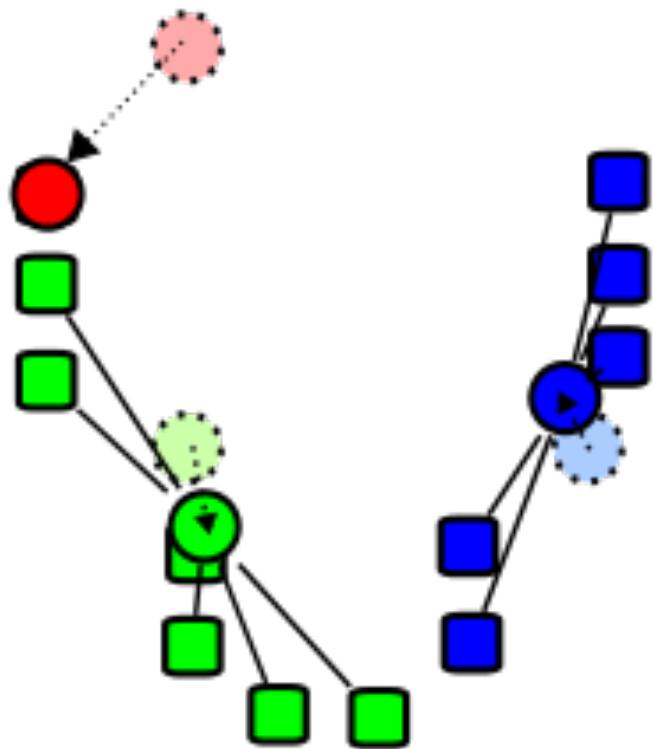


Se hallan las distancias desde cada centroide a cada muestra x.

Se agrupan los más cercanos

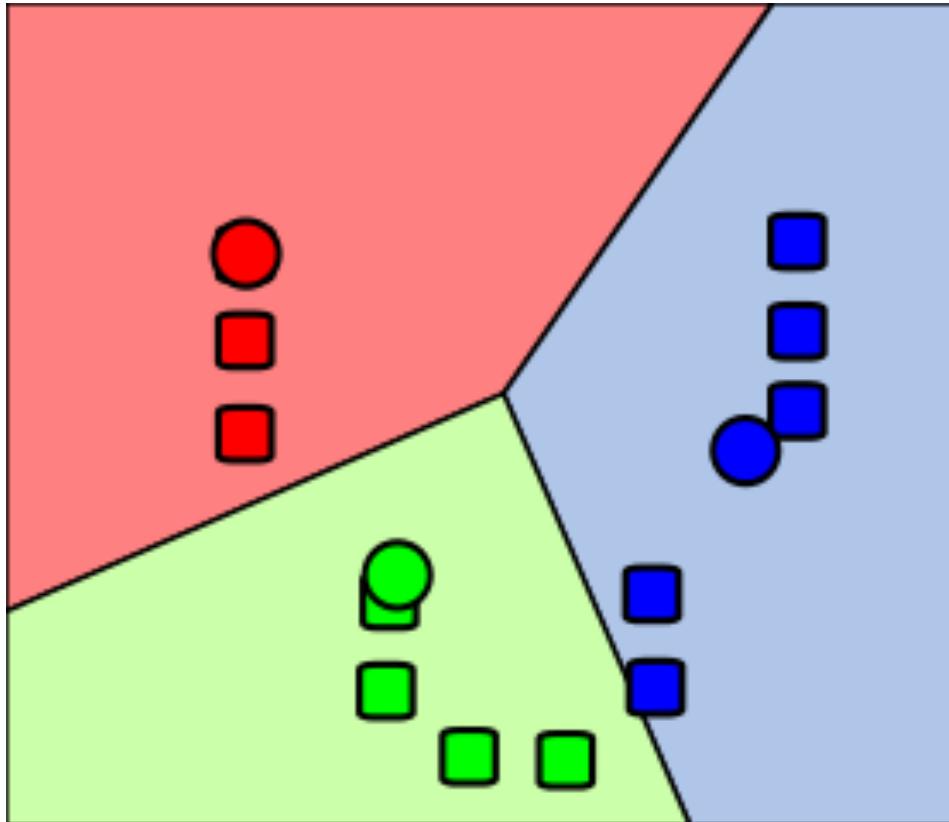
- Este proceso genera un diagrama de voronoi.
- Un conjunto de fronteras de decisión a nuestro problema.

K-medias o K-Means PASO 3



Los nuevos centroides serán entonces la media entre los x correspondientes a cada región.

K-medias o K-Means Ciclo.



Se repite hasta llegar a una condición de convergencia:

- No cambios apreciables.
- Número de iteraciones.
- No convergencia.

Consideraciones con las K- medias.



- El algoritmo no garantiza la convergencia al óptimo global.
- El resultado puede depender de los grupos iniciales.
- Como el algoritmo suele ser rápido, es común ejecutarlo varias veces con diferentes condiciones de inicio.

Consideraciones con las K- medias.

- Existen diferentes variantes para inicializar los centroides:
 - El método Forgy elige aleatoriamente k observaciones del conjunto de datos y las utiliza como centroides iniciales.
 - El método de partición aleatoria asigna k grupos, sin reglas de distancia, tomados aleatoriamente del conjunto x , similar al paso 2, a partir de esos se calcula la media como en el paso 3, y estos centroides son usados como los iniciales.

<https://www.sciencedirect.com/science/article/abs/pii/S0167865599000690>

Ejercicio en clase

- Partiendo del código dado en clase, implementar el algoritmo de K-means en su versión más simple, comparar el resultado con el k-means de algún módulo de python ya implementado.