DIPLOMADO
Ciencia de Datos

Módulo 7
Aprendizaje de máquina no supervisado
1. Clustering, k-medias

Eduardo Espinosa Avila

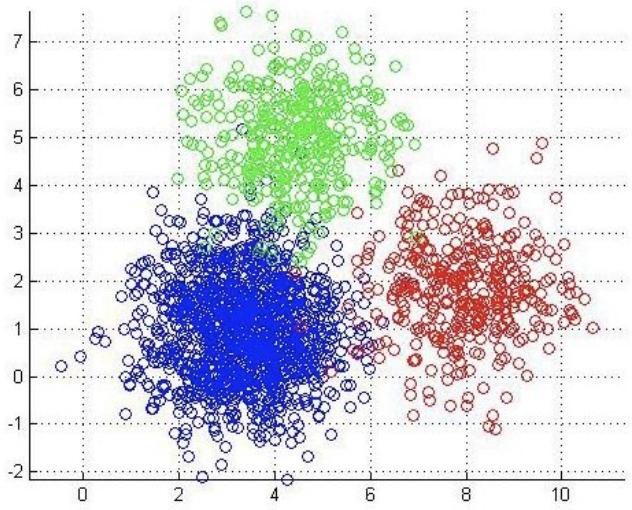
DGTIC
Universidad Nacional Autónoma de México
Dirección General de Cómputo y de Tecnologías de Información y Comunicación

Agrupamiento (*clustering*)

En ocasiones es posible dividir una colección de observaciones en distintos subgrupos, basados únicamente en los atributos de las observaciones.

La intención es realizar división de datos en grupos (clusters) de observaciones, que son más similares dentro de un grupo que entre varios grupos.

Los grupos son formados, ya sea agregando observaciones o dividiendo un gran grupo de observaciones en una colección de grupos más pequeños.



Agrupamiento por *k-means*

Imagen original

Imagen segmentada con K = 3

Imagen original

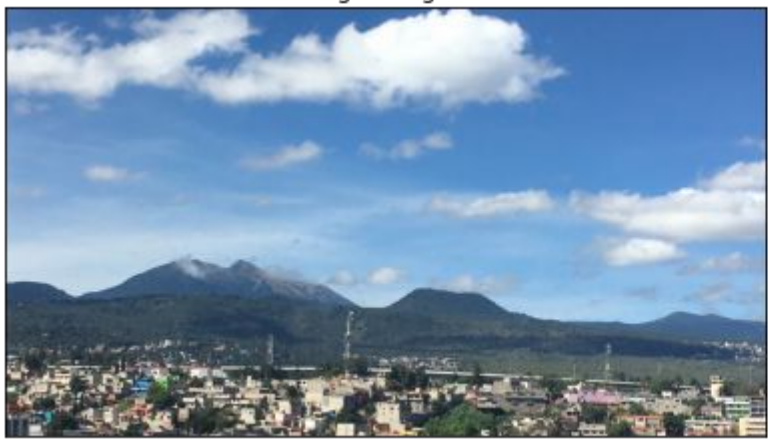
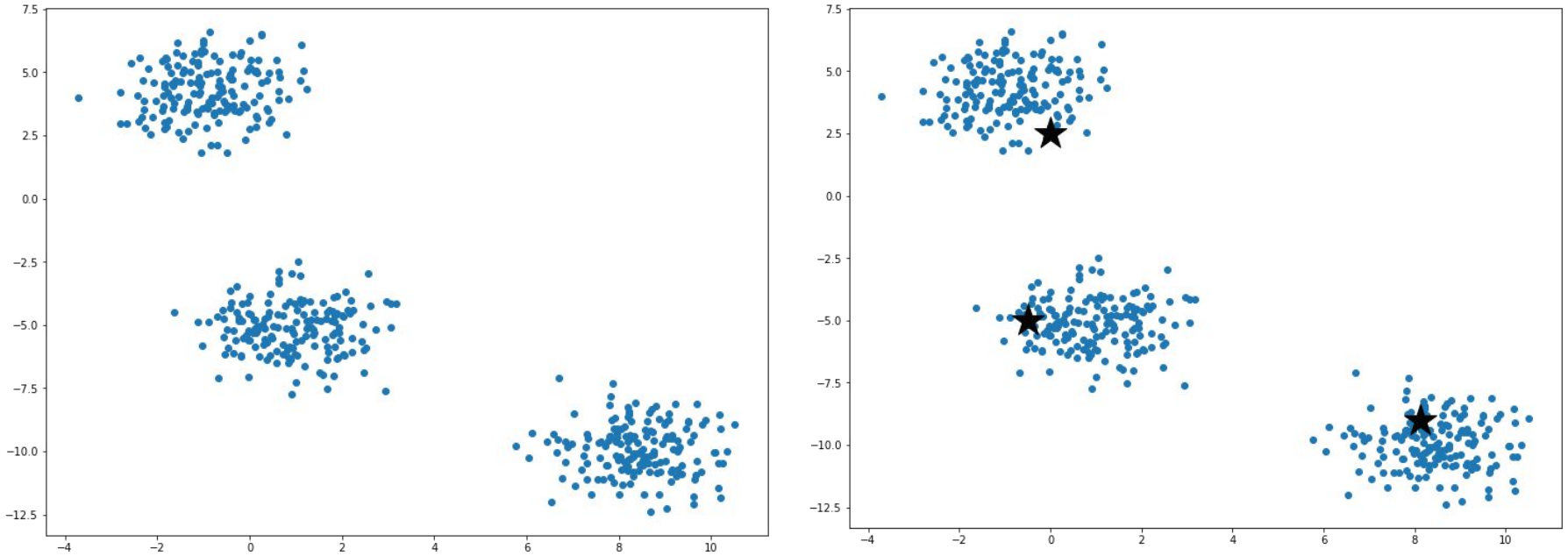


Imagen segmentada con K = 5

Agrupamiento por *k-means*, definición

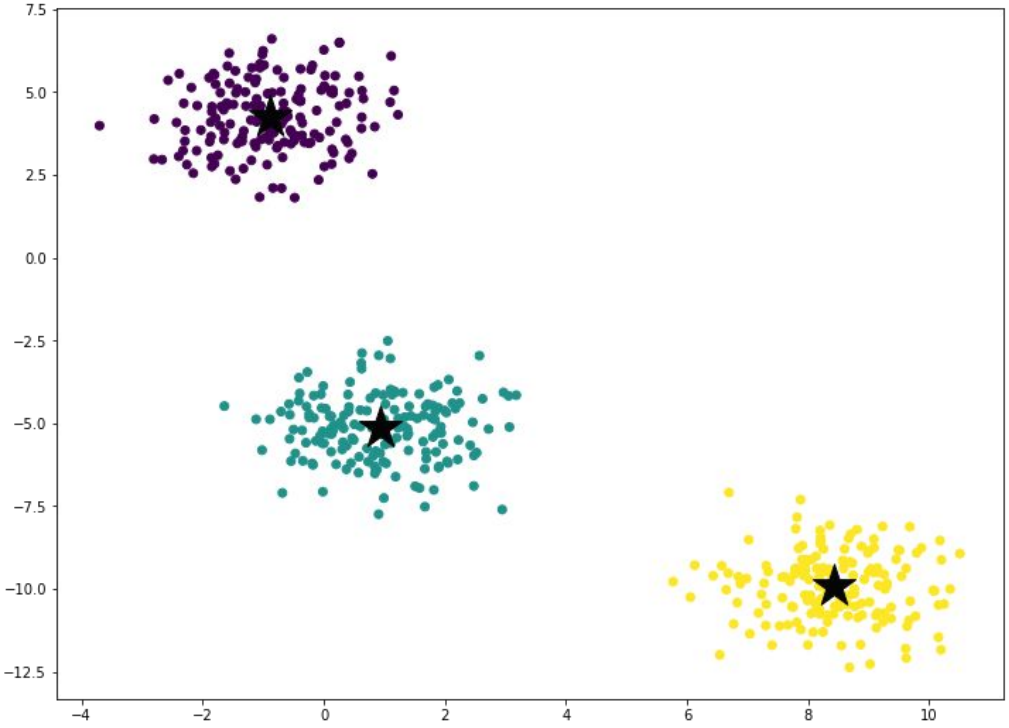


En este algoritmo se establece una cantidad *k* de grupos y se eligen *k* centroides iniciales.

Agrupamiento por *k-means*, definición

Es muy poco probable que la configuración inicial sea una buena solución. Por tanto, el algoritmo itera entre dos pasos hasta que no haya cambio en los centroides:

- Asignar cada observación al grupo más cercano
- Recalcular los centroides de cada grupo



Agrupamiento por *k-means*, métricas

k-medias utiliza por omisión la distancia euclidiana para la asignación de las observaciones a los grupos, en cuyo caso minimiza la suma de las distancias cuadráticas (línea verde).

Pero también es posible utilizar la distancia Manhattan (de taxi). En la imagen las líneas roja, azul y amarilla tienen la misma distancia, de acuerdo con esta métrica.

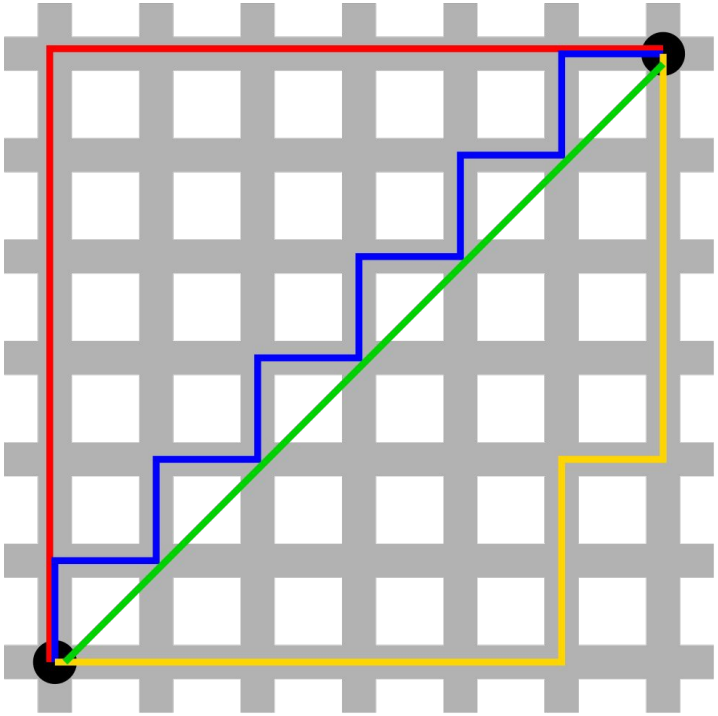


Imagen de https://bit.ly/3tTHmvF

Agrupamiento por *k-means*, aplicaciones

k-medias puede ser usado en muchas situaciones, debido a la sencillez de su implementación, por ejemplo:

- Segmentación de imágenes
- Segmentación de mercados
- Astronomía
  - https://arxiv.org/pdf/1003.3186.pdf
  - https://arxiv.org/abs/1404.3097
- Paso previo para algoritmos más sofisticados (redes neuronales)

Agrupamiento por *k-means*, limitaciones

k-medias siempre termina, pero no garantiza la configuración óptima para los centroides encontrados. El algoritmo tiene algunas debilidades:

- El resultado final depende de los valores iniciales y, frecuentemente, devuelve soluciones subóptimas: es recomendable ejecutar el algoritmo para diferentes inicializaciones.

- No se especifica cómo inicializar las medias. Existen varias opciones. Una de las más comunes, es elegir al azar los *k* valores iniciales. Otra es utilizar *k-means*++ .

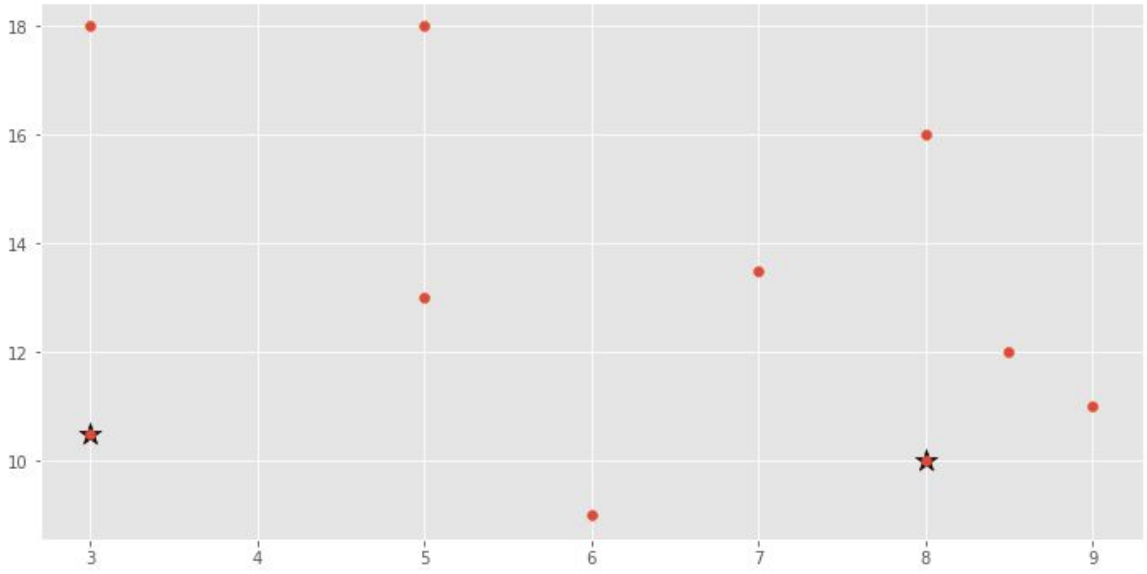
Agrupamiento por *k-means*, ejemplo

Clasificar las muestras siguientes utilizando k=2:

[8,10],[3,10.5],[7,13.5],[5,18],[5,13],[6,9],[9,11],[3,18],[8.5,12],[8,16]

Tomando como medias iniciales:

[8,10],[3,10.5]



Referencias

- Steele, Brian, Chandler, John & Reddy, Swarna
  Algorithms for Data Science / Brian Steele, John Chandler and Swarna Reddy -- Switzerland : Springer, 2016
  1 recurso en línea (430 páginas)
  https://link-springer-com.pbidi.unam.mx:2443/book/10.1007/978-3-319-45797-0

- David Arthur and Sergei Vassilvitskii
  k-means++: The Advantages of Careful Seeding
  http://ilpubs.stanford.edu:8090/778/1/2006-13.pdf

- https://es.wikipedia.org/wiki/Geometr%C3%ADa_del_taxista

Contacto

Dr. Eduardo Espinosa Avila

laloea@fisica.unam.mx

Tels: 5556225000 ext. 5003

Redes sociales:

https://twitter.com/laloea

https://www.linkedin.com/in/eduardo-espinosa-avila-84b95914a/