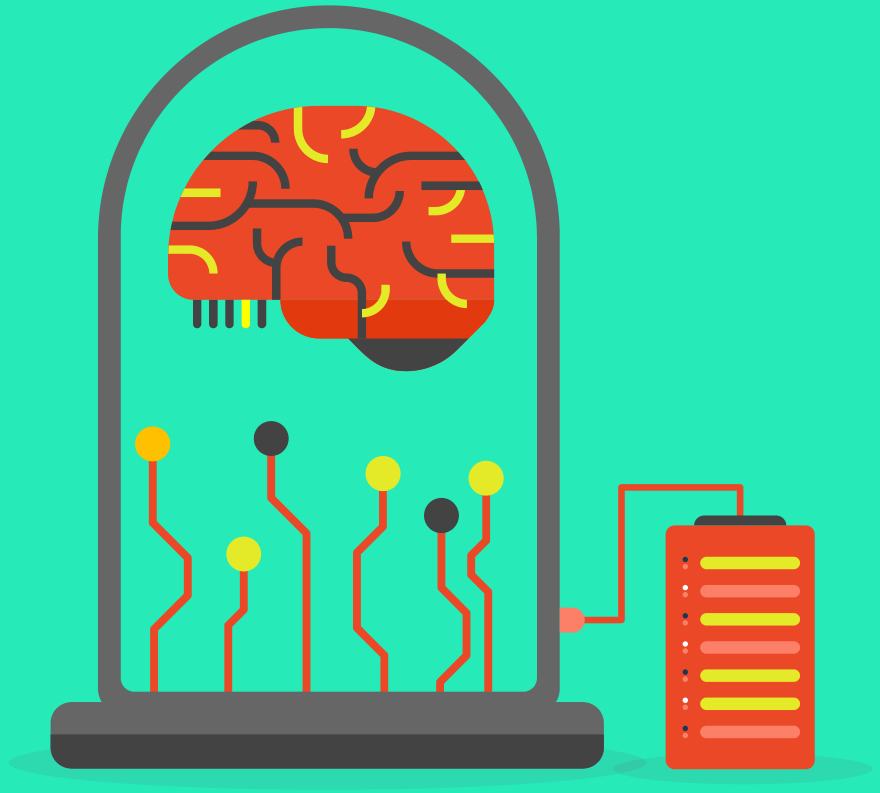
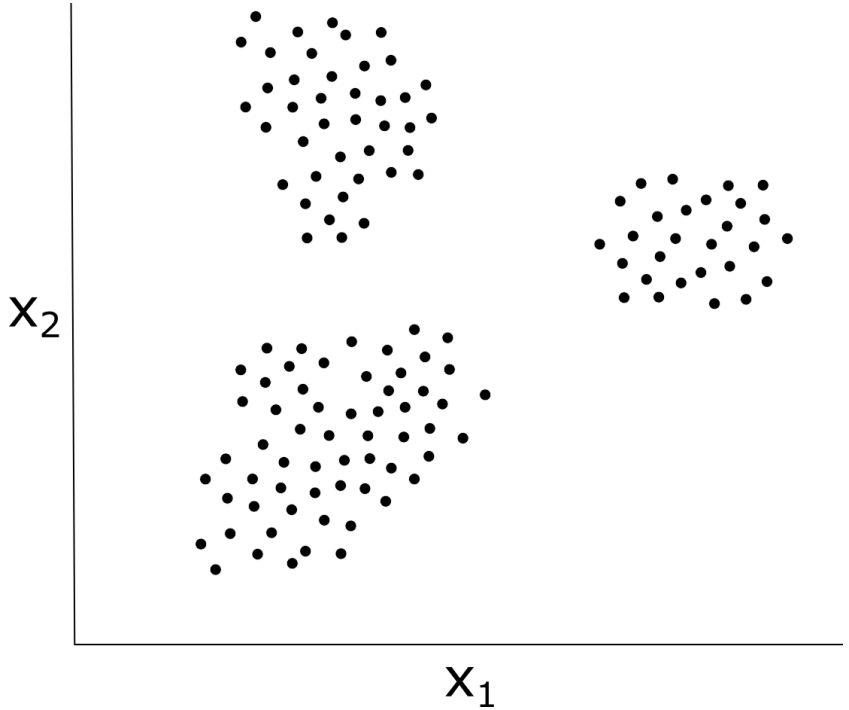
16. Algoritmo de K Medios



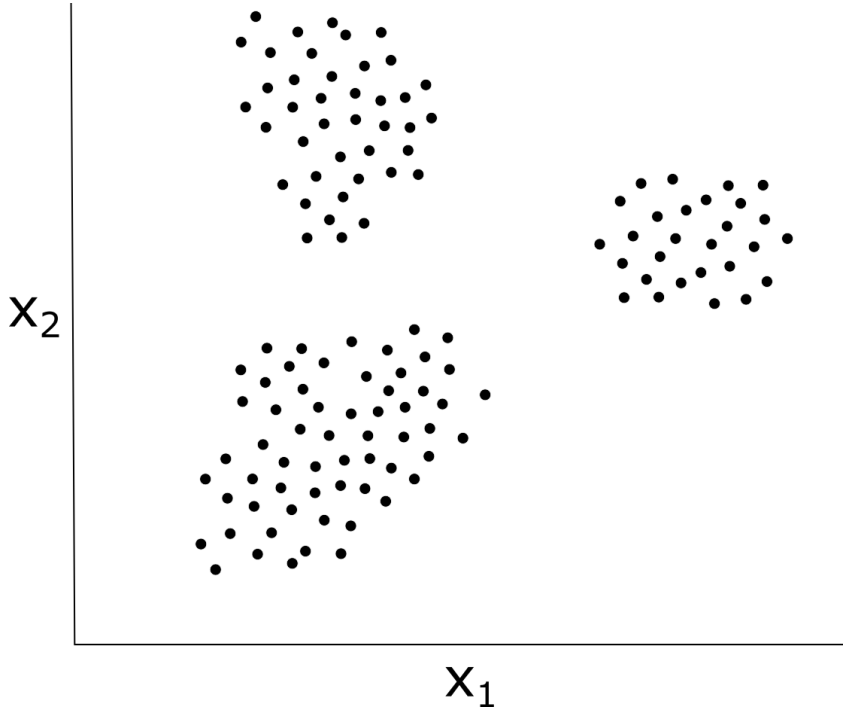
Planteamiento del Problema

Supongamos que tenemos un conjunto de datos no etiquetados que tienen la distribución que se observa en la figura:



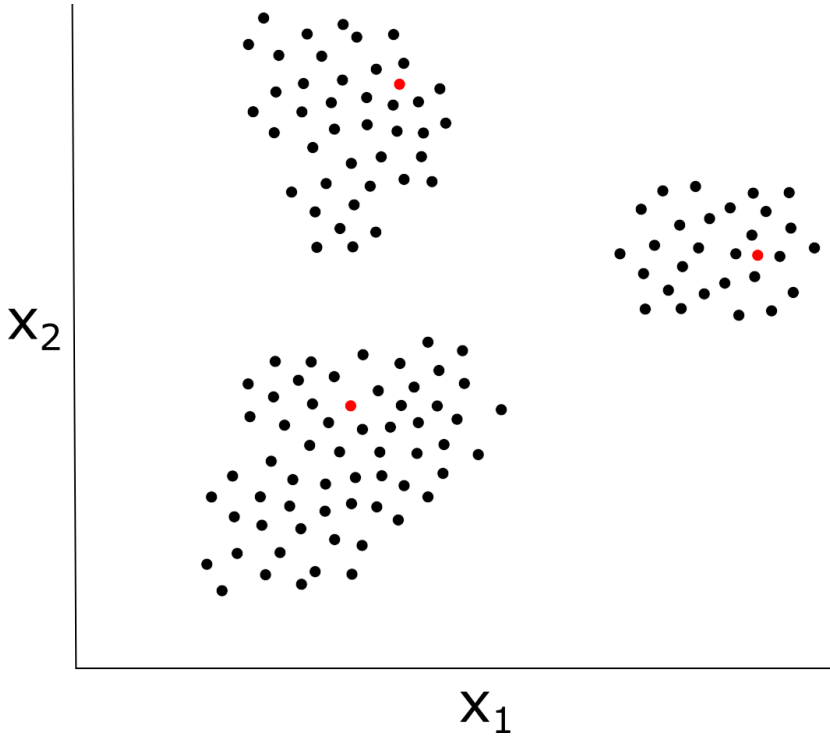
Planteamiento del Problema

Resulta evidente la existencia de tres agrupamientos de datos intrínsecos. ¿Cómo podemos encontrar dichos agrupamientos y etiquetarlos en clases distintas?



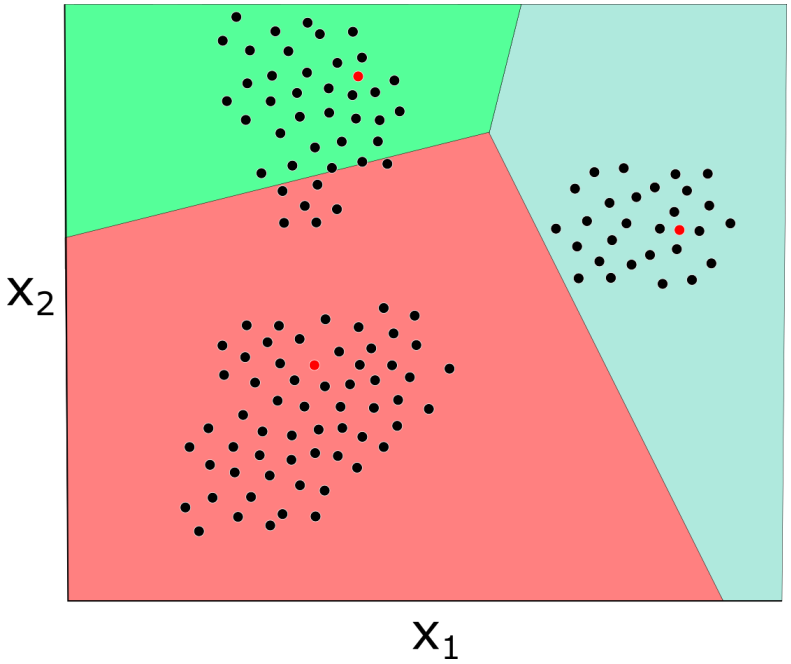
Algoritmo de K Medios (K-Means)

Escojamos K=3 puntos al azar, dentro del conjunto de datos, que llamaremos *centroides*:



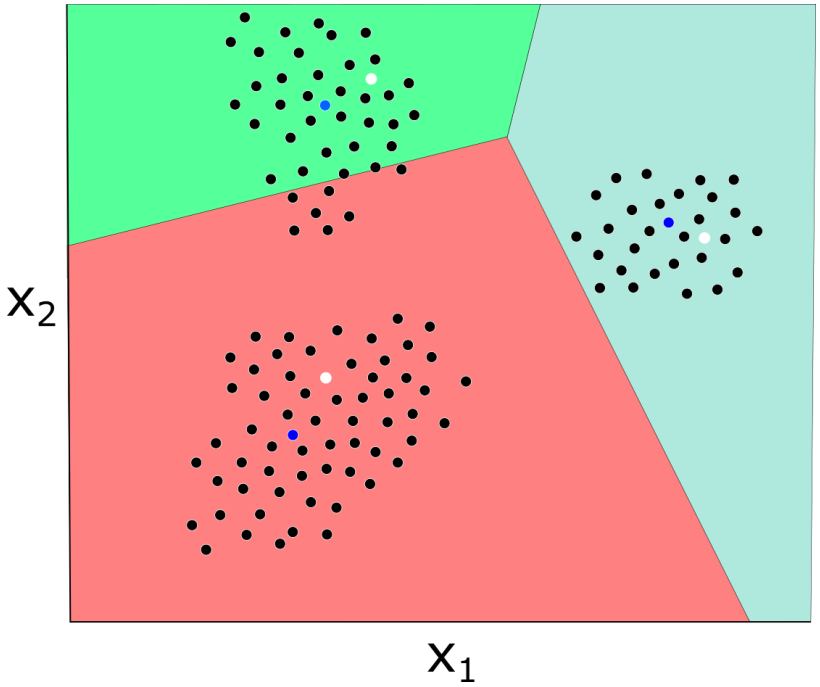
Algoritmo de K Medios (K-Means)

En el primer paso del algoritmo, vamos a calcular la distancia que hay desde cada punto a cada centroide generado, y vamos a etiquetar cada punto como perteneciente a la clase del centroide más cercano.



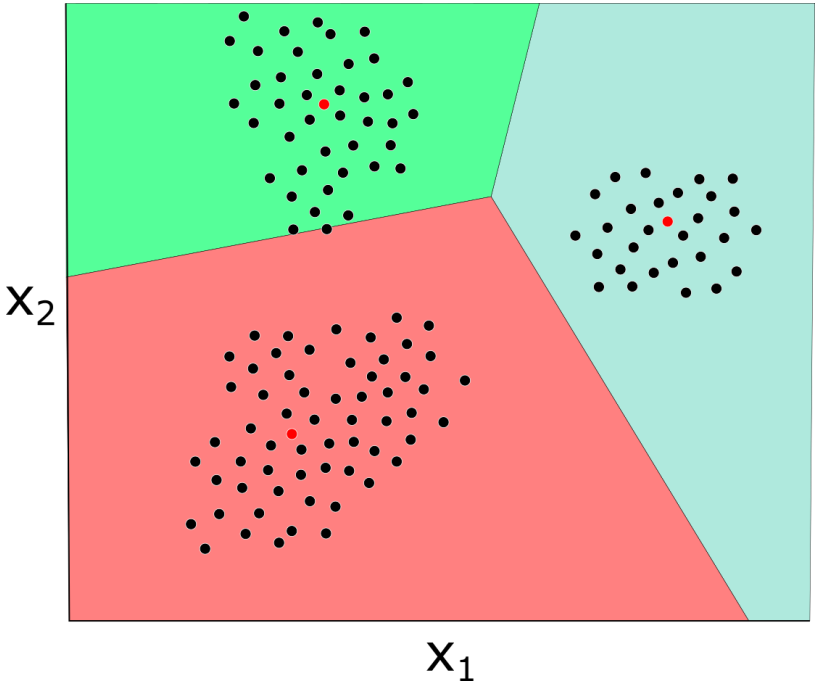
Algoritmo de K Medios (K-Means)

En el siguiente paso, vamos a recalcular la posición de los centroides a partir del promedio de los puntos que fueron agrupados en cada clase:



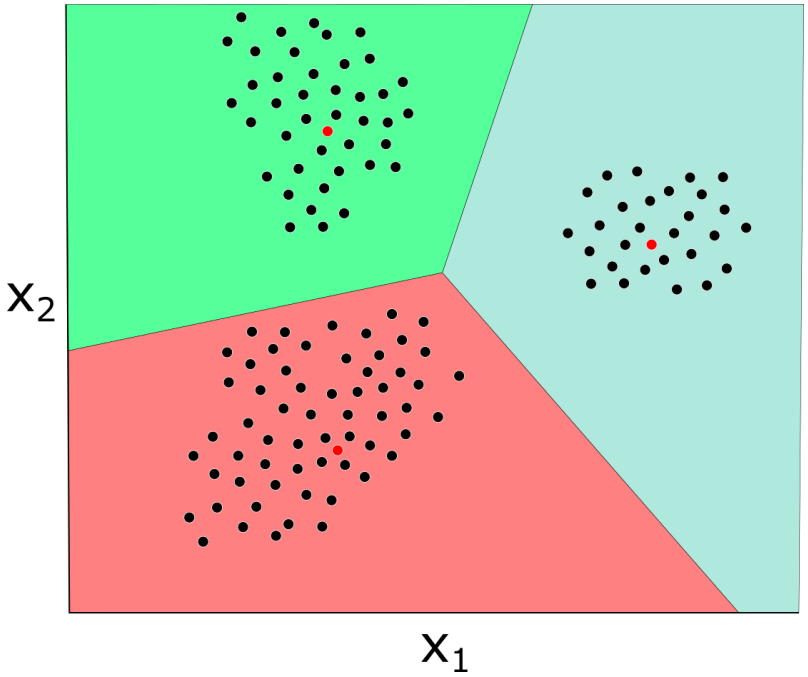
Algoritmo de K Medios (K-Means)

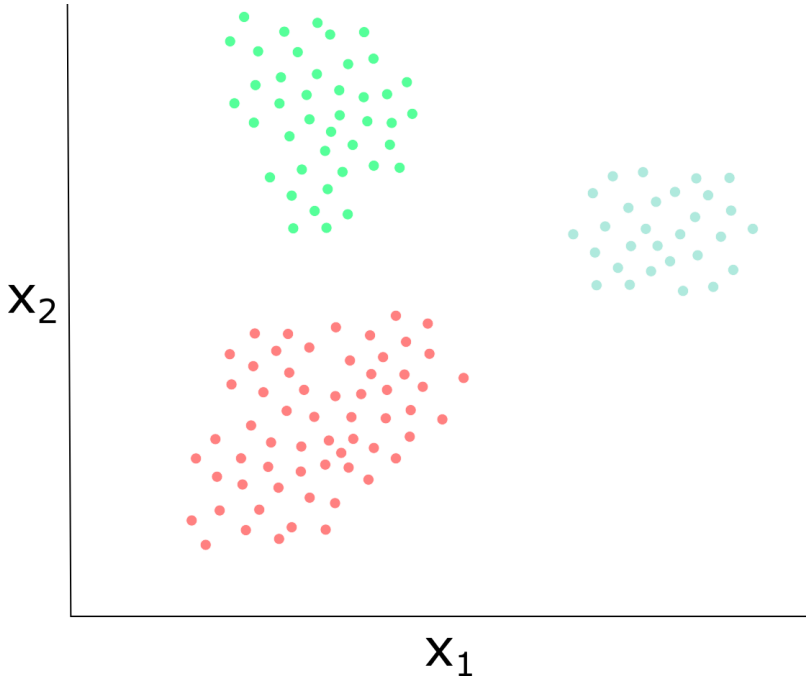
Luego, vamos mover el centroide a dichas nuevas posiciones, y entonces repetimos el primer paso: calcular las distancias de todos los puntos a los centroides y etiquetar las clases. Esto modificará las regiones de clasificación:



Algoritmo de K Medios (K-Means)

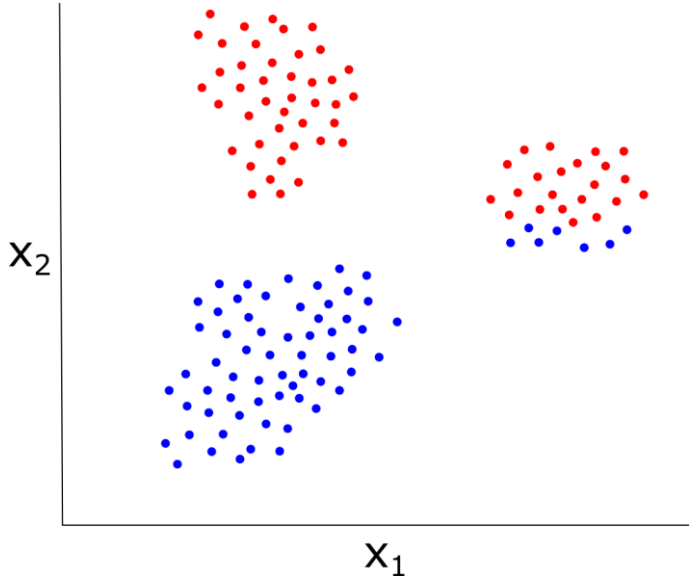
Finalmente, todo este proceso se repite de manera iterativa hasta que los centroides ya no cambien de posición, o hasta que se alcance un número máximo de iteraciones. El resultado, será el agrupamiento de las clases en regiones bien delimitadas:

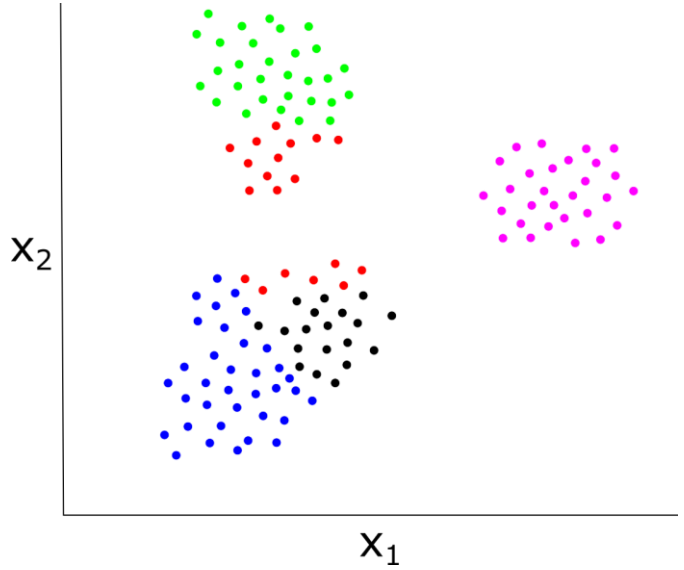




Algoritmo de K Medios (K-Means)

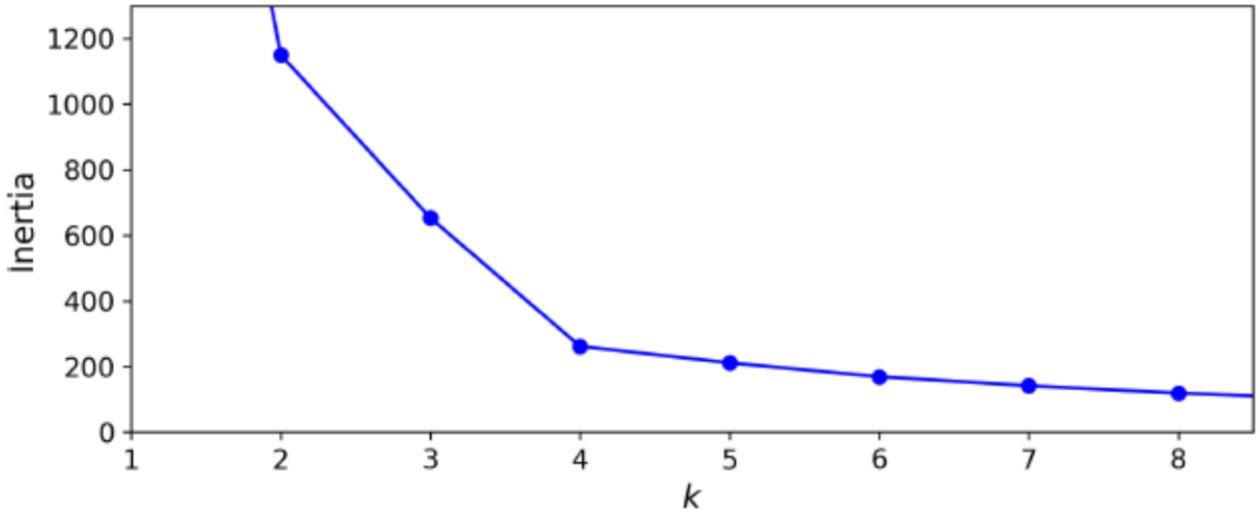
Ahora bien, el número de centroides K seleccionados inicialmente puede ser cualquiera. Así como seleccionamos 3 para el ejemplo anterior, pudimos haber tomado 2 o 5, y eventualmente el algoritmo convergerá y dividirá el conjunto de datos en igual número de clases. Sin embargo, ¿cómo podemos saber la cantidad óptima de K clases a seleccionar para un problema dado?





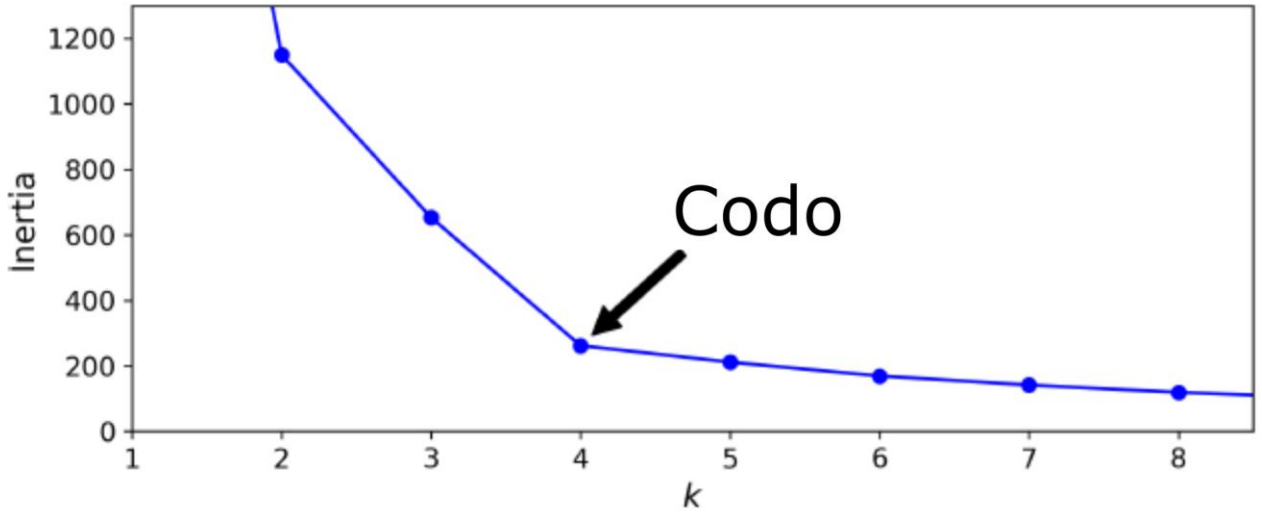
Algoritmo de K Medios (K-Means)

Una manera de encontrar el valor óptimo de K parte de emplear una cantidad llamada **inercia** del modelo, que representa la distancia cuadrática media entre cada punto y su centroide más cercano. A medida que la inercia sea menor, menor es la distancia promedio que existe entre cada centroide y sus grupos. El **Método del Codo** consiste en aplicar el algoritmo variando el valor de K de manera consecutiva, y graficar cómo cambia la inercia en función de K:



Algoritmo de K Medios (K-Means)

El valor óptimo para K se seleccionará como aquel en donde se encuentra el “codo” de la figura, pues es la menor cantidad de Clases que minimiza la inercia:



Algoritmo de K Medios (K-Means)

Si bien el algoritmo de K Medios es muy sencillo de comprender y de implementar, además de ser computacionalmente rápido, también tiene sus limitaciones. Por ejemplo:

- Puede ofrecer resultados distintos dependiendo de la inicialización de los centroides.
- Es necesario especificar el número de centroides de antemano.
- Puede ofrecer malos resultados con datos de diferentes densidades o datos cuya distribuciones no sean esféricas.
- Es sensible a la escala de las variables (por lo que es necesario aplicar reescalado).

Décimo Sexto Notebook Práctico

Una vez que conocemos el fundamento teórico del algoritmo de K Medios, veamos el cómo se implementa el mismo en un dataset generado artificialmente, y luego realicemos una implementación práctica del mismo sobre un dataset conocido para conocer su desempeño como clasificador.

Décimo Sexto Notebook Práctico