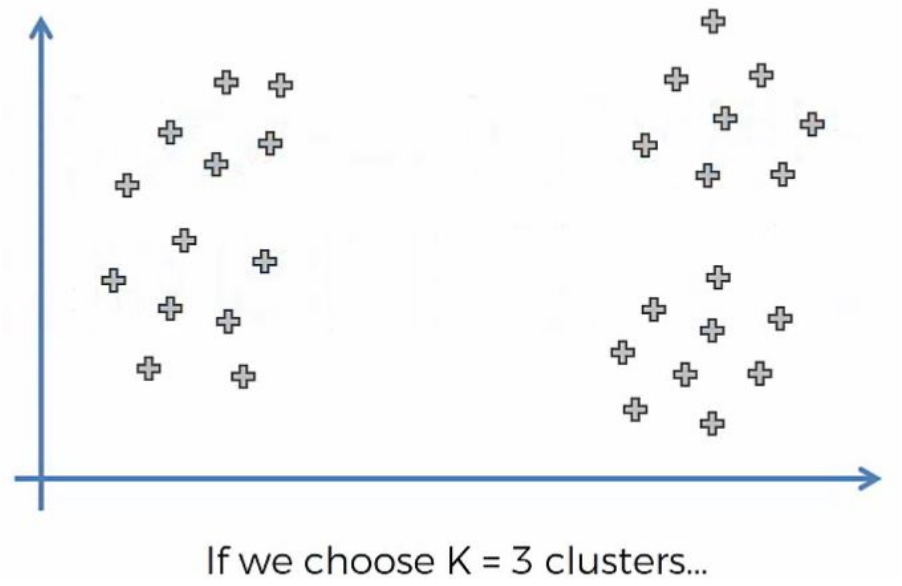
K-Means Intuition: Random initialization Trap

Hablaremos de aspectos específicos de K-Means como la inicialización aleatoria

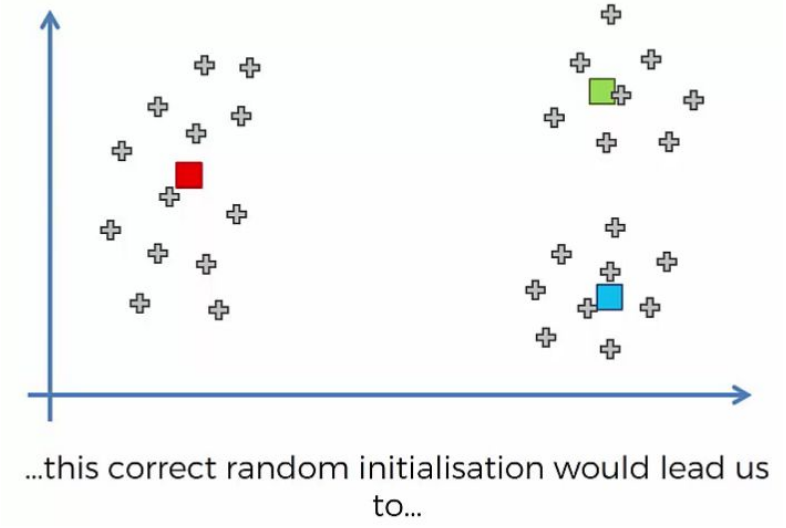
Trampa de inicialización aleatoria, veamos en que consiste:

Tenemos un diagrama de dispersión, de nuevo con dos variables de coordenadas x e y

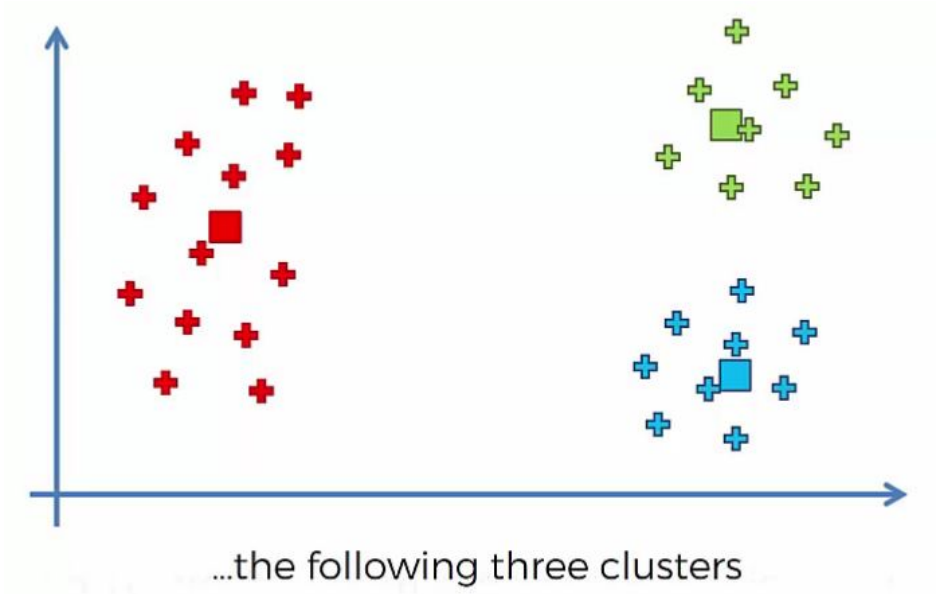
Si tenemos un diagrama de K=3 clústeres:



De imediato no podemos decir cual es el resultado o como se verá. Los grupos de datos se ven muy identificables, así que inicializamos así nuestros centroides, para que el algoritmo converja más rápido en este ejemplo y no tengamos que hacer varios pasos.

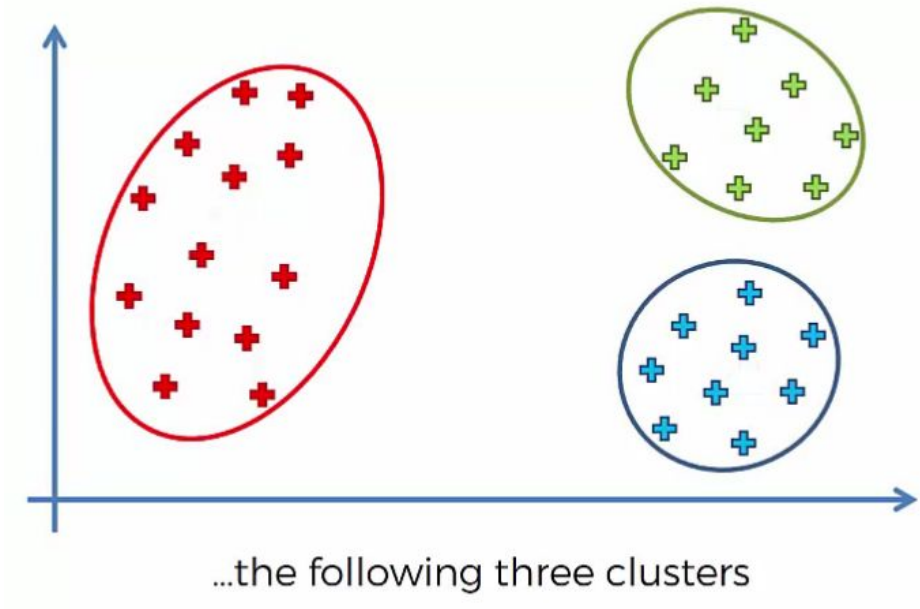


Y si ejecutamos K-Means lograremos estos clústeres



Este ya el resultado final, se ha obviado el proceso para no abordarlo aquí, ya se hizo acá

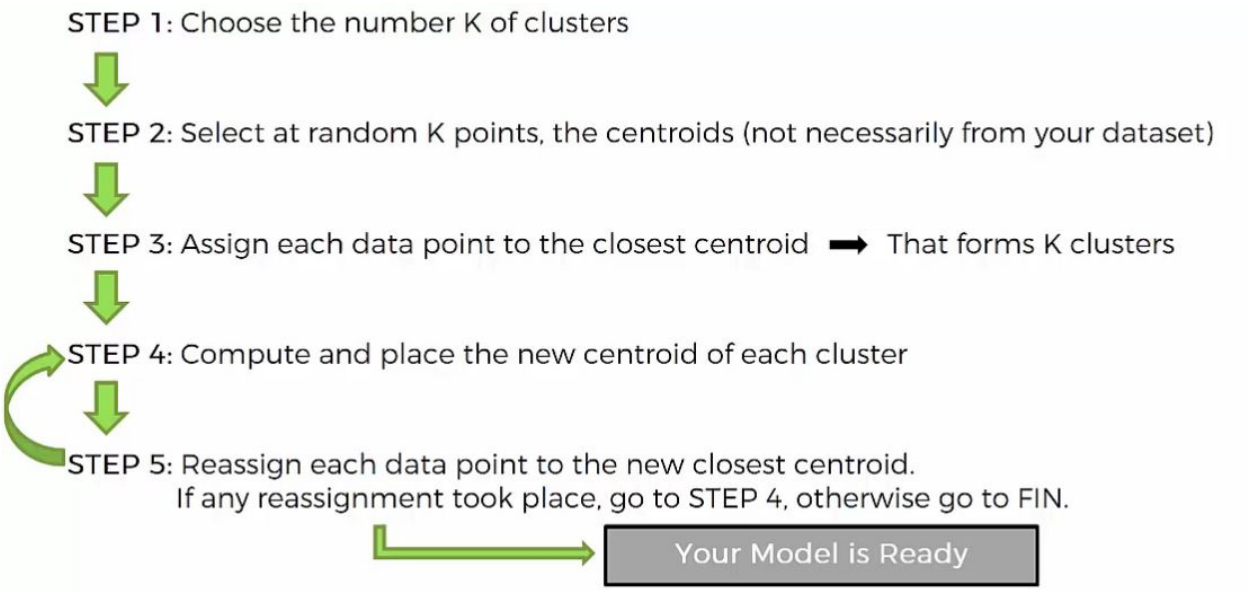
Estos serán los grupos con losque vamos a terminar



La pregunta es: ¿Qué pasa si seleccionamos un centroide en diferentes ubicaciones, lo cuál sera capaz de cambiar el resultado?
Porque se podría seleccionar el centroide al azar, pero no queremos que la selección de los centroides no afecten la forma en como se conformarán los clústeres.

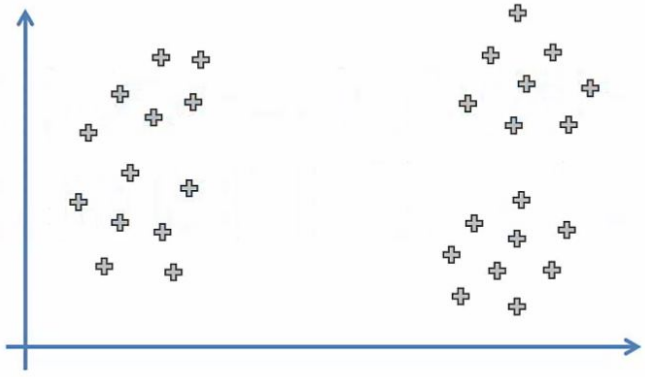
¿Qué sucede si tenemos una mala inicialización aleatoria?

Entonces de nuevo seguimos los pasos de K-Means

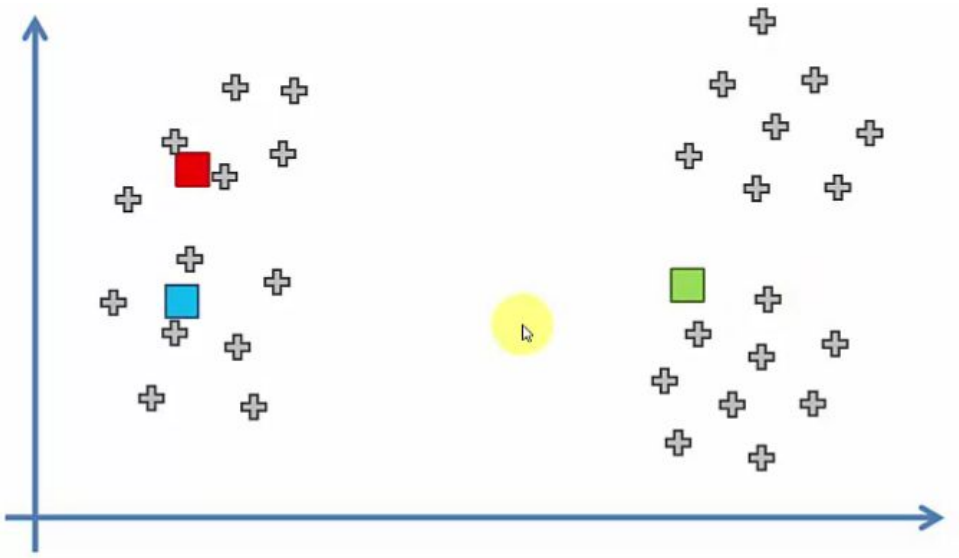


Escogeremos un número de clústers K = 3, todos seleccionados aleatoriamente, tres puntos que serán nuestros centroides y esto formará K clusters que serán calculados y ubicarán nuevos centroides para cada clase de pensar en terminos de centro de massa o centro de gravedad y se reasignarán los puntos cuando toque y de ser necesario volveremos a calcular el lugar nuevo para los centroides y asi hasta que el modelo converga

Procedemos:
Seleccionaremos los centroides de forma diferente

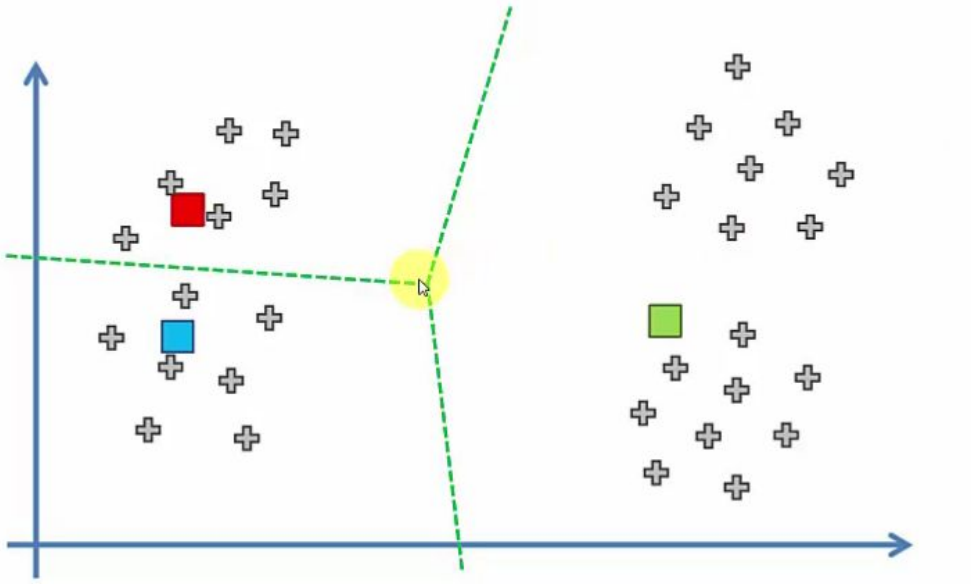


If we choose K = 3 clusters...

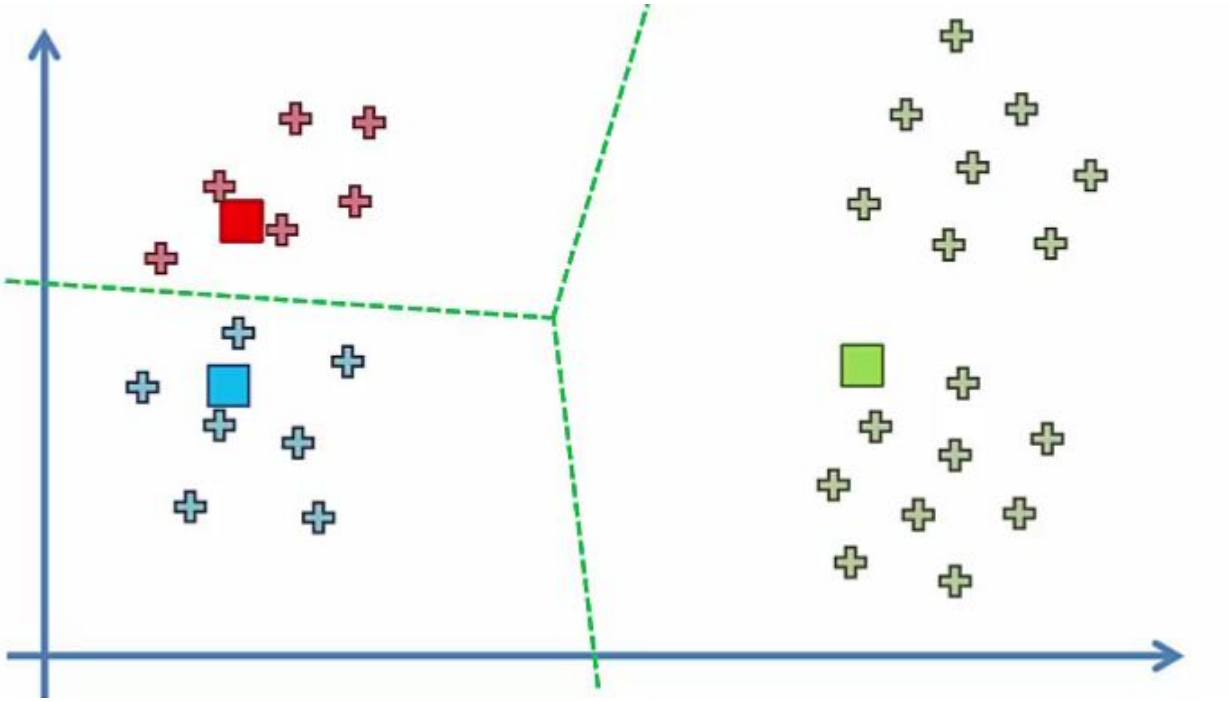


Y miremos que sucede

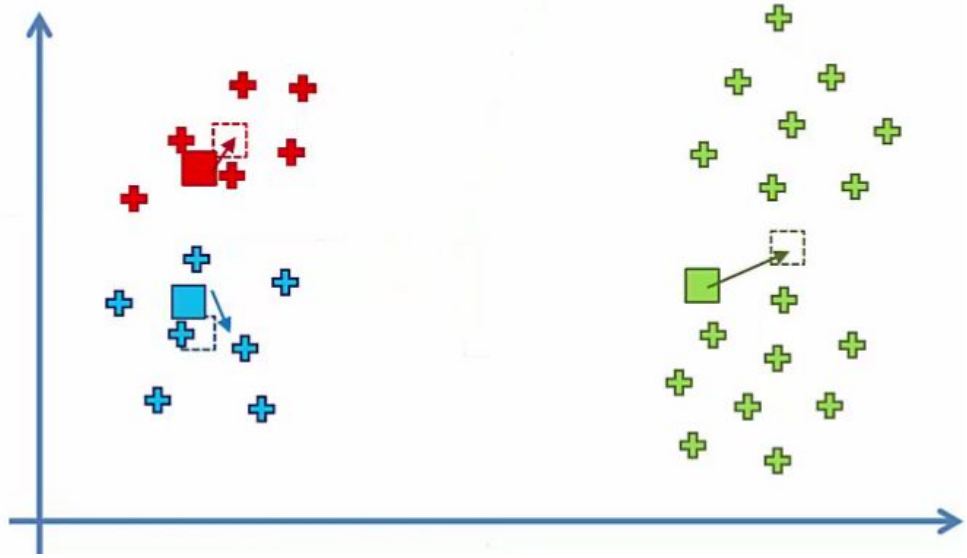
Si trazamos una línea y tenemos tres clústers tres centroides, la linea es equidistante a todos los centroides asi que los puntos de datos en la linea verde estan a igual distancia de los centroides rojos y azules y en el verde con respecto al rojo y azul.

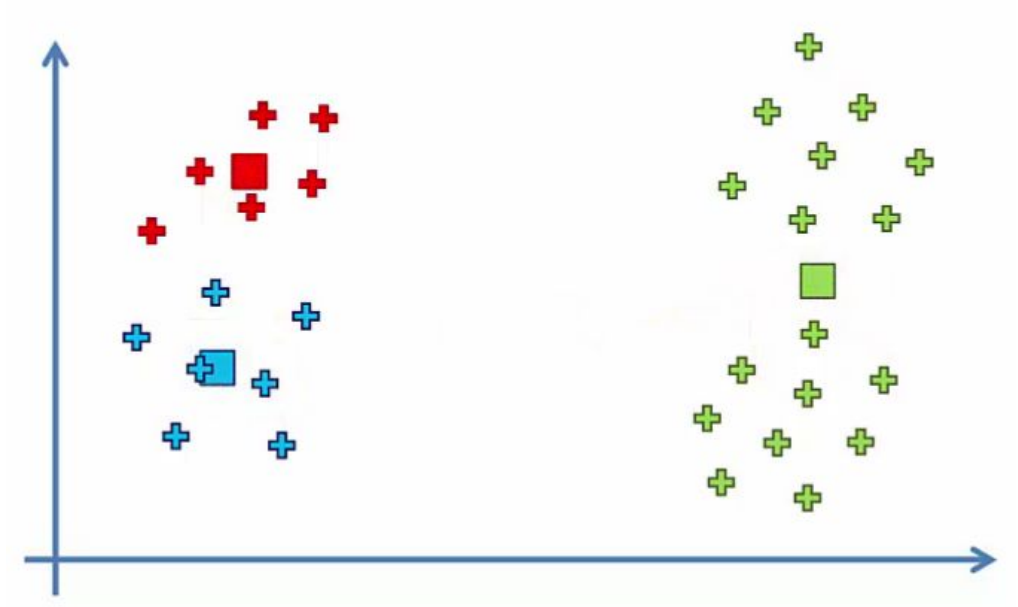


Acorde a este método podemos diferenciar cada clúster, hay uno rojo, uno amarillo y uno verde

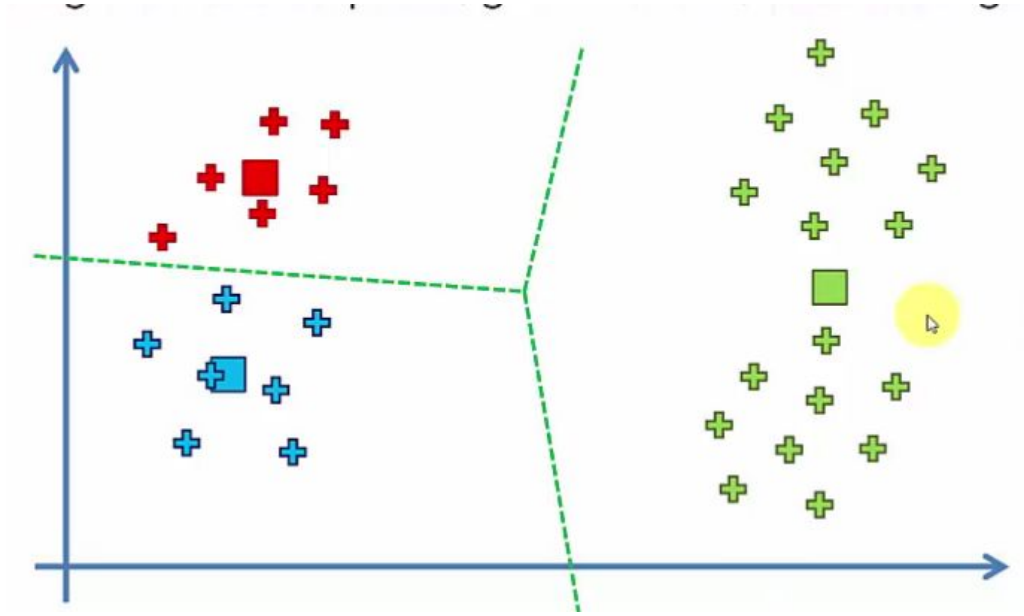


Asignamos cada punto de datos al centroide más cercano y es genial ya están.
ENtonces vamos al paso 4 y vamos a remover estos clústers o moverlos acorde al centro de gravedad de los datos



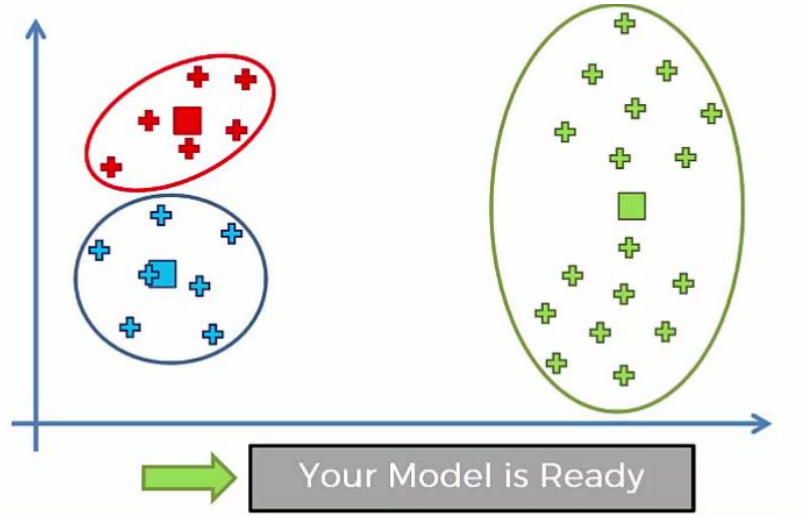


Ahora ejecutamos el paso 5 re-asignando los puntos de datos al nuevo centroide más cercano mirando la equidistancia de puntos y lineas con respecto al centroide

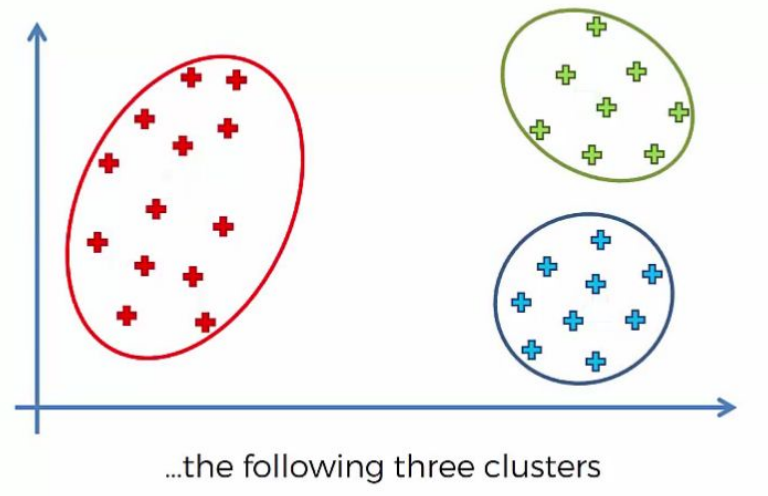


Nada va a cambiar, todo está en su lugar lo que significa que el algoritmo convergio.

Como resultado tenemos tres diferentes clusters



Y este resultado es diferente a lo que miramos al principio que era esto



Aqui tenemos una situación o fenómeno en donde la selección del centroide al comienzo del algoritmo puede potencialmente dictar el resultado del algoritmo. Y esto no es bueno, pirque el centroide es seleccionado al azar.

Entonces ¿Cómo combatir esto?
La respuesta no es tan simple
Existe una adición o modificación al algoritmo de K-Means que permite seleccionar correctamente el centroide y es el K-Means ++

Pero no porfundizaremos en su estructura, es un enfoque bastante complicado en cómo se produce esa selección, pero la buena noticia es que todo esto sucede en un segundo plano, de modo que KMeans ++ sucede en pYthon o cualquier otra herramienta que usemos y actualmente no necesitamos implementarlo

Por lo tanto, es una buena idea estar al tanto de este problema de que hay un verdadero resultado de clúster que está buscando y que puede haber algunos resultados de agrupamiento falsos o no deseables en clústeres de los K-means

Es bueno conocer este issue y es bueno también conocer las herramientas que podemos usar o asegurarnos de que las herramientas que estemos usando seran adecuadas para obtener un buen resultado

Si queremos aprender más sobre K-MEANS ++ hayq ue leer acerca de el, pero no es algo de lo que nos debamos preocupar