K-Means Intuition: Understanding K-Means

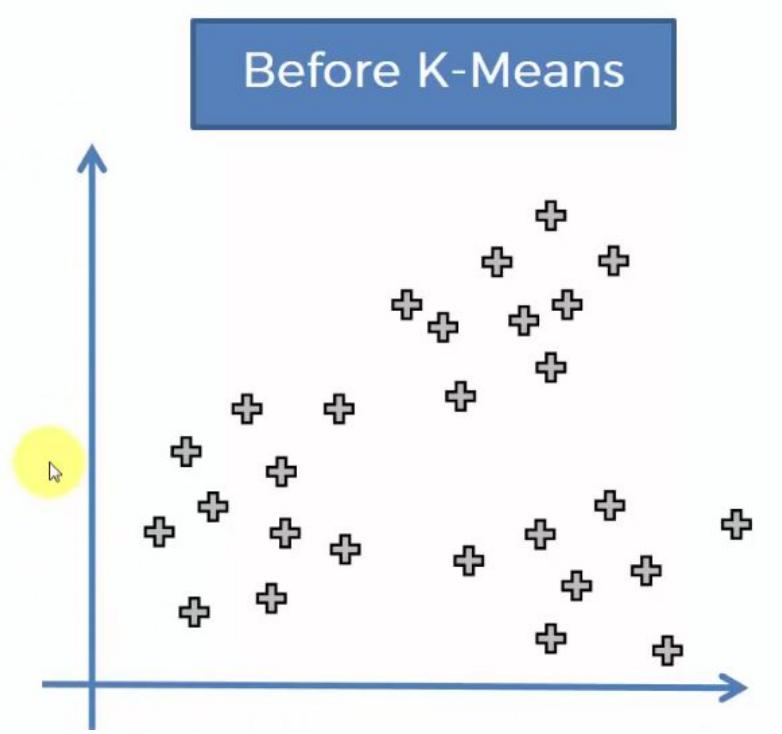
Permite agrupar nuestros datos y es muy conveniente para descubrir categorías de grupos en un dataset, que de otro modo no hubiésemos pensado

Aprenderemos como entender K-Means en un nivel muy intuitivo. Haremos que esto sea sencillo

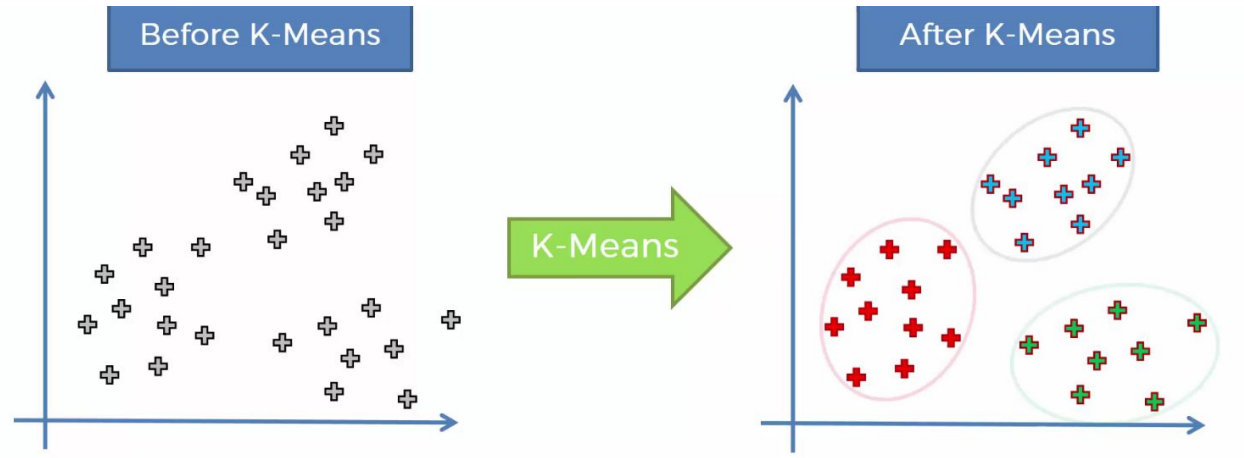
Tenemos un diagrama de dispersión. Imaginemos que tenemos dos variables en nuestro dataset y que decidimos dibujar estas dos variables sobre el eje X y el eje Y. Y es si como nuestros datos son configurados acorde a estas dos variables.

La pregunta es ¿Podemos identificar ciertos grupos entre estas variables?
POdemos identificar 2 grupos, tres grupos, 4 o 5 grupos aquí

¿Como podemos identificar el número de grupos? ¿Cómo podemos identificar los grupos en sí mismos?



Lo que K-Means hace por nosotros es eliminar la complejidad de este proceso de toma de decisiones y nos permite identificar muy fácilmente esos clústeres que en realidad se llaman clústeres de puntos de datos en mi dataset



Tenemos 3 clases y es un ejemplo muy simplificado, solo tenemos dos dimensiones aquí, asi que dos variables de K-Means pueden funcionar con objetos multidimensionales, con 3, 4, 5 , 10 dimensiones
Esto es solo para fines de ilustración, de modo que podemos ver visualmente lo que está sucediendo, pero en realidad pueden ser hasta 10, o 100 cualquier cantidad de variables y K-Means hará ese cálculo complejo por nosotros.
Es un algoritmo diseñado para encontrar estos grupos por nosotros

¿Cómo funciona K-Means?

1. Elegir la cantidad de clústers K (se hablara de como seleccionar la cantidad óptima de ellos)

Por ahora asumamos que acordamos varios clústers para un ejemplo específico y acordamos que existen 3 clústers de estos 5 clústers o dos clases

2. Seleccionar aleatoriamente K puntos, los centroides, que serán el centro de gravedad de mis clústeres, y no necesariamente estos puntos deben ser del conjunto de datos.

El número de centroides debe corresponder con el numero de clústers decidido

3. Asignar cada punto de datos al centroide más cercano y eso formará K clústers inmediatamente

Comienzo formado clústers (paso 1) y aca en el tres se da un proceso iterativo para refinar esos clústers asignando cada punto de datos a su centroide más cercano, Esto se repite

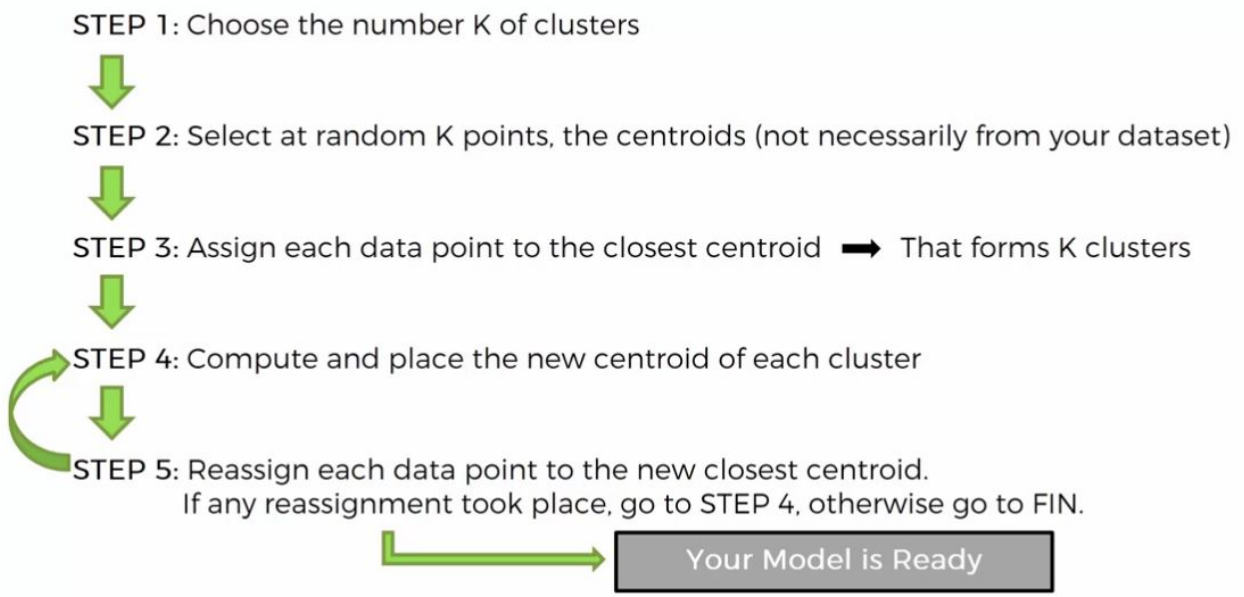
La distancia a la que se mide un punto de dato con el centroide es a distancias Euclidianas, existen otro tipo de distancias que pueden ayudarnos mejor, pero hablaremos de distancias Euclidianas, lo que es básicamente distancias geométricas en términos simples.

4. Se calcula y ubica el nuevo centroide de cada clúster.

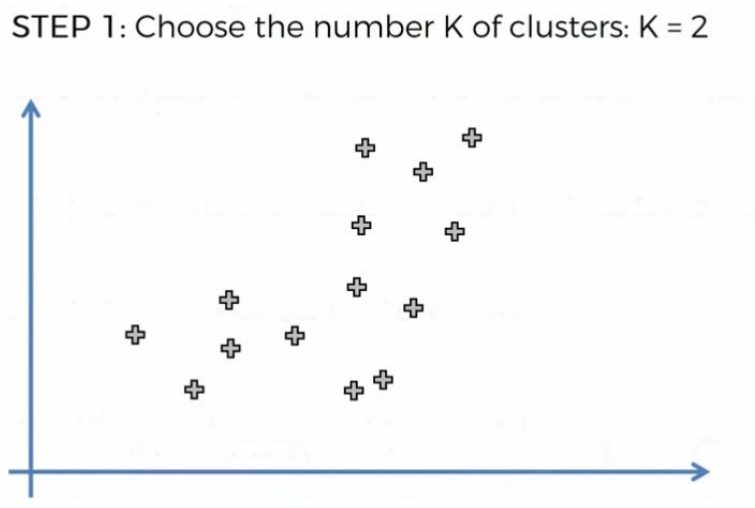
Una vez que se encuentren los clústeres, producto de asignar puntos de datos al centroide más cercano

5. Reasignar cada punto de datos al nuevo centroide

Aqui se realiza el paso 3 nuevamente solo que lo enumeramos como 5 y si tiene lugar una nueva asignación, repita el paso 4



Entendamos esto a través de un ejemplo visual



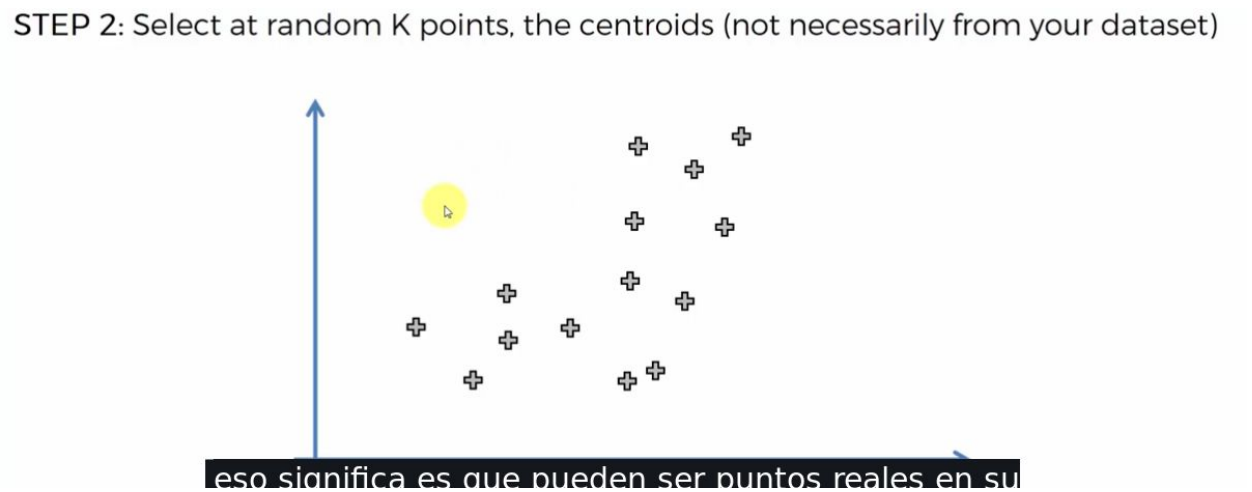
NUestros datos están ubicados contra dos variables y de inmediato, la primera pregunta es ¿puedes visualmente identificar rápidamente los clusters finales que crees que terminarán formándose? Es algo dificil, incluso solo con dos variables, imaginemos una situación compleja de 3 0 5 variables, tendríamos que pensar en como trazar un diagrama de dispersión de cinco dimensiones como ese.

Entonces, aquí es donde entra en juego K-Means clustering y es en donde este algoritmo nos ayudará a simplificar el proceso. Veremos como funciona, realizando manualmente ese mismo algoritmo de agrupamiento de clústeres. Entonces:

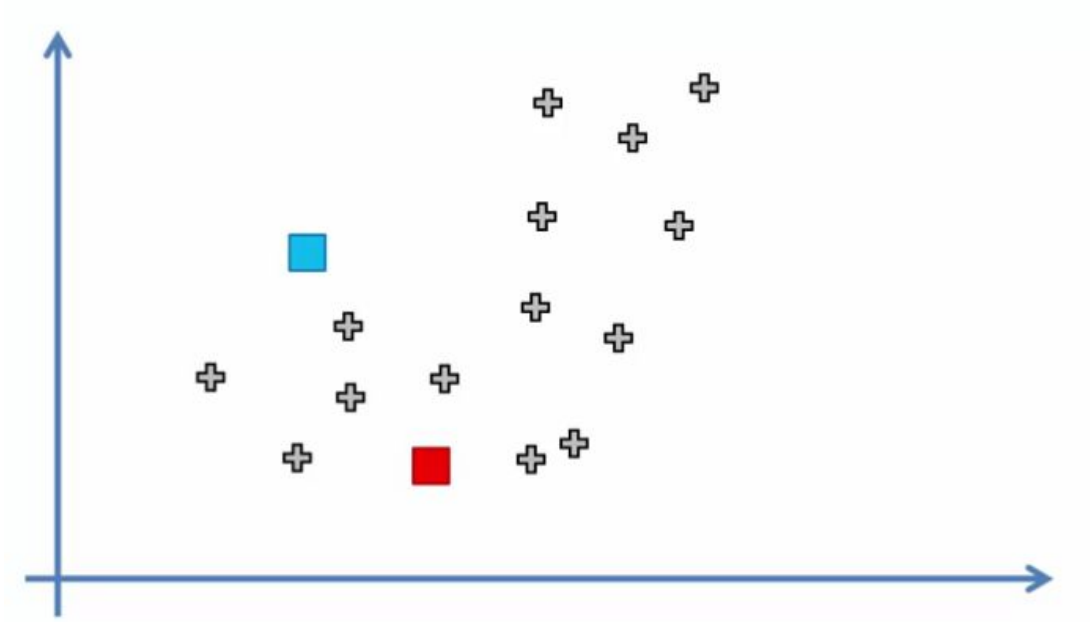
PASO 1: Escogiendo el número K de clústers

Identificamos DE ALGUNA MANERA que la cantidad óptima de clústeres es K=2

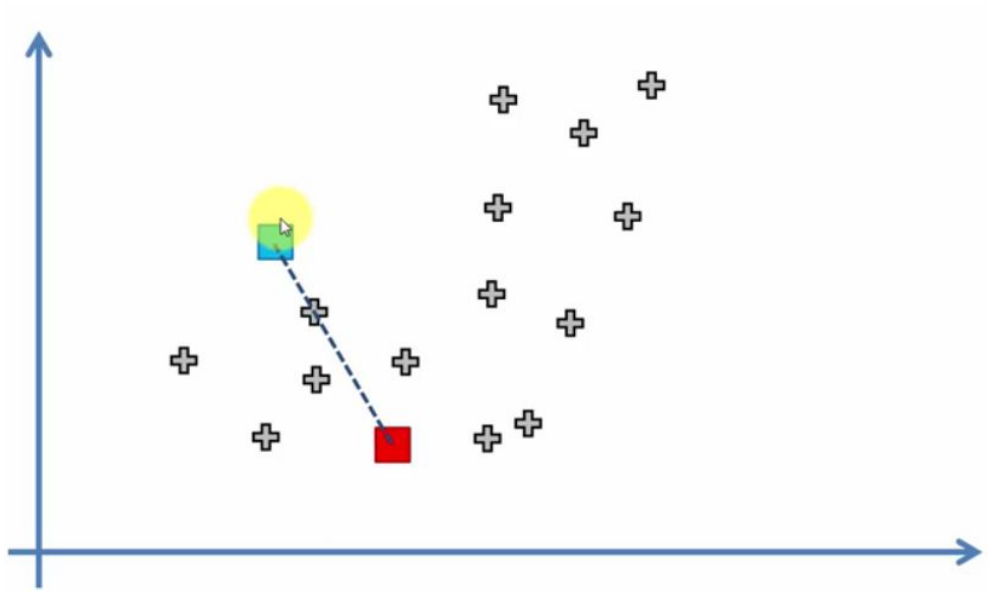
PASO 2: Escoger aleatoriamente K puntos que serán los centroides de los clústers. No necesariamente deben ser de mi conjunto de datos, lo que significa que pueden ser puntos reales en nuestro conjunto de datos o simplemente pueden ser puntos aleatorios en nuestro diagrama de dispersión



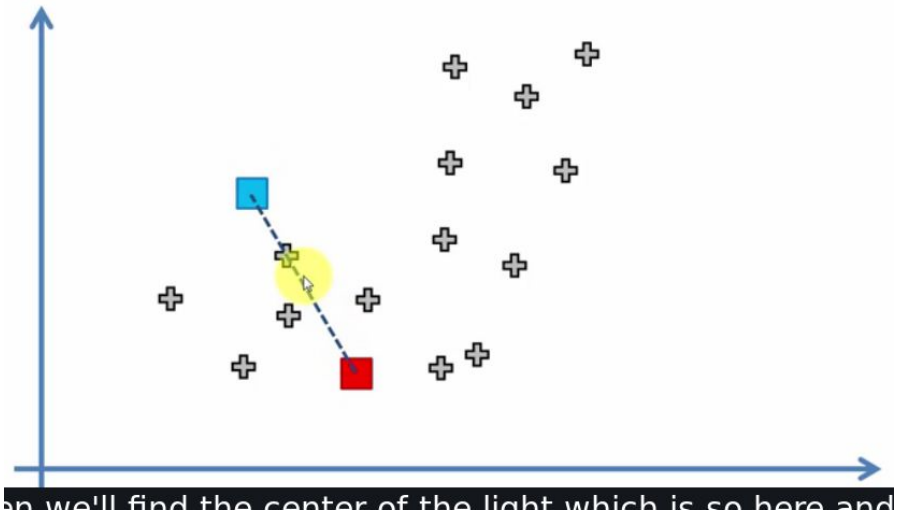
PASO 3 Asignar cada punto de datos al centroide más cercano
Entonces seleccionamos estos dos centroides, uno azúl y uno rojo y se asignarán cada punto de datos al centroide más cercano, por lo que hay que identificar para cada punto de datos cual de los dos centroides es el más cercano



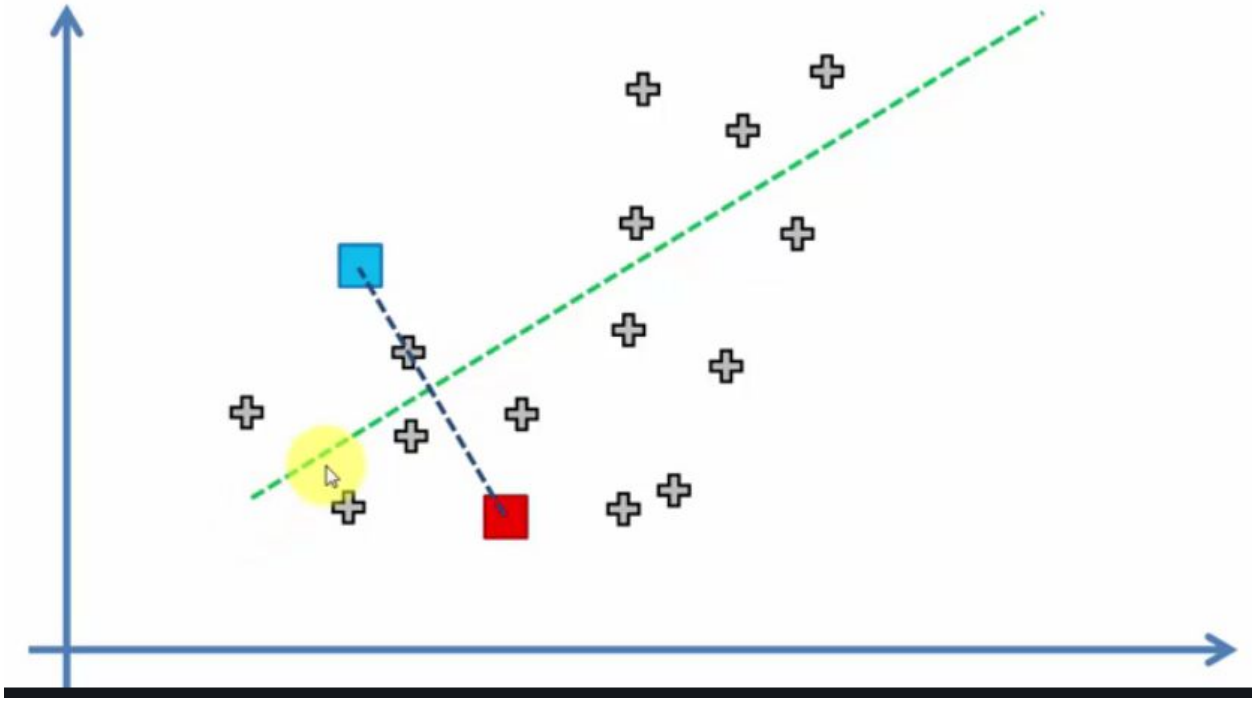
Usaremos un truco rápido aquí y es algo que aprendimos en geometría, y es conectar estos dos centroides con una línea como esta:



Y encontramos el centro de los dos centroides conectados, la luz amarilla que es esta:

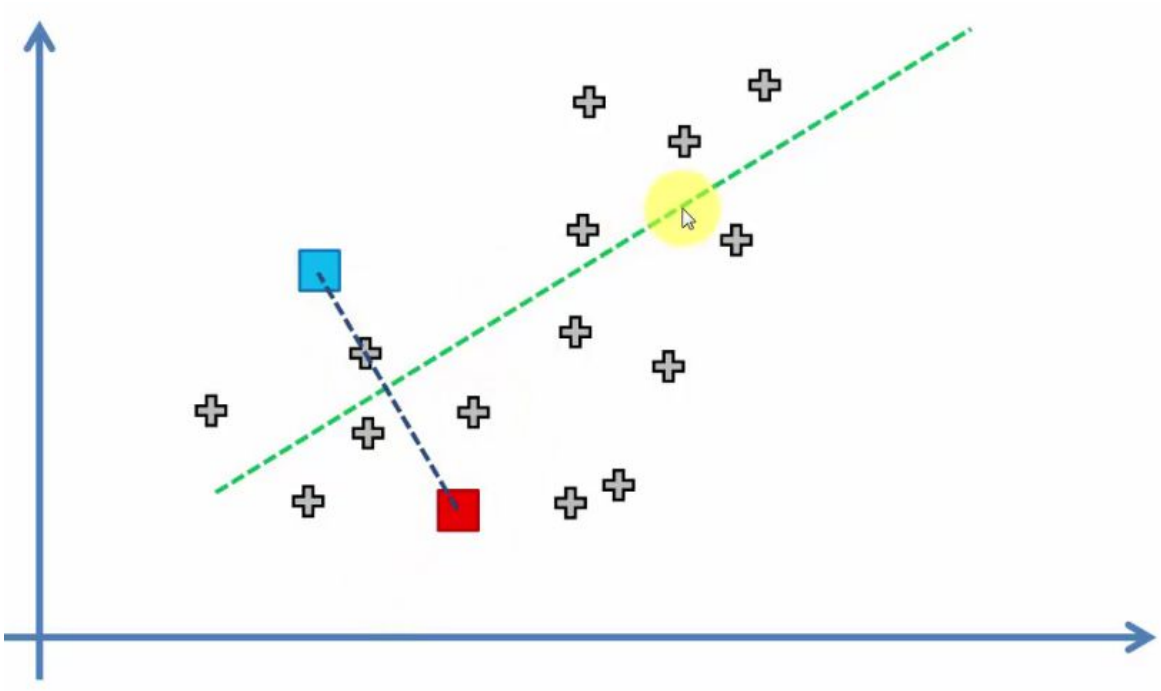


Y colocaremos una luz perpendicular exactamente que atraviese ese centro



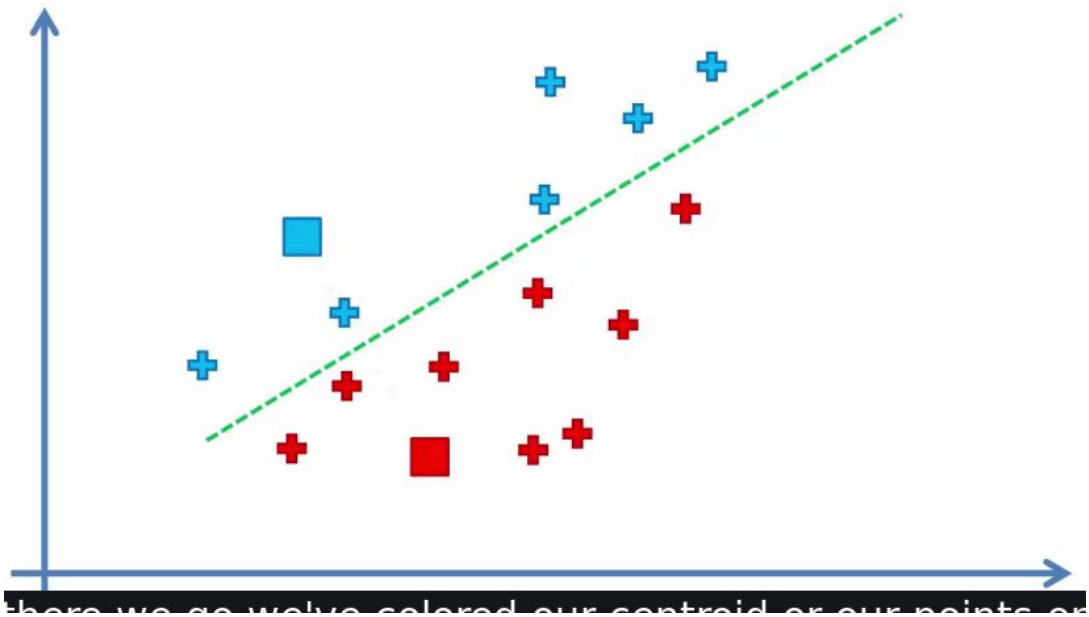
El concepto es que cualquier punto sobre esta línea verde es equidistante del azúl y del rojo .

Asi que si tomo este punto, será la misma distancia al centroide azul y al rojo.



Esta linea entera es equidistante a los centroides rojo y azúl.
Y analizado esto, es obvio que cualquiera de nuestros puntos sobre el diagrama de dispersión arriba de la linea verde esta más cerca al centroide azúl y cualquiera de los puntos debajo de la linea verde es más cerca al centroide rojo. ESta es una forma muy rápida de colorearlos e identificar el centroide más cercano.

Entonces coloreamos nuestros puntos de datos acorde a su centroide más cercano



Quiero mencionar algo aquí que el término "está más cerca" es muy parejo aunque parece sencillo, es un término bastante ambiguo porque "está más cerca" cuando estás visualizando cosas en un diagrama de dispersión. Sí, es bastante sencillo que sea la distancia

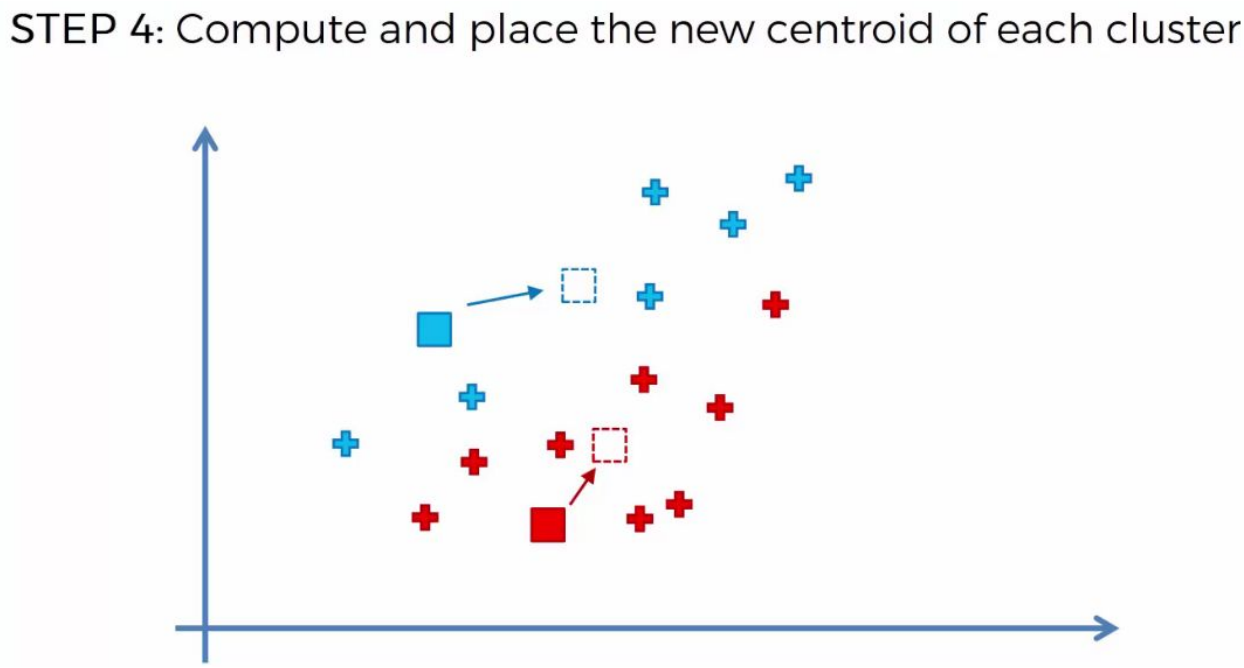
En Matematicas y CIencia de datos existe un monton de diferentes tipos de distancia, como la que estamos usando, que es la Euclidiana. Y la pregunta es:

¿ Deberíamos usar distancias Euclidianas o debo usar otro tipo de distancias definidas para mi problema ?
Eso es algo que uno mismo decide y existe algo que me permite especificar al algoritmo que tipo de distancia va a utilizar
Por nuestro bien usaremos distancias Euclidianas para propósitos de ilustración, Y, básicamente, esas son las distancias geométricas simples muy simples entre dos puntos diferentes.

Para algunos problemas es posible que queramos usar otras distancias y no las Euclidianas, asi que si se desea investigar más sobre las distancias, y que tipo de distancias se pueden usar. Con esto en mente, pasemos al paso 4

4. Calcular y ubicar el lugar del nuevo centroide para cada clúster

Así que, básicamente, ahora mismo tenemos los nuevos puntos: los antiguos, los azules, excluyendo el centroide mismo y todos los rojos, excluyendo el centroide mismo

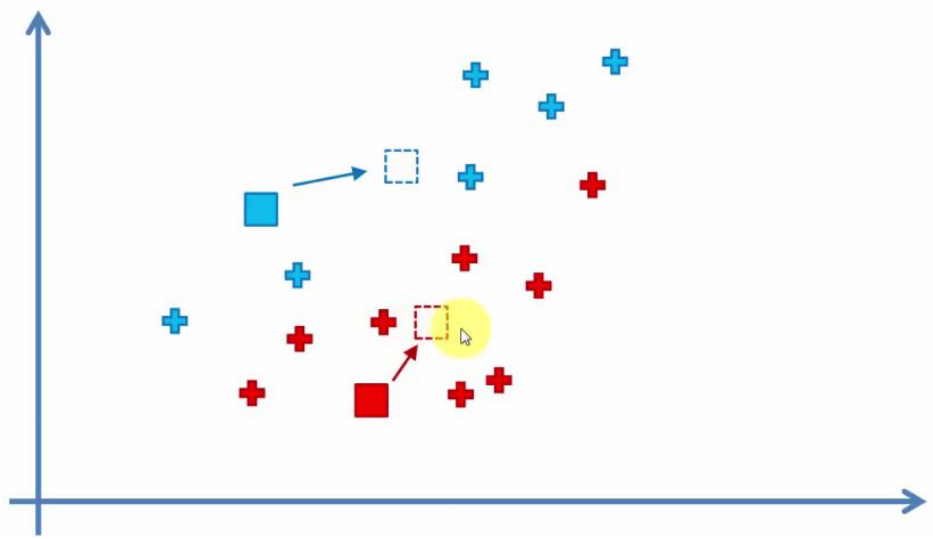


Y necesitamos encontrar el nuevo centroide para los puntos azules y el nuevo centroide para los puntos rojos, y averiguar en donde esta el centroide.

La forma de averiguar esto, es que el centroide no tiene peso, pero estos puntos azules y rojos de datos tienen un cierto peso. Digamos un kilo cada uno.

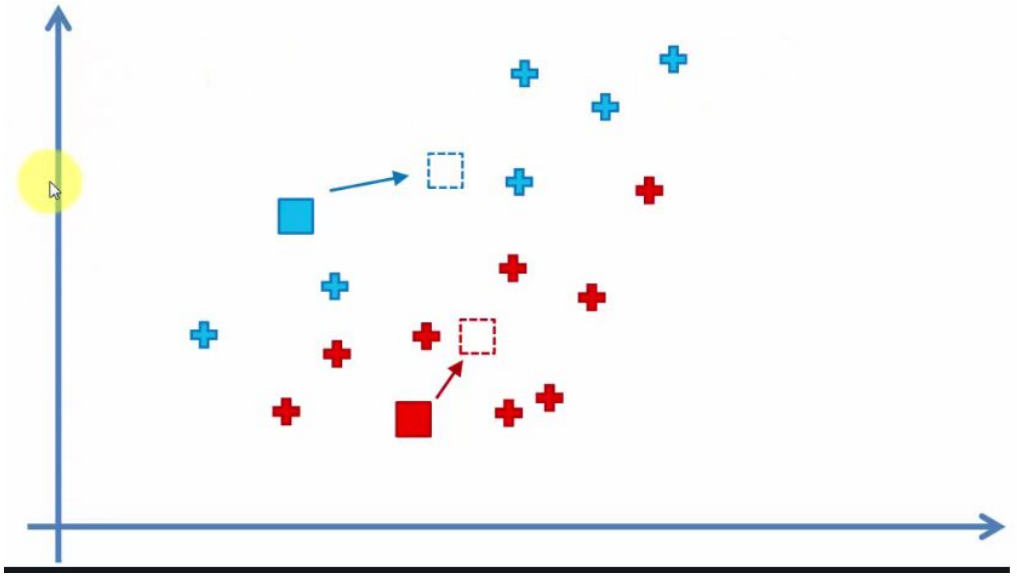
Lo que necesitamos hacer es encontrar el centro de la masa o el centro de gravedad de ese centroide y necesitamos dibujarlo en nuestro diagrama de dispersión.

Para el centroide azul sera aqui y para el centroide rojo será aquí:

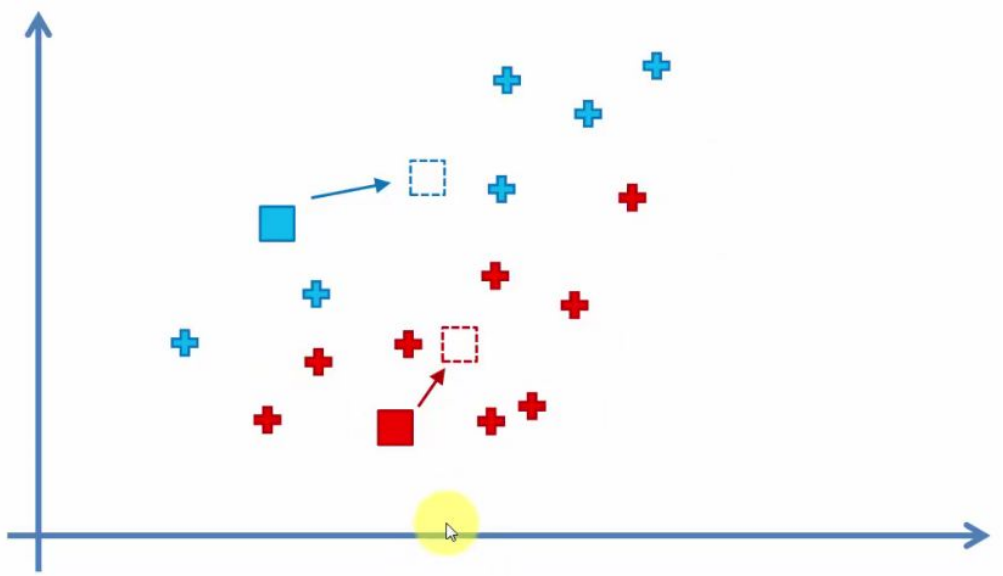


Y la forma de pensar sobre esto, en un diagrama de dispersión bidimensional puede ser ver visualmente donde esta.

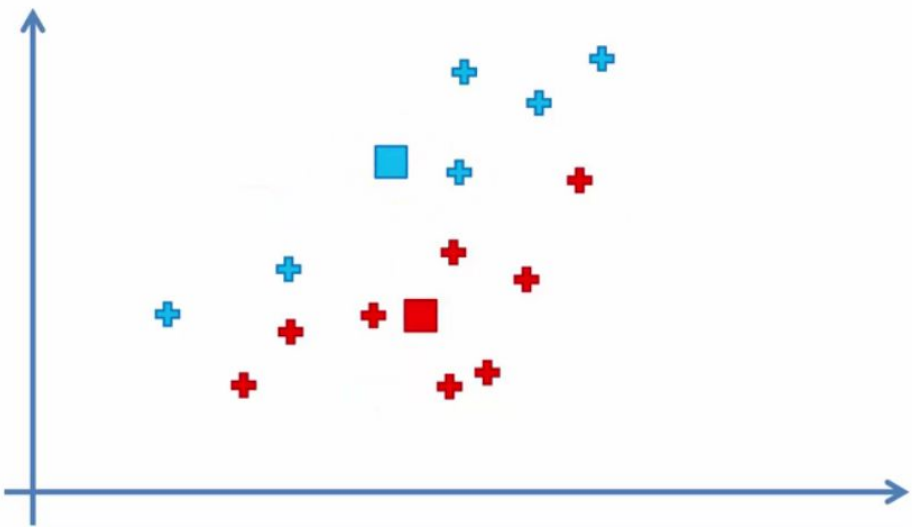
O simplemente se puede mirar las coordenadas X, Y para todos los puntos azúles y encontrar el centro de gravedad para las coordenadas Y que sería aquí:



O el centro de gravedad para las coordenadas X que sería aquí:



O el promedio de la esquina X y es asi como se obtiene los nuevos centroides para los datos azul y rojo

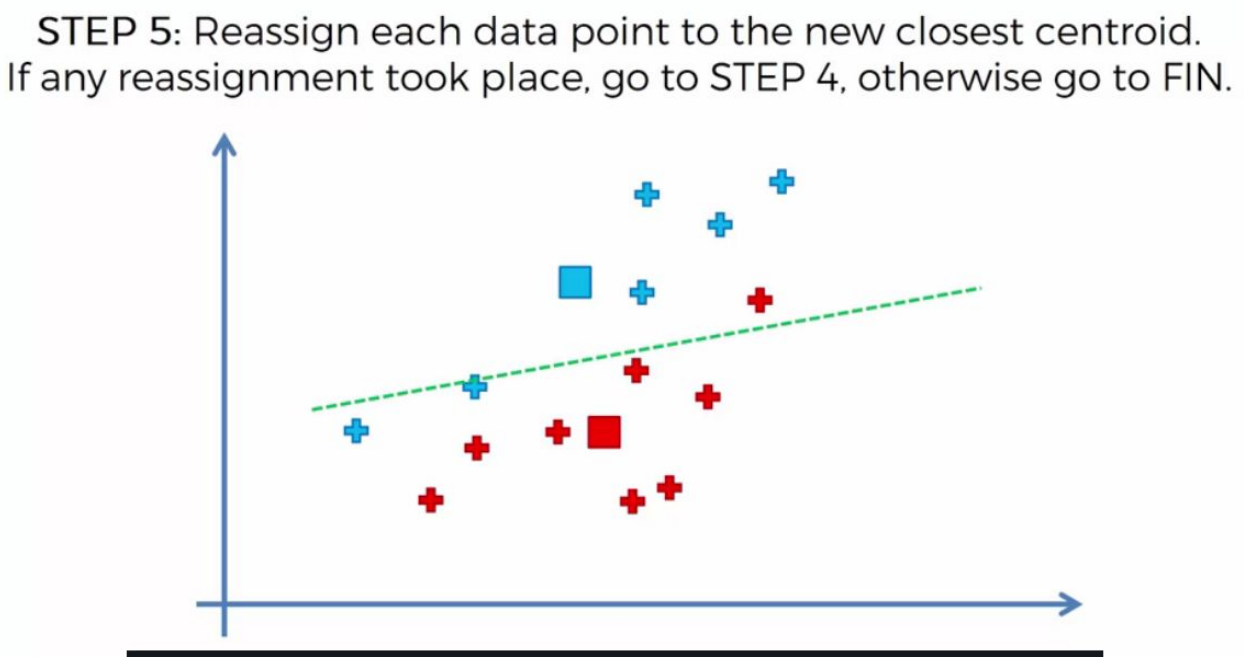


PASO 5: Reasignar cada punto de datos al nuevo centroide más cercano

Es repetir el paso 3 y 4 hasta que no se muevan más centroides, lo que significa que nuestro algoritmo ha convergido.

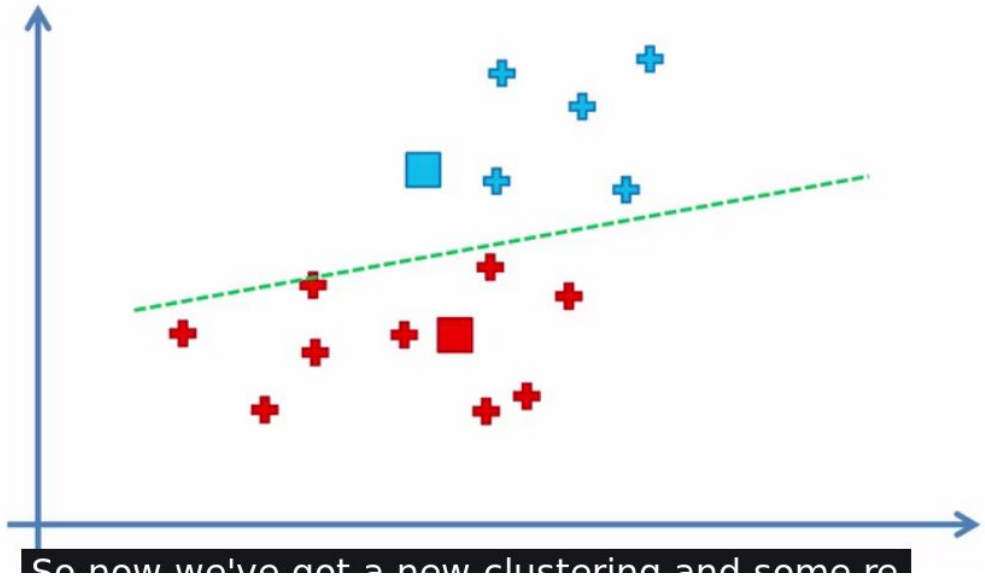
MIremos como ahora los puntos de datos son reasignados a los nuevos centroides

Coloca os la linea a través de nuestro diagrama de dispersión, y vemos que:

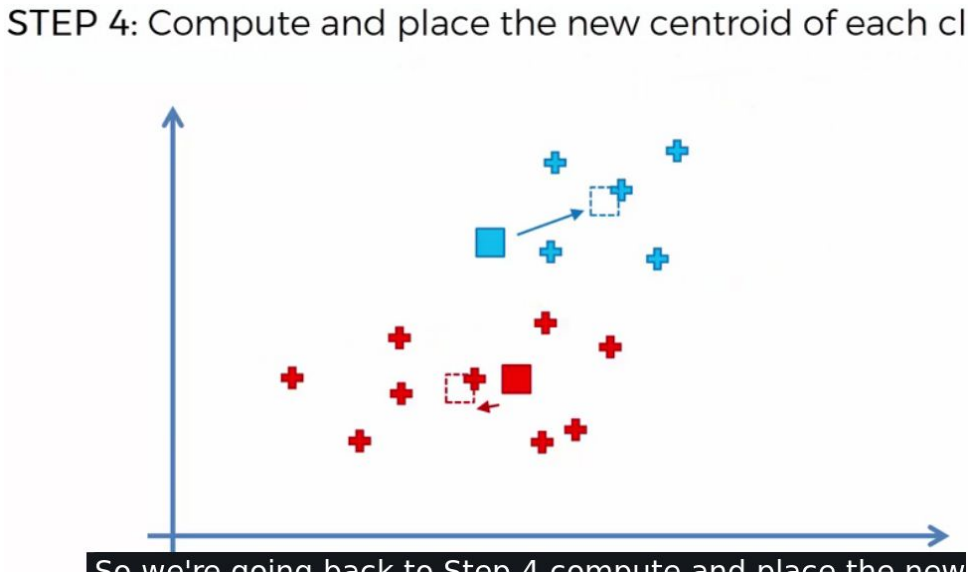


Vemos que existe un punto en ella, en realidad tres puntos entre rojos y azules que deben ser reasignados. UNo para el centroide azul y dos para el centroide rojo.

Asi que los recoloreamos

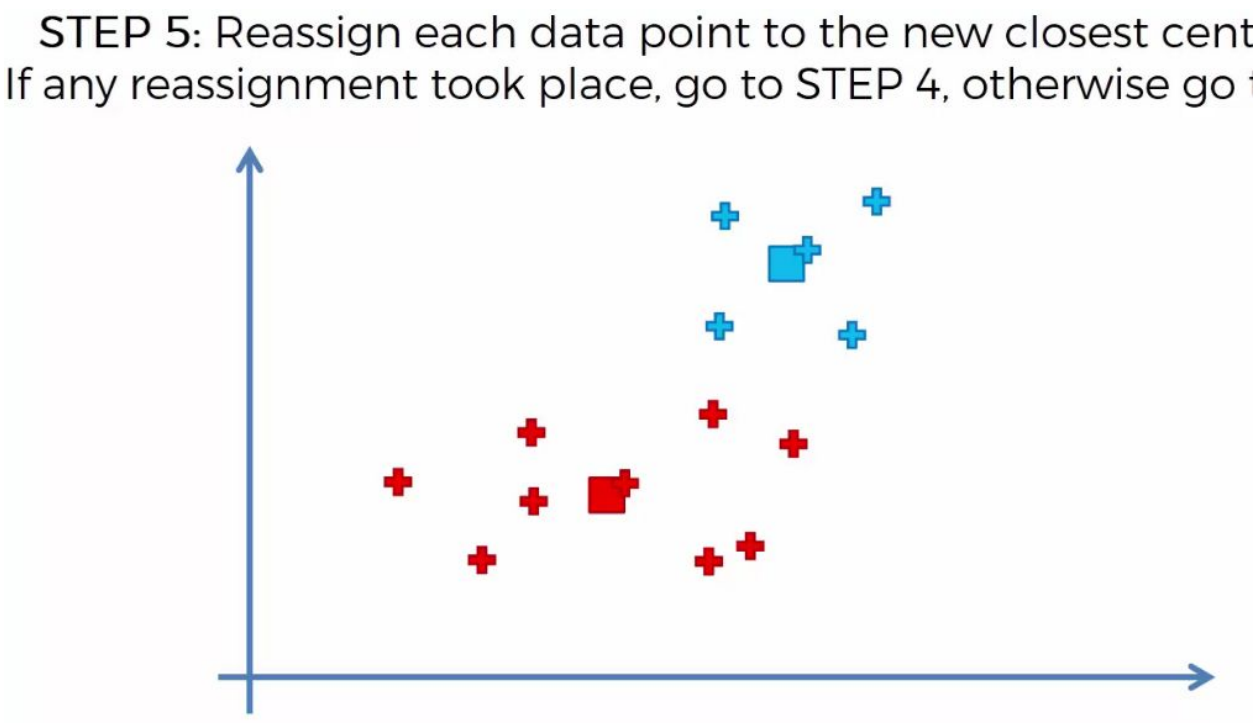


Volvemos al paso 4. Calculamos los nuevos centroides:
Encontrando los centros de masa o gravedad para cada centroide,

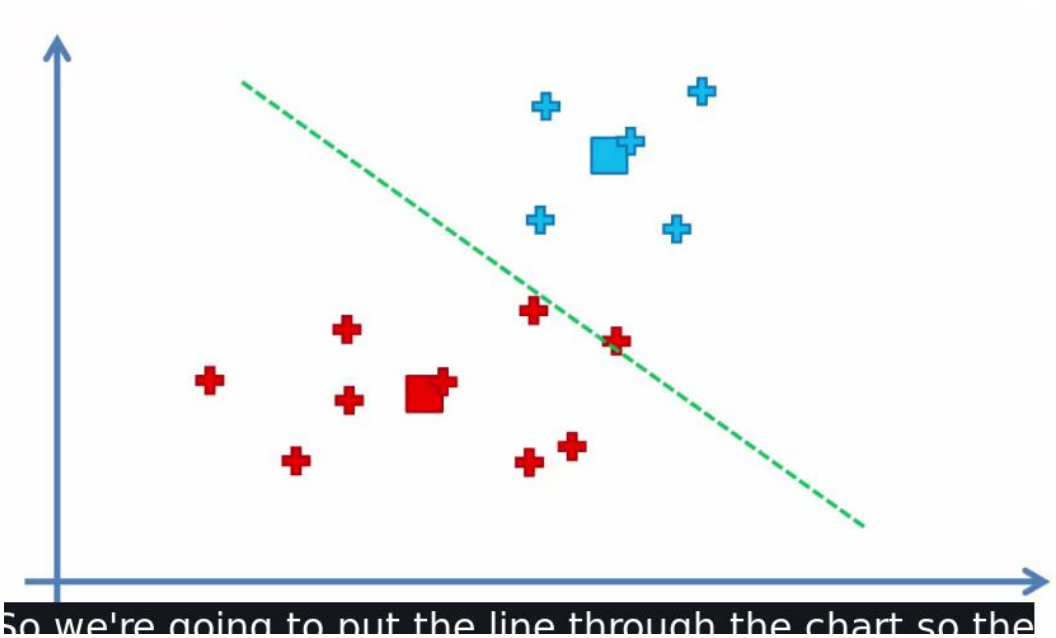


ubicando el centroide en esas nuevas locaciones

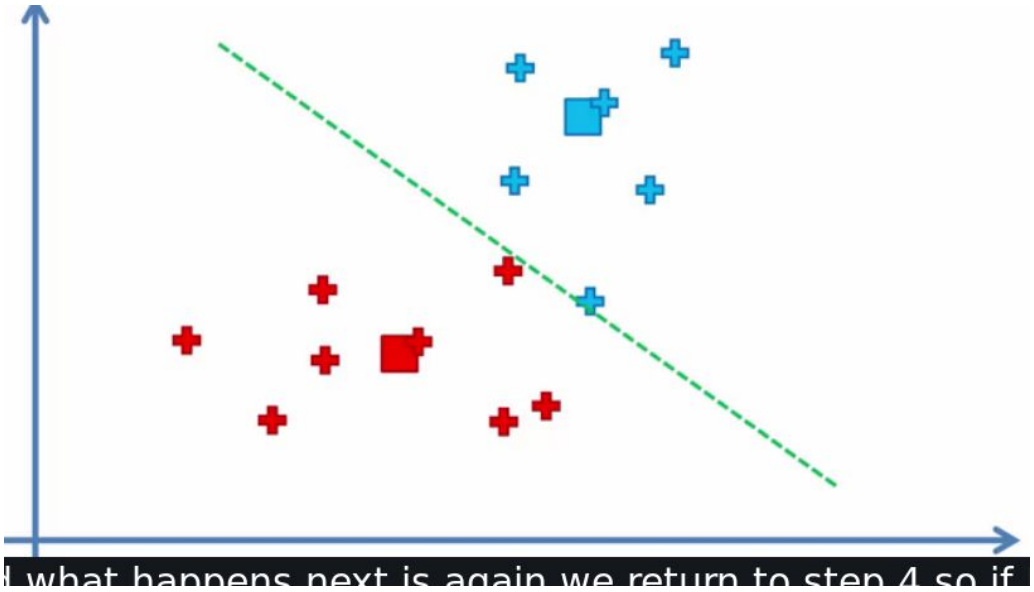
Y ahora repetir el paso 5, re asignar o mover los centroides



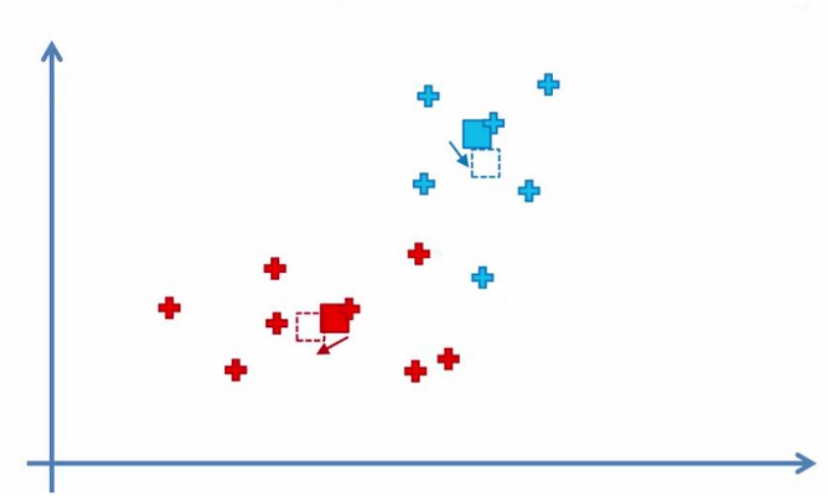
DE Nuevo trazamos la linea verde equidistante para ver que puntos de datos estan más cerca a esa nueva ubicación de los centroides



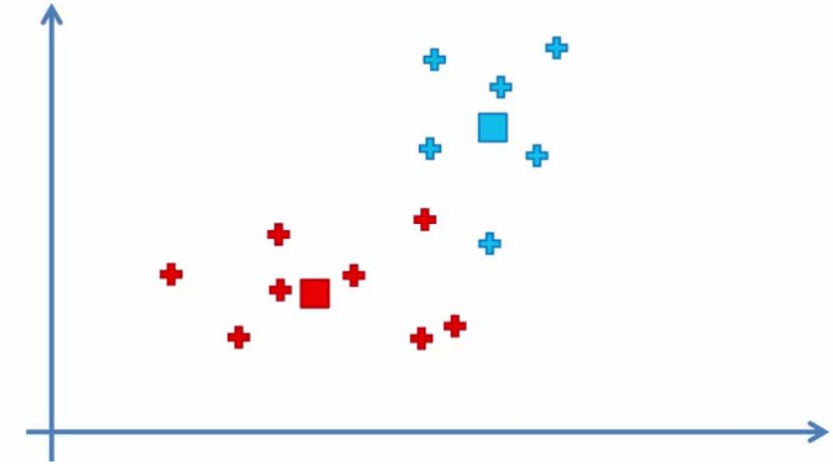
Vemos que hay puntos d edatos que deben ser reasignados



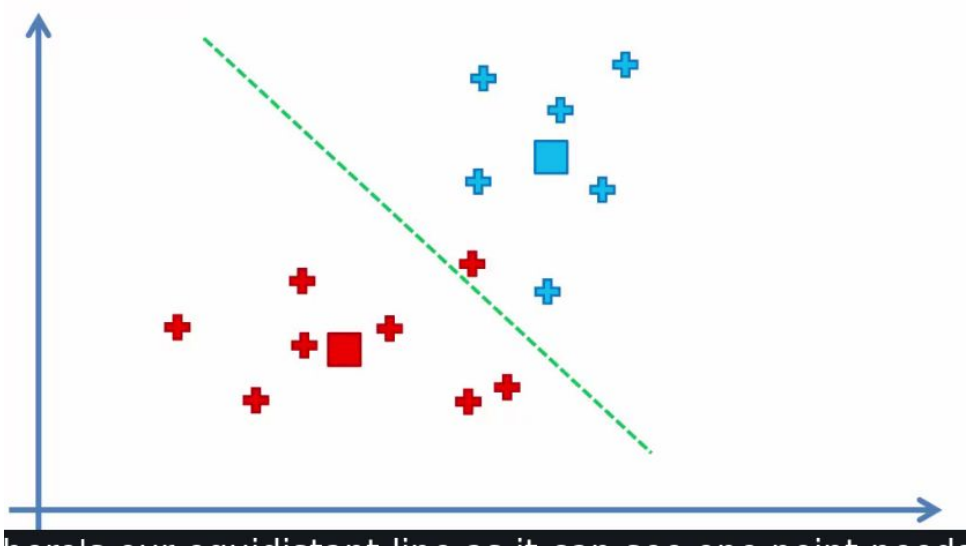
Retornamos al paso 4 y calculamos un nuevo lugar para el centroide basado en el centro de masa o gravedad de los datos.



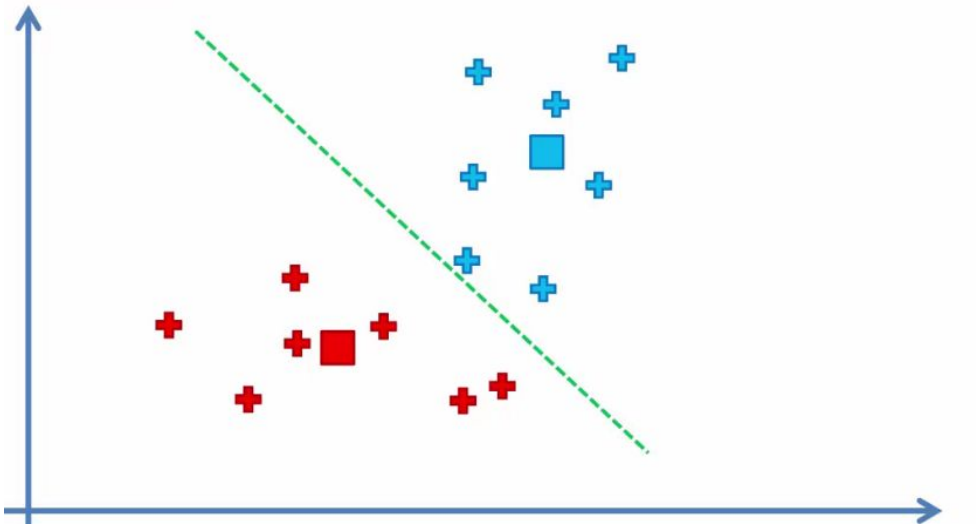
Se mueven los centroides



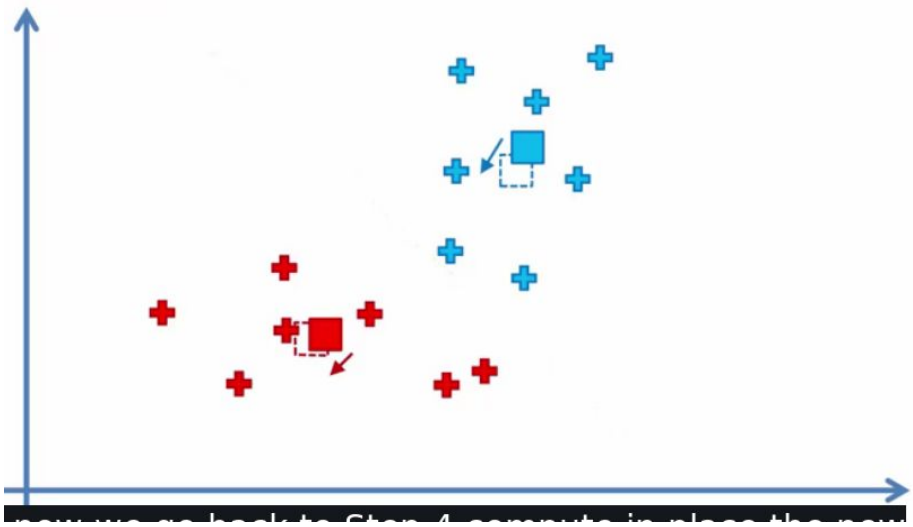
Trazamos la line equidistante para mirar si tenemos puntos de datos por reasignar.



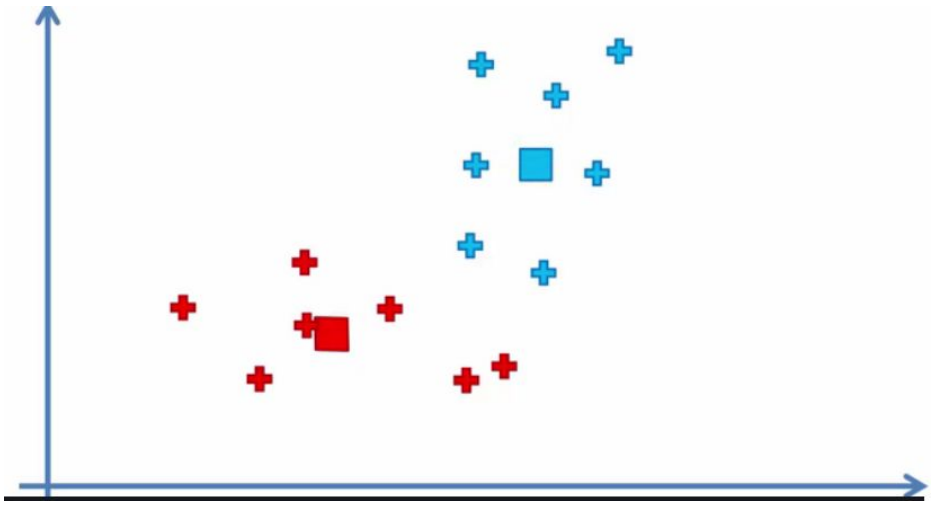
Reasignamos esos puntos al centroide azul



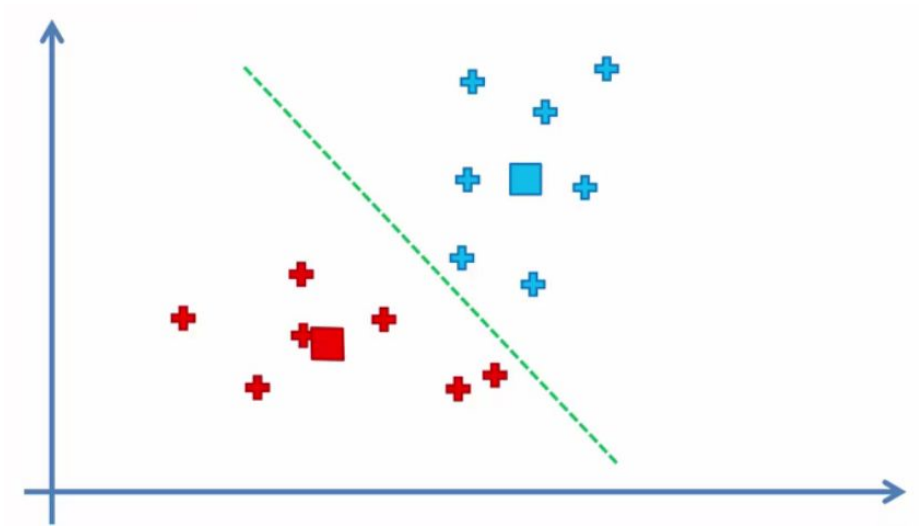
Calculamos un nueov centro de gravedad para los datos y tener una nueva locación de los centroides



Los movemos



Trazamos la linea y vemos que no hay puntos de datos por asignar



Esto quiere decir que todos los puntos estan en sus correctos clústers y esto significa que no hay reasingnacion de ellos por ende no cambiará el centro de gravedad de los datos y por ende no se moverán más los centroiides, lo que significa que nuestro algoritmo ha convergido.

FIN: Your Model Is Ready