DATA SCIENCE

Leandro Ferrado, Javier Lezama, Valentina Rubiolo

ACÁMICA

2 de Julio de 2019

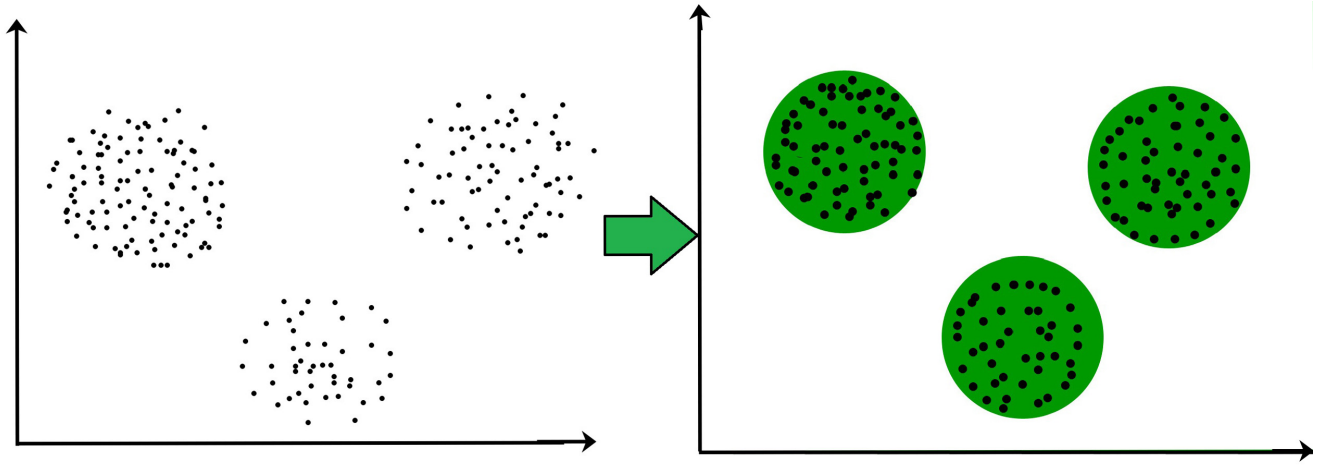
CLUSTERING

- El objetivo es agrupar los datos que presenten ciertas semejanzas entre sus miembros, es decir que se 'parezcan'.

CLUSTERING

- El objetivo es agrupar los datos que presenten ciertas semejanzas entre sus miembros, es decir que se 'parezcan'.
- También buscamos los datos que pertenezcan a grupos diferentes tengan rasgos lo suficientemente diferentes entre si.

CLUSTERING



CLUSTERING

- Los clusters deben ser identificables y de tamaño considerable.

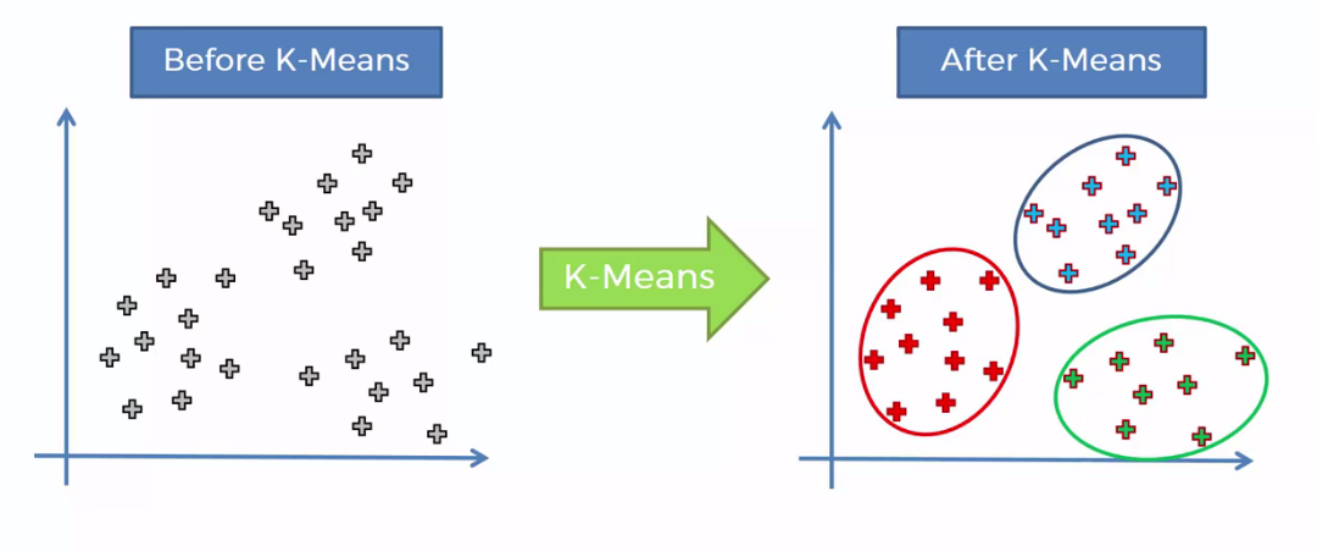
CLUSTERING

- Los clusters deben ser identificables y de tamaño considerable.
- Los puntos de un mismo cluster deben ser compactos y tener intersección mínimacon cualquier otro cluster.

CLUSTERING

- Los clusters deben ser identificables y de tamaño considerable.
- Los puntos de un mismo cluster deben ser compactos y tener intersección mínimacon cualquier otro cluster.
- Los cluster deben tener sentido desde el contexto del análisis. Los puntos de un mismo cluster deben tener propiedades comunes en el contexto estudiado.

CLUSTERING



Tipos de clustering

- Jerarquicos.

Tipos de clustering

- Jerarquicos.
- Planos.

K-means

- es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano.

K-means

- El algoritmo trabaja iterativamente para asignar a cada «punto» (las filas de nuestro conjunto de entrada forman una coordenada) uno de los «K» grupos basado en sus características. Son agrupados en base a la similitud de sus features (las columnas)

K-means

- Como resultado de ejecutar el algoritmo tendremos:

K-means

- Como resultado de ejecutar el algoritmo tendremos:
- Los «centroids» de cada grupo que serán unas «coordenadas» de cada uno de los K conjuntos que se utilizarán para poder etiquetar nuevas muestras.

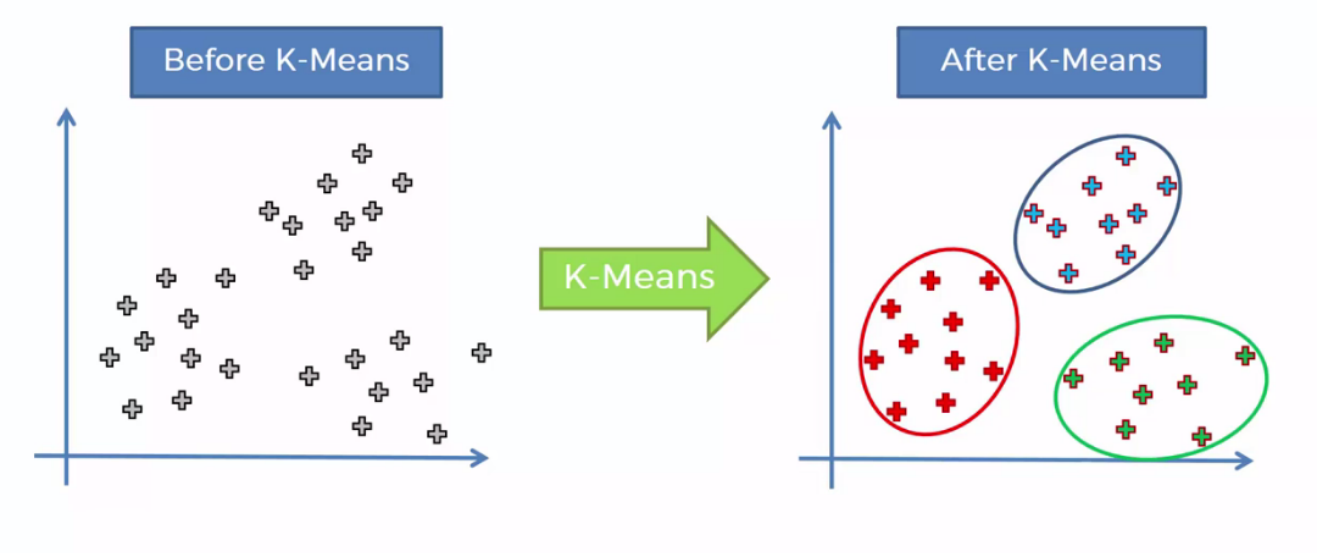
K-means

- Como resultado de ejecutar el algoritmo tendremos:
- Los «centroids» de cada grupo que serán unas «coordenadas» de cada uno de los K conjuntos que se utilizarán para poder etiquetar nuevas muestras.
- Etiquetas para el conjunto de datos de entrenamiento. Cada etiqueta perteneciente a uno de los K grupos formados.

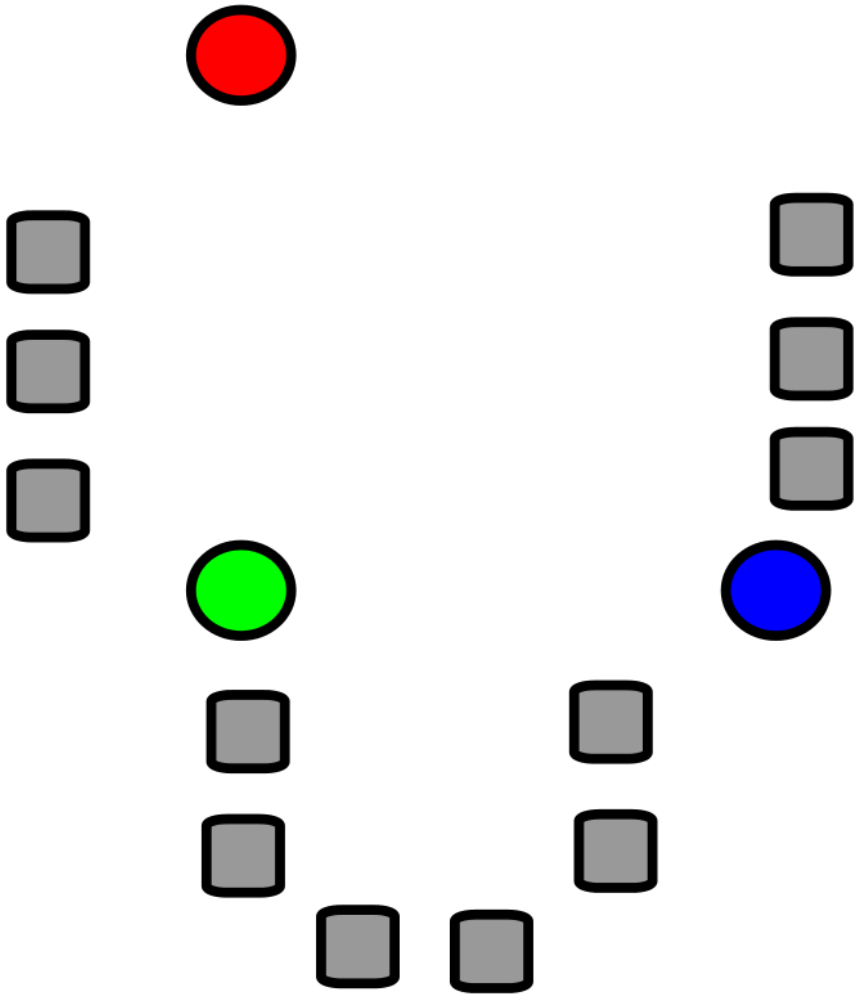
K-means

- Los grupos se van definiendo de manera «orgánica», es decir que se va ajustando su posición en cada iteración del proceso, hasta que converge el algoritmo. Una vez hallados los centroids deberemos analizarlos para ver cuales son sus características únicas, frente a la de los otros grupos. Estos grupos son las etiquetas que genera el algoritmo.

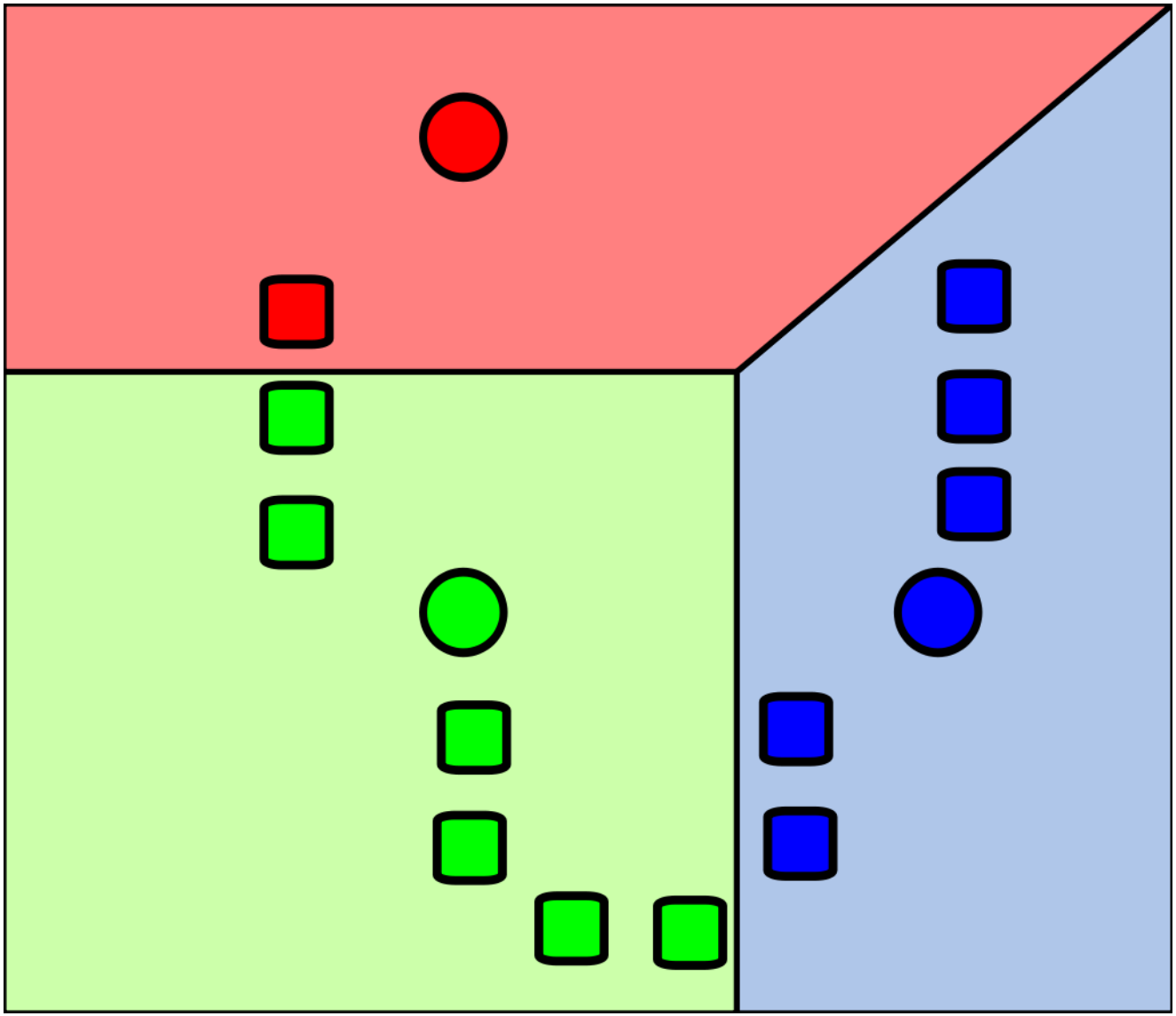
K-means



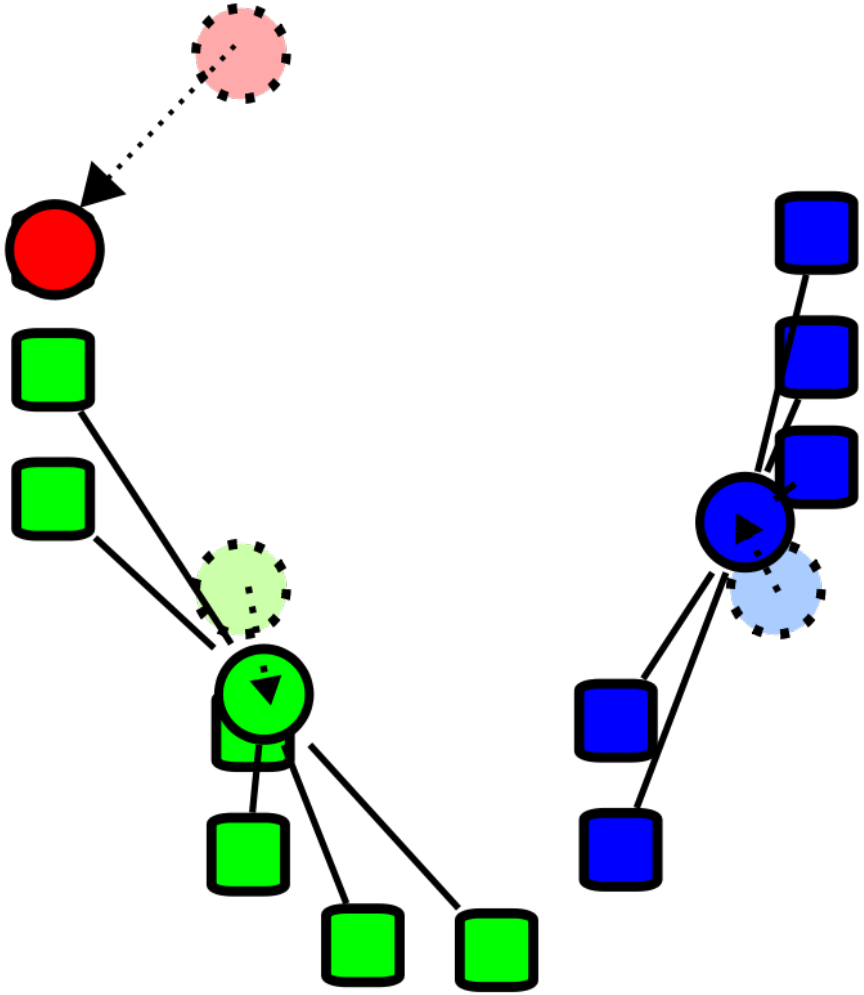
K-means



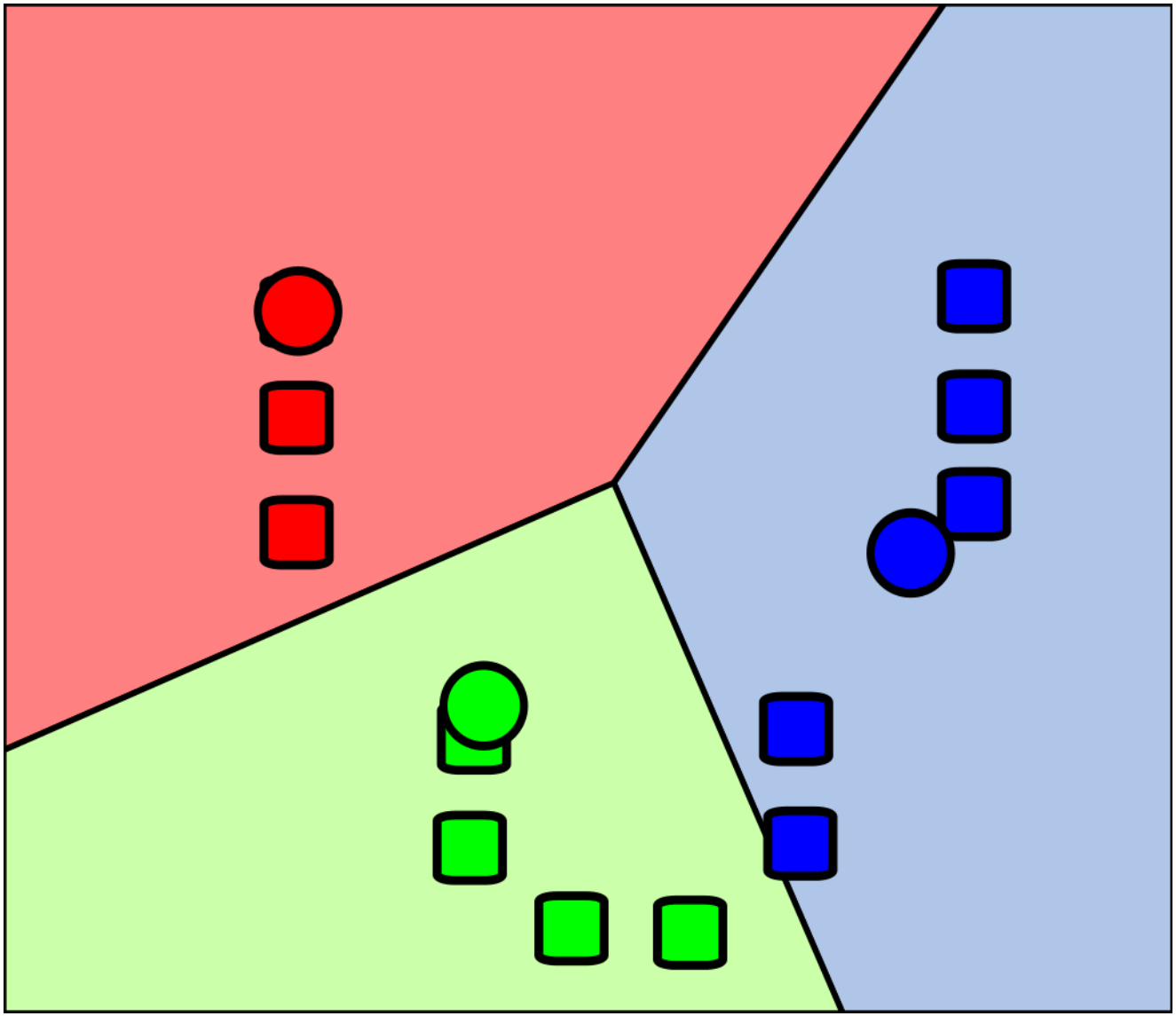
K-means



K-means



K-means



El Algoritmo K-means

- El algoritmo utiliza una proceso iterativo en el que se van ajustando los grupos para producir el resultado final. Para ejecutar el algoritmo deberemos pasar como entrada el conjunto de datos y un valor de K. El conjunto de datos serán las características o features para cada punto. Las posiciones iniciales de los K centroids serán asignadas de manera aleatoria de cualquier punto del conjunto de datos de entrada. Luego se itera en dos pasos:

El Algoritmo K-means

- 1- Paso de Asignación de datos

El Algoritmo K-means

- 1- Paso de Asignación de datos

El Algoritmo K-means

- 1- Paso de Asignación de datos

En este paso, cada «fila» de nuestro conjunto de datos se asigna al centroide más cercano basado en la distancia cuadrada Euclideana. Se utiliza la siguiente fórmula (donde dist() es la distancia Euclideana standard):
$min_{ci \in C} dist(c_i, x)^2$

El Algoritmo K-means

- 2-Paso de actualización de Centroid

El Algoritmo K-means

- 2-Paso de actualización de Centroid

El Algoritmo K-means

- 2-Paso de actualización de Centroid

En este paso los centroid de cada grupo son recalculados. Esto se hace tomando una media de todos los puntos asignados en el paso anterior.
$c_i = \frac{1}{S_i} \sum_{x_i \in S_i} x_i$

El Algoritmo K-means

- El algoritmo itera entre estos pasos hasta cumplir un criterio de detención:

El Algoritmo K-means

- El algoritmo itera entre estos pasos hasta cumplir un criterio de detención:
- si no hay cambios en los puntos asignados a los grupos,

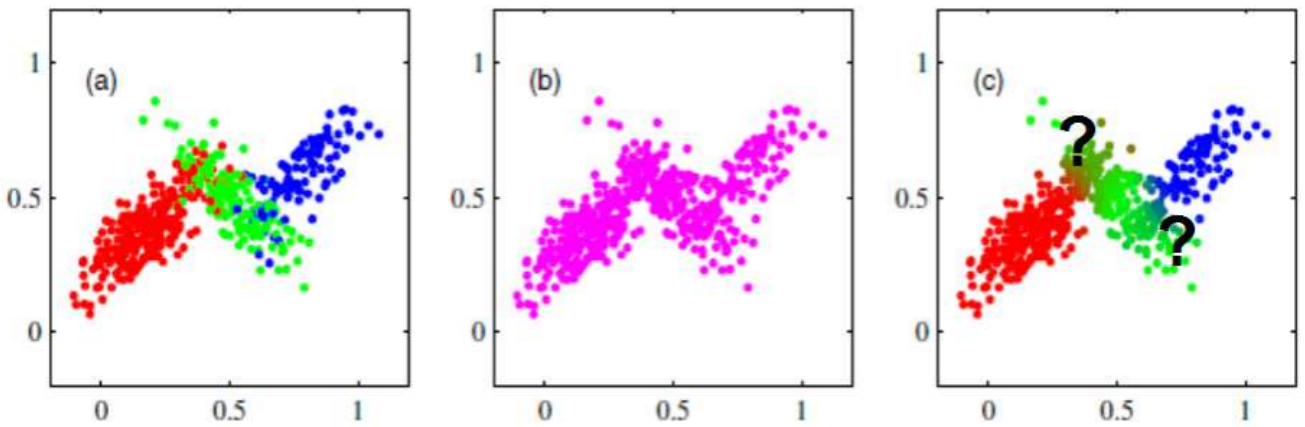
El Algoritmo K-means

- El algoritmo itera entre estos pasos hasta cumplir un criterio de detención:
- si no hay cambios en los puntos asignados a los grupos,
- o si la suma de las distancias se minimiza,

El Algoritmo K-means

- El algoritmo itera entre estos pasos hasta cumplir un criterio de detención:
- si no hay cambios en los puntos asignados a los grupos,
- o si la suma de las distancias se minimiza,
- o se alcanza un número máximo de iteraciones.

K-means



El Algoritmo K-means

- Gaussian Mixture Models(GMM):

El Algoritmo K-means

- Gaussian Mixture Models(GMM):
- Mean shift