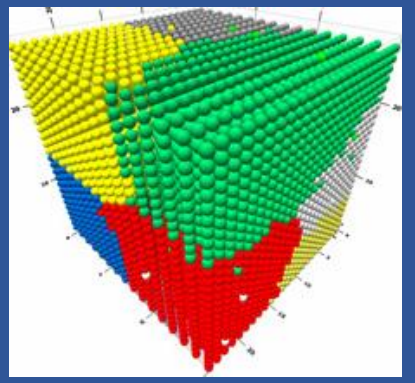
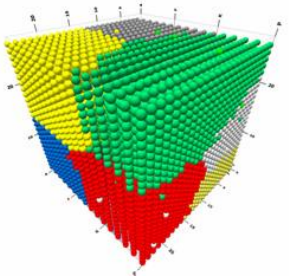
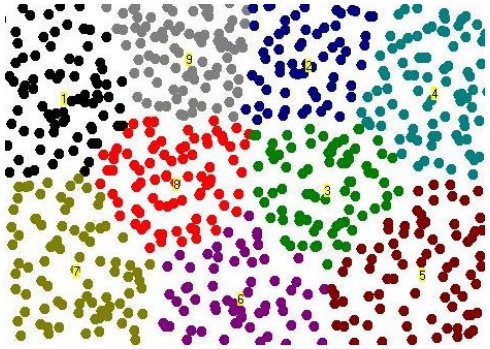
Modelos de Agrupamiento

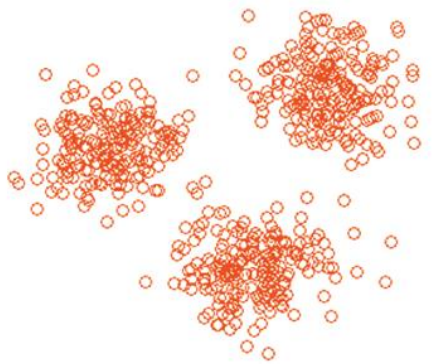
Clustering



Agrupamiento



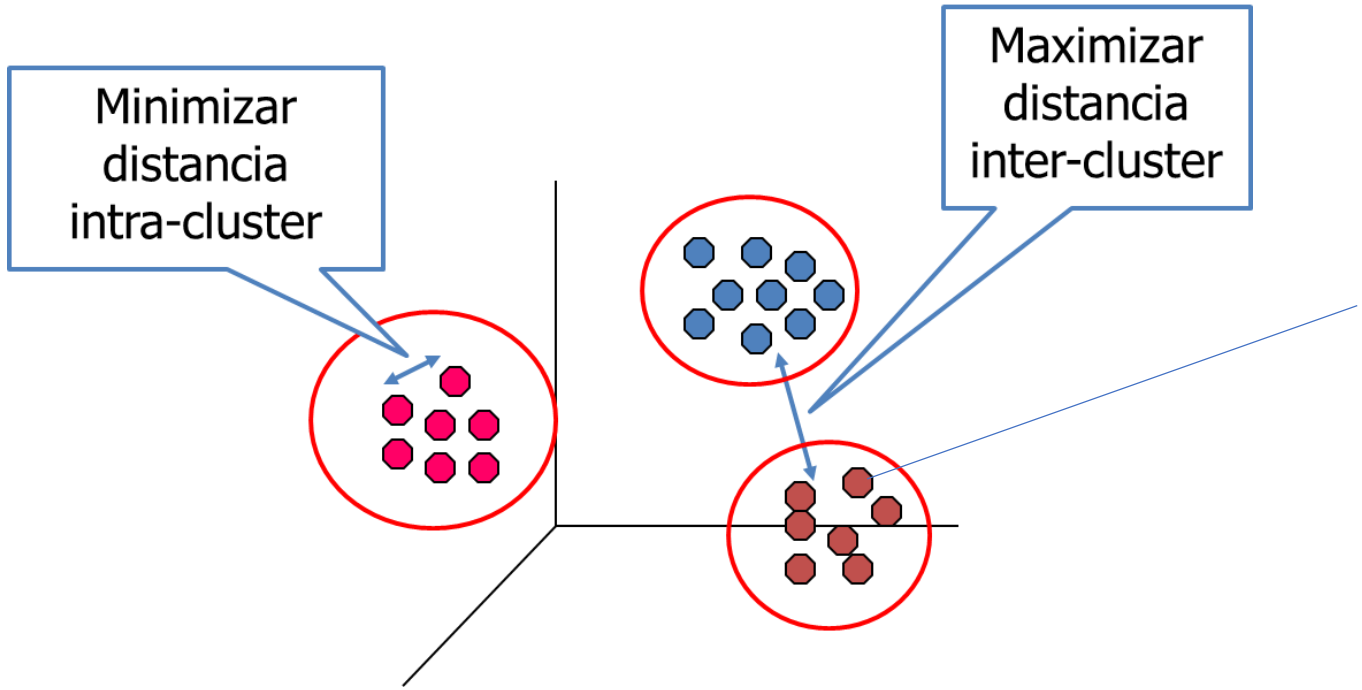




- Se basa en intentar responder cómo es que ciertas instancias (instancias, casos, vectores) pertenecen o "caen" naturalmente en cierto número de clases o grupos, de tal manera que estas instancias compartan ciertas características.
- Es un método de agrupación de una serie de instancias (puntos) dados en un espacio multidimensional de acuerdo con un criterio de cercanía.
- La cercanía se define en términos de una determinada función de distancia
- Las instancias de un mismo grupo tienen propiedades comunes.

Agrupamiento

- Se trata de encontrar agrupamientos de tal forma que los objetos de un grupo sean similares entre sí y diferentes de los objetos de otros grupos

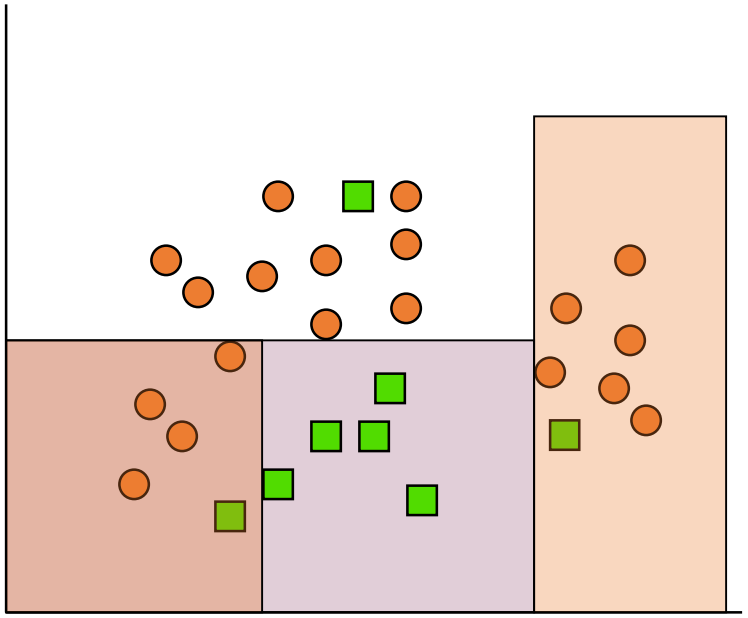


Una **instancia** (un conjunto finito de atributos)

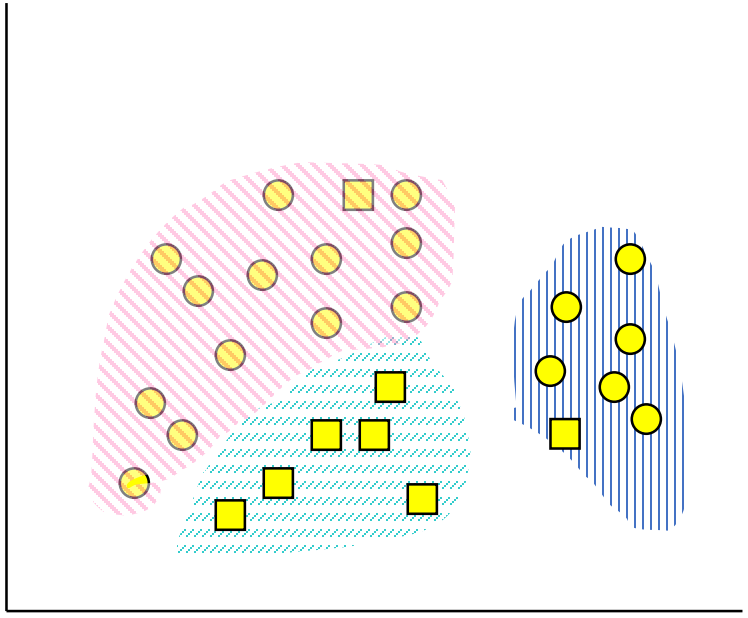
- Numéricas: son números reales en general
- Nominales : Son variables discretas pero que no tienen un orden especificado (color de ojos)
- Ordinales: Son variables discretas con una relación de orden (Alta, Media, baja)
- Binarias: solo pueden tomar dos estados posibles (dicotómicas)

Clasificación y Agrupamiento

Clasificación
- Aprendizaje Supervisado
- Aprende un método para predecir las clases de instancias



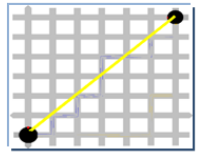
Agrupamiento
- Aprendizaje no Supervisado
- Encuentra grupos naturales de instancias dado instancias no clasificadas



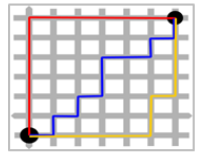
Medidas de Similaridad

- Se usan para confirmar la similaridad entre las instancias que pertenecen a un grupo y la diferencia que existe entre los grupos.
  - Una columna numérica A
    - Distancia(X, Y) = A(X) – A(Y)
  - Varias columnas numéricas:
    - Distancia(X, Y) = distancia euclidiana entre X, Y
  - Columnas nominales:
    - 1 si son diferentes, 0 sin son iguales.
  - Todos los atributos son importantes?
    - Ponderación podría ser necesario.

- Euclídea $d_{ij} = \sqrt{\sum_{k=1}^{p} W_k \left(x_{ik} - x_{jk}\right)^2}$



- City-Block $d_{ij} = \sum_{k=1}^{p} W_k \left|x_{ik} - x_{jk}\right|$

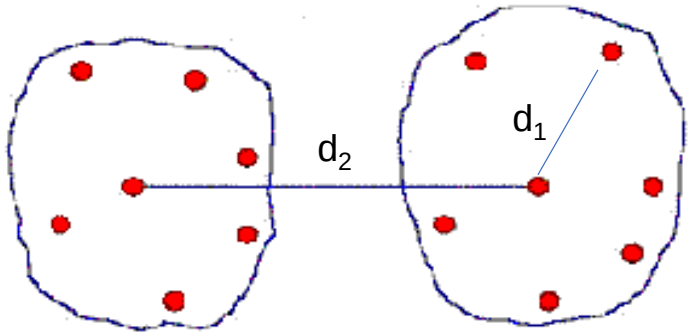


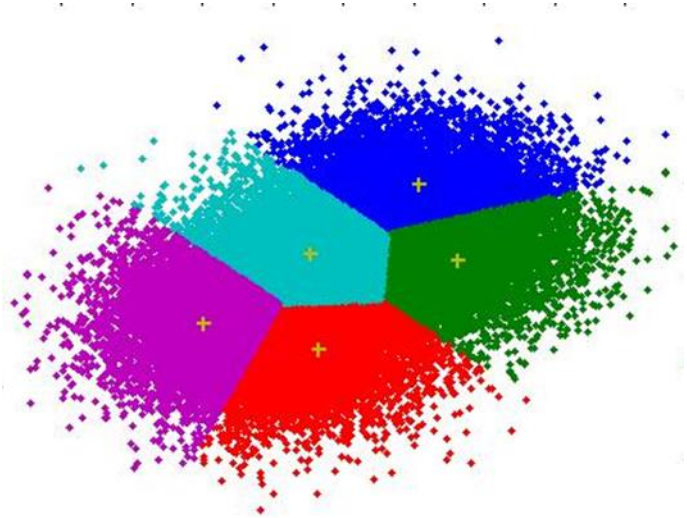
- Minkowski $d_{ij} = \sqrt[\lambda]{\sum_{k=1}^{p} W_k \left(x_{ik} - x_{jk}\right)^{\lambda}} \quad \lambda > 0$

- Coseno $d_{ij} = \dfrac{\sum_{k=1}^{p} x_{ik} \cdot x_{jk}}{\sqrt{\sum_{k=1}^{p} x_{ik}^2} \cdot \sqrt{\sum_{l=1}^{p} x_{jl}^2}}$

Evaluación de Resultados

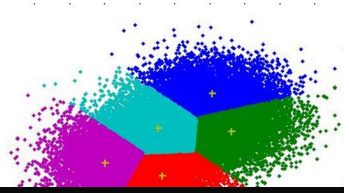
- No existe un indicador de qué tan bien están formados los grupos.
  - Inspección manual.
  - Comparación con etiquetas reales de cada instancia.
  - Medidas de calidad del grupo
    - Medidas de distancia
    - Alta similaridad dentro de un grupo (d_1), baja entre grupos (d_2).

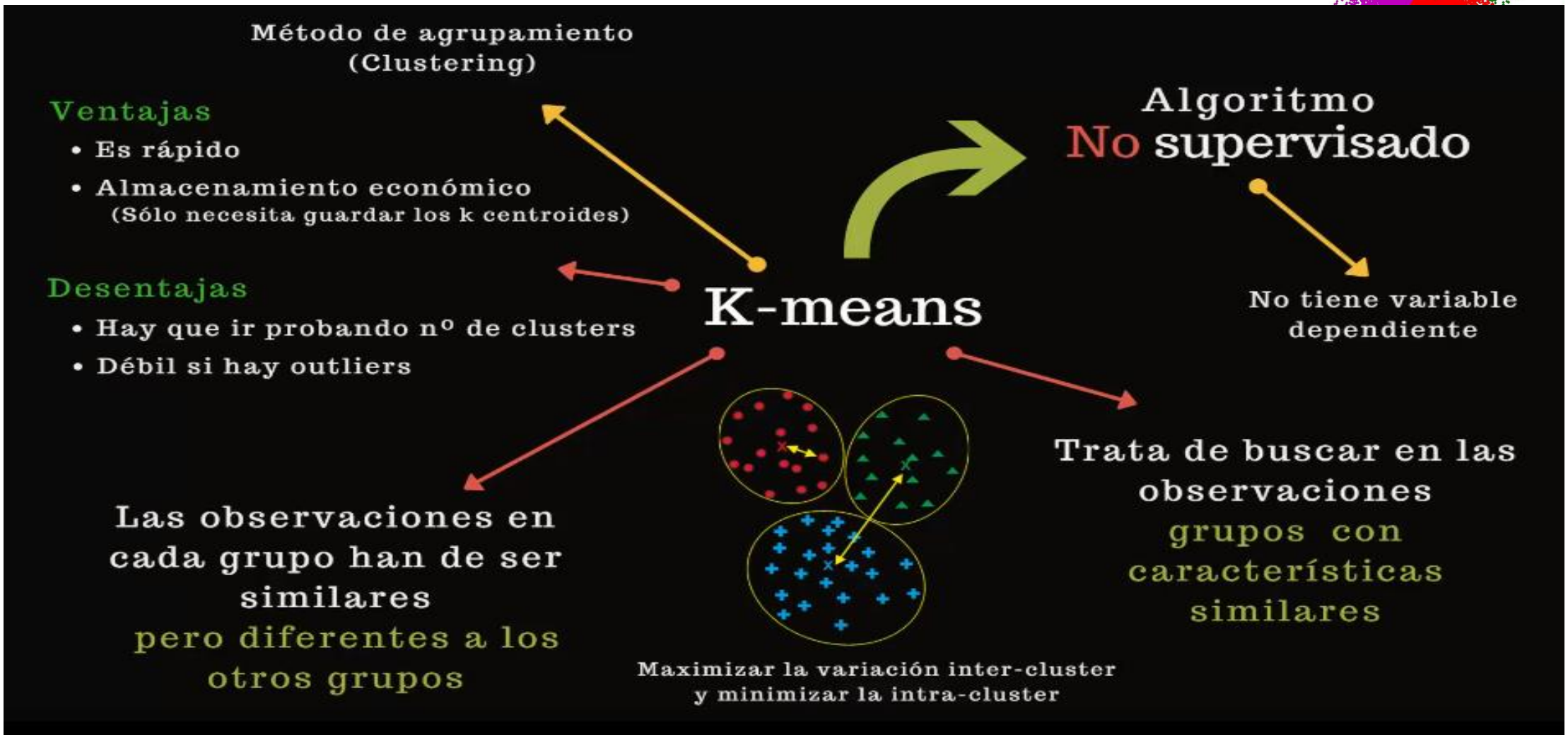




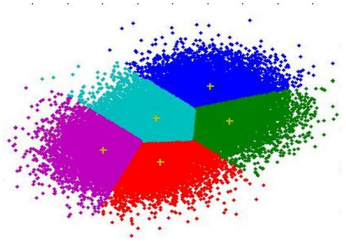
Cluster K-Means

K-Means




Método de agrupamiento
(Clustering)
Ventajas
• Es rápido
• Almacenamiento económico
(Sólo necesita guardar los k centroides)
Desentajas
• Hay que ir probando nº de clusters
• Débil si hay outliers
K-means
Algoritmo
No supervisado
No tiene variable
dependiente
Trata de buscar en las
observaciones
grupos con
características
similares
Las observaciones en
cada grupo han de ser
similares
pero diferentes a los
otros grupos
Maximizar la variación inter-cluster
y minimizar la intra-cluster

K-Means – Algoritmo



1. Seleccionar un número (K) de centros de cluster aleatoriamente (centroides iniciales de los K-grupos).

2. Calcular la distancia de las instancias a cada centroide.

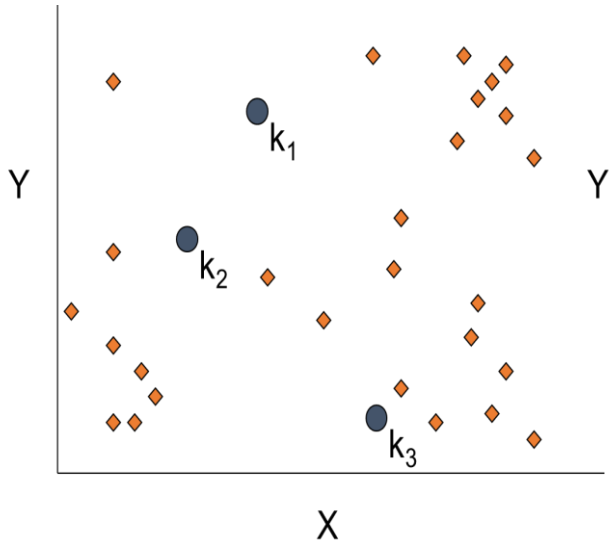
3. Asignar cada instancia al centro de cluster más cercano

4. Mover cada centro de cluster hacia la media de los instancias asignadas a dicho cluster.

5. Repetir pasos 2, 3 y 4 hasta converger (cambio en la asignación a clusters menor que un nivel predefinido)

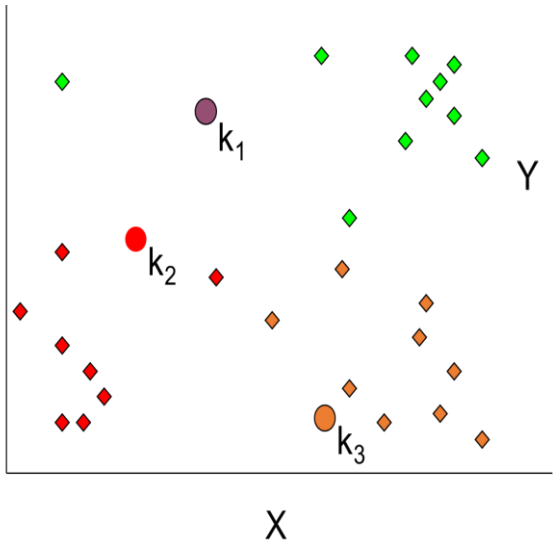
Ejemplo de K-means

paso 1



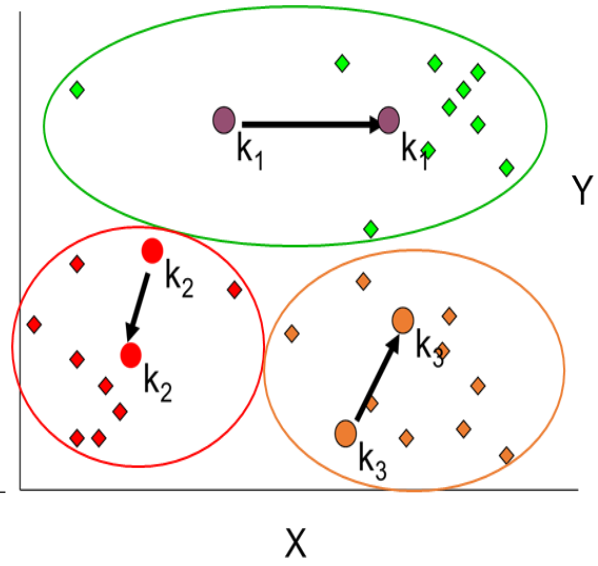
Selecciona 3 centros de clusters iniciales (aleatoriamente)

paso 2 Y 3



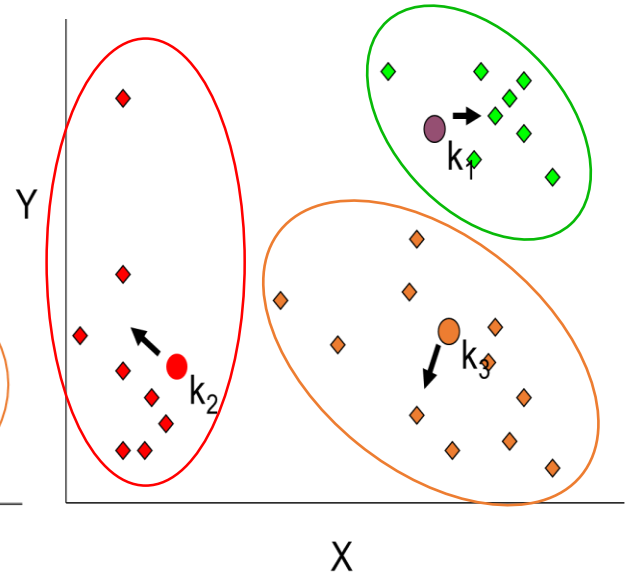
- Calcular la distancia de las instancias a cada centroide
- Asignar cada punto al centro de cluster más cercano

paso 4



Mover cada centro de cluster hacia su respectiva media

paso 2,3 y4



Reasignar los puntos a sus más cercanos centros de cluster

Re calcular las medias de cada cluster

Evaluación del Modelo