Tema 11

IV. Minería de datos descriptiva

Agrupamiento

Tema 11. Agrupamiento

1. *Clustering*/agrupamiento/segmentación
   - 1.1. Definición de *clustering*
   - *1.2. Clustering* versus clasificación
   - 1.3. Aplicaciones
   - 1.4. Bondad de un análisis *cluster*
   - 1.5. Propiedades deseables en un método de *clustering* en minería de datos
2. Medidas de distancia y similaridad
3. Distintas aproximaciones al *clustering*
4. Métodos basados en particionamiento
5. Métodos jerárquicos

1.1. Definición de *clustering*

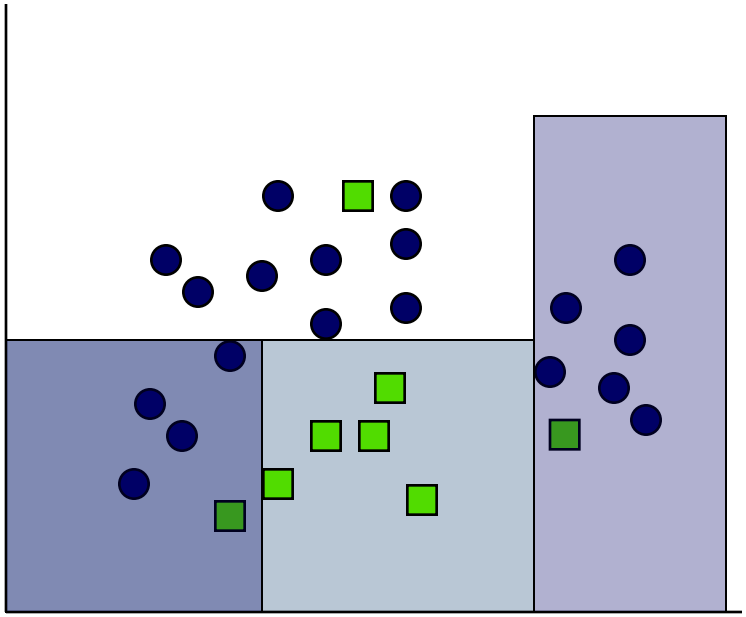
- *Cluster:* un grupo o conjunto de objetos
  - Similares a cualquier otro incluido en el mismo *cluster*
  - Distintos a los objetos incluidos en otros grupos
- *Clustering* (análisis *cluster*):
  - Segmentar una población heterogénea en un número de subgrupos homogéneos o *clusters*
- *Clustering* puede verse como clasificación no supervisada, las clases no están predefinidas
- Aplicaciones típicas:
  - Como una tarea de preprocesamiento antes de aplicar otra técnica de descubrimiento del conocimiento
  - Como técnica de descubrimiento del conocimiento para obtener información acerca de la distribución de los datos (p.e.: encontrar clientes con hábitos de compra similares)

1.1. Definición de *clustering*

Cuando se aplican algoritmos de *clustering* a problemas reales, nos enfrentamos a:

- Dificultad en el manejo de *outliers*
  - Se pueden ver como *clusters* solitarios
  - Se puede forzar a que estén integrados en algún *cluster* → suele implicar que la calidad de los *clusters* obtenidos es baja
- Si se realiza en BBDD dinámicas implica que la pertenencia a *clusters* varía en el tiempo
  - Los resultados del *clustering* son dinámicos
- Interpretar el significado de cada *cluster* puede ser difícil
- No hay una única solución para un problema de *clustering*. No es fácil determinar el número de *clusters*

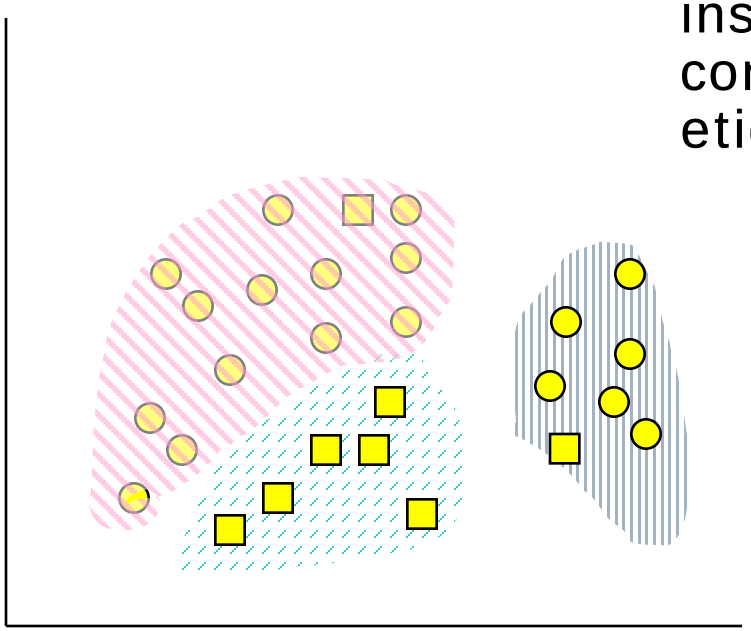
1.2. Clustering vs. clasificación



Clasificación: Aprendizaje supervisado:

Aprende, a partir de un conjunto de instancias pre-etiquetadas un metodo para predecir la clase a que pertenece una nueva instancia

1.2. Clustering vs. clasificación



Aprendizaje no supervisado:

Encuentra un agrupamiento de instancias “natural” dado un conjunto de instancias no etiquetadas

1.3. Aplicaciones

- Marketing: descubrimiento de distintos grupos de clientes en la BD. Usar este conocimiento en la política publicitaria, ofertas, …
- Uso de la tierra: Identificación de áreas de uso similar a partir de BD con observaciones de la tierra (cultivos, …)
- Seguros: Identificar grupos de asegurados con características parecidas (siniestros, posesiones, ….). Ofertarles productos que otros clientes de ese grupo ya poseen y ellos no
- Planificación urbana: Identificar grupos de viviendas de acuerdo a su tipo, valor o situación geográfica
- WWW: Clasificación de documentos, analizar ficheros .log para descubrir patrones de acceso similares, …

1.4. Bondad de un análisis *cluster*

- Un buen método de *clustering* debe producir *clusters* en los que:
  - Se maximize la similaridad *intra-cluster*
  - Se minimize la similaridad *inter-cluster*
- La calidad del *clustering* resultante depende tanto de la medida de similaridad utilizada como de su implementación
- Medidas de similaridad/disimilaridad: normalmente una función de distancia: d(i, j)
- Las funciones de distancia son muy sensibles al tipo de variables usadas, así su definición puede cambiar para variables: medidas por intervalos, booleanas, categóricas (nominales), ordinales, …
- Es posible dar peso a ciertas variables dependiendo de distintos criterios (relativos a su aplicación, …)
- En general, es complicado dar definiciones para términos como “suficientemente similar”, asi que algunas respuestas serán subjetivas y dependientes de umbrales

1.5. Propiedades deseables en un método de *clustering* en minería de datos

- Escalables
- Capacidad para tratar distintos tipos de variables
- Capacidad para descubrir *clusters* con formas arbitrarias
- Requisitos mínimos de conocimiento del dominio para determinar los parámetros de entrada
- Capacidad para tratar datos con ruido y *outliers*
- Insensible al orden de los registros de entrada
- Capacidad para incorporar restricciones del usuario
- Válido para registros de alta dimensionalidad
- Resultados interpretables

Tema 11. Agrupamiento

1. *Clustering*/agrupamiento/segmentación
2. Medidas de distancia y similaridad
3. Distintas aproximaciones al *clustering*
4. Métodos basados en particionamiento
5. Métodos jerárquicos

2. Medidas de distancia y similaridad

- La propiedad más importante que debe verificar un *cluster* es que haya más cercanía entre las instancias que están dentro del *cluster* que respecto a las que están fuera del mismo

- La definición de la medida de distancia depende normalmente del tipo de variable:

  - Variables intervalares
    - Variables continuas para las que se utiliza una discretización: peso, edad, ...
  - Variables binarias/boolenas
  - Variables nominales/categóricas
  - Variables ordinales
  - Variables mixtas

2. Medidas de distancia y similaridad

- El caso más simple: un único atributo numérico A

  Distancia(X,Y) = A(X) – A(Y)

- Varios atributos numéricos:
  - Distancia(X,Y) = Distancia euclídea entre X,Y

- Atributos nominales: La distancia se fija a 1 si los valores son diferentes, a 0 si son iguales

- ¿Tienen todos los atributos la misma importancia?
  - Si no tienen igual importancia, será necesario ponderar los atributos

2. Medidas de distancia y similaridad

- Distancia de *Minkowski*:

$$d(i,j)=\sqrt[q]{(|x_{i_1}-x_{j_1}|^q+|x_{i_2}-x_{j_2}|^q+...+|x_{i_p}-x_{j_p}|^q)}$$

donde $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ y $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ son dos objetos p-dimensionales, y q es un entero positivo

- Si $q = 1$, d es la distancia de Manhattan

$$d(i,j)=|x_{i_1}-x_{j_1}|+|x_{i_2}-x_{j_2}|+...+|x_{i_p}-x_{j_p}|$$

- *Si* $q = 2$, d es la distancia Euclídea
- Como ya hemos comentado se pueden usar pesos. Por ejemplo, Euclídea con pesos:

$$d(i,j)=\sqrt{w_1|x_{i_1}-x_{j_1}|^2+w_2|x_{i_2}-x_{j_2}|^2+...+w_p|x_{i_p}-x_{j_p}|^2}$$

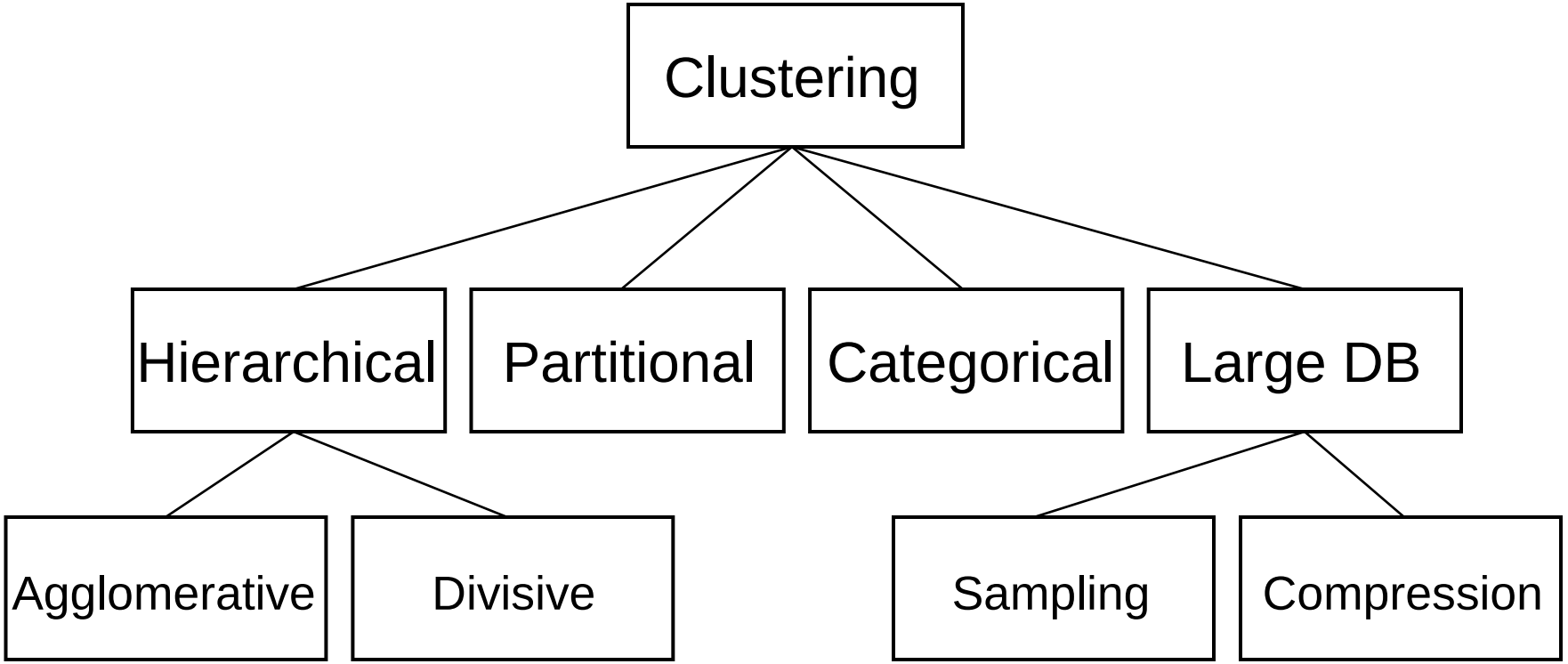
Tema 11. Agrupamiento

1. *Clustering*/agrupamiento/segmentación
2. Medidas de distancia y similaridad
3. Distintas aproximaciones al *clustering*
4. Métodos basados en particionamiento
5. Métodos jerárquicos

3. Distintas aproximaciones al *clustering*

- Algoritmos de particionamiento: Construir distintas particiones y evaluarlas de acuerdo a algún criterio
- Algoritmos jerárquicos: Crear una descomposición jerárquica del conjunto de datos (objetos) usando algún criterio
- Otros:
  - Basados en densidad, utilizan funciones de conectividad y densidad
  - Basados en rejillas, utilizan una estructura de granularidad de múltiples niveles
  - Basados en modelos. Se supone un modelo para cada uno de los *clusters* y la idea es encontrar el modelo que mejor ajuste

3. Distintas aproximaciones al clustering



Tema 11. Agrupamiento

1. *Clustering*/agrupamiento/segmentación
2. Medidas de distancia y similaridad
3. Distintas aproximaciones al clustering
4. Métodos basados en particionamiento
5. Métodos jerárquicos

4. Métodos basados en particionamiento

- Métodos basados en particionamiento: Construyen una partición de la base de datos *D* formada por *n* objetos en un conjunto de *k clusters*

- Dado un valor para *k*, encontrar la partición de D en *k clusters* que optimice el criterio de particionamiento elegido

- Métodos Heurísticos:
  - *k-means* (k medias): Cada *cluster* se representa por el centro del *cluster*
  - *k-medoids* o PAM (particionamiento alrededor de los *medoides*): cada *cluster* se representa por uno de los objetos incluidos en el *cluster*

4. Métodos basados en particionamiento

Algoritmo *k-means*

- Necesita como parámetro de entrada el número de *clusters* deseado

- Es un algoritmo iterativo en el que las instancias se van moviendo entre *clusters* hasta que se alcanza el conjunto de *clusters* deseado

4. Métodos basados en particionamiento

Algoritmo *K-Means*

- Para k=1,...,K hacer
  - r[k] = punto seleccionado arbitrariamente de D
- Mientras haya cambios en los *clusters* $C_1,...,C_k$ hacer

  - Para k=1,...,K hacer // construir los *clusters*

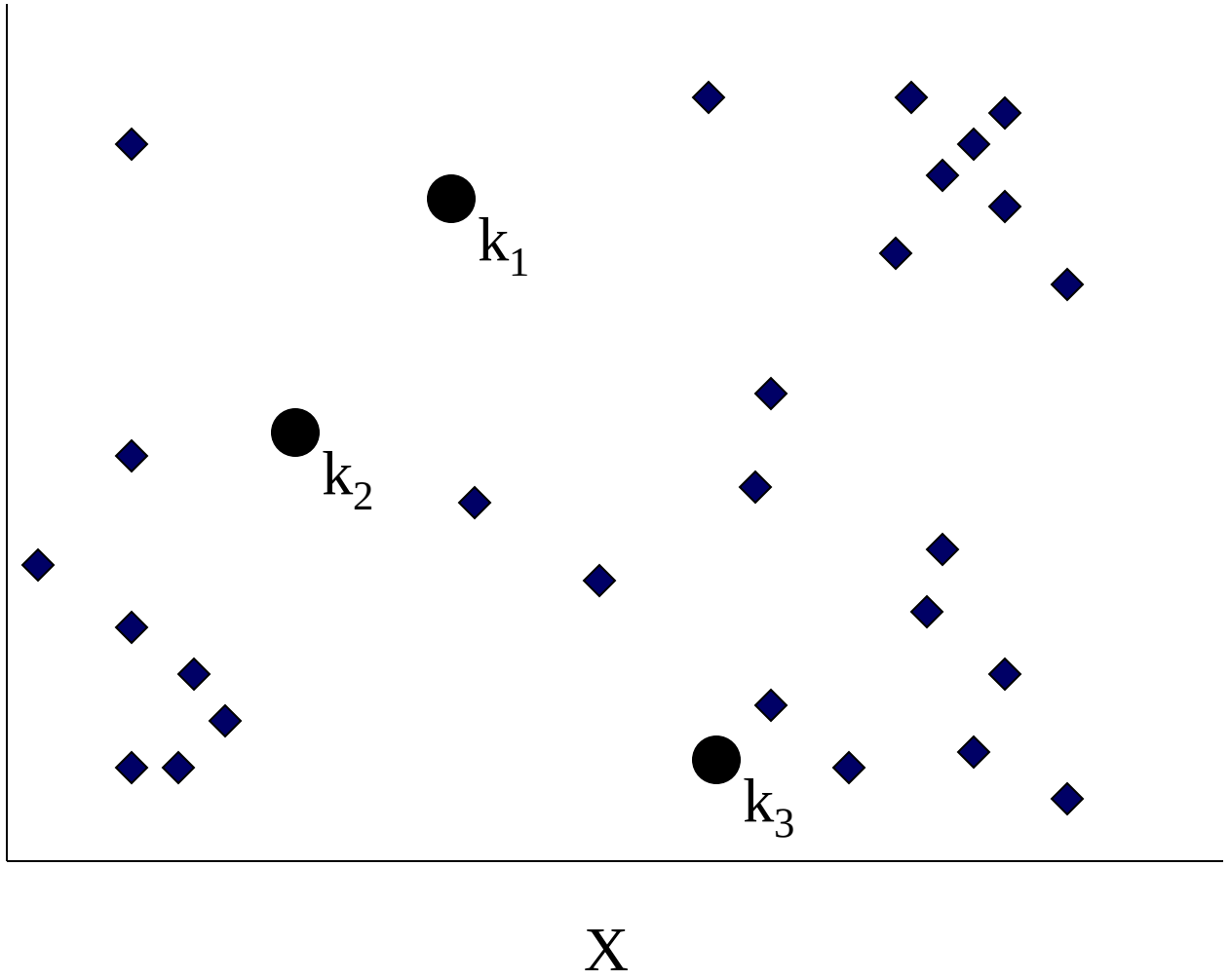
    $C_k = \{ x \in D \mid d(r[k],x) \leq d(r[j],x)$ para todo $j=1,...,K, \ j \neq k\}$

  - Para k=1,...,K hacer // calcular los nuevos centros

    r[k] = el punto medio de los objetos en C_k

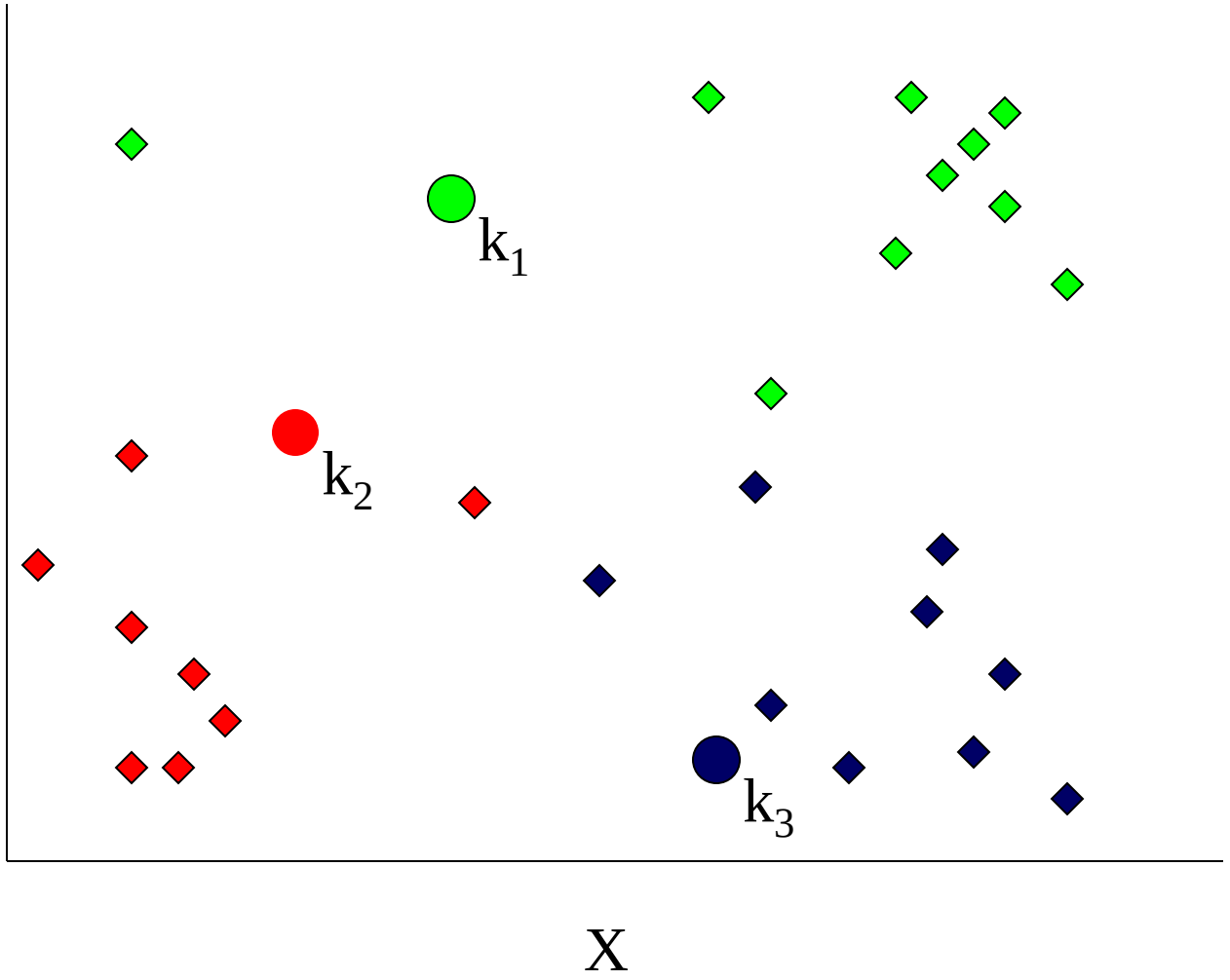
4. Métodos basados en particionamiento.
Ejemplo de *K-means*

Elige 3 centros de *clusters* (aleatoriamente)



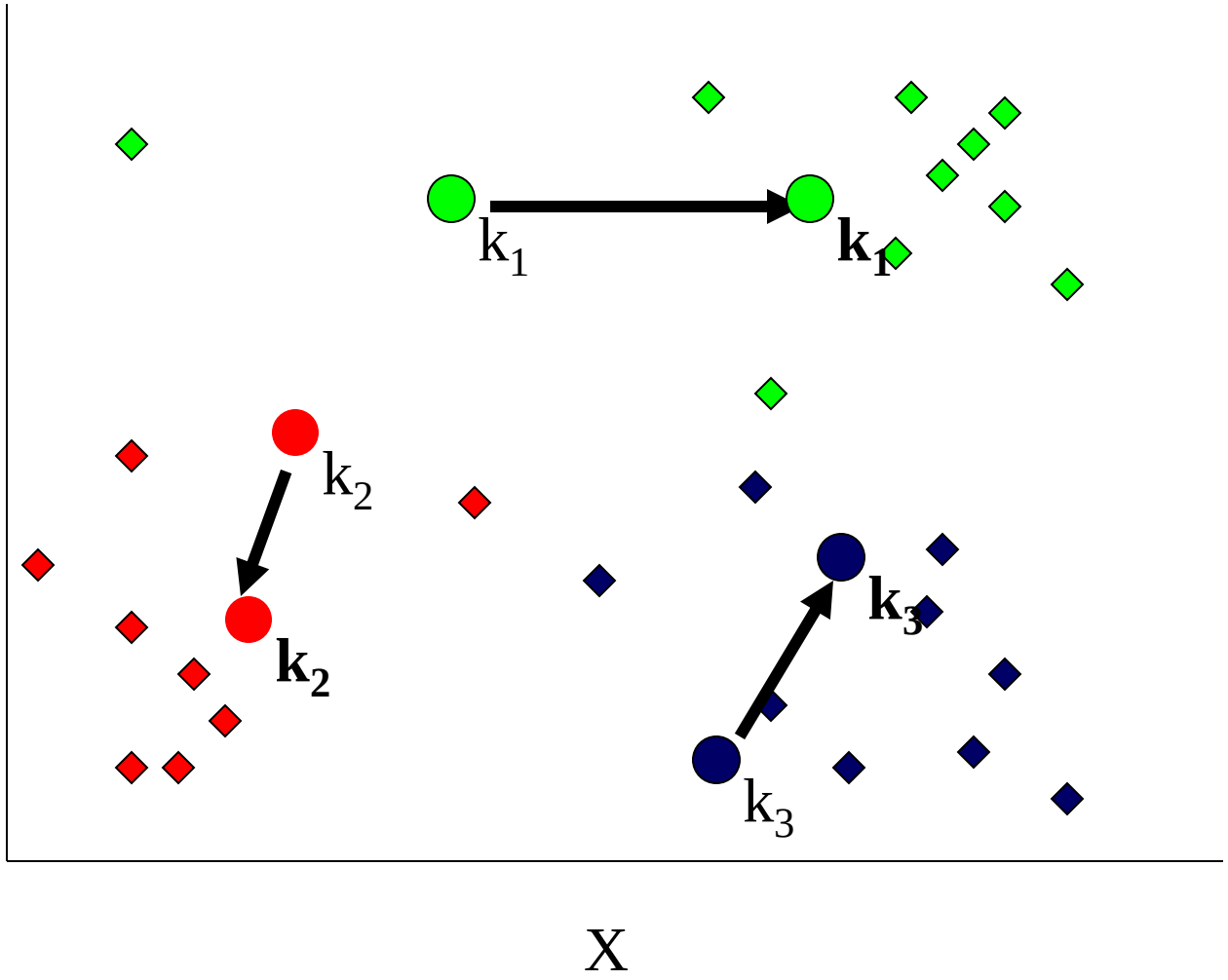
4. Métodos basados en particionamiento.
Ejemplo de *K-means*

Asigna cada punto al centro de *cluster* más cercano



4. Métodos basados en particionamiento.
Ejemplo de *K-means*

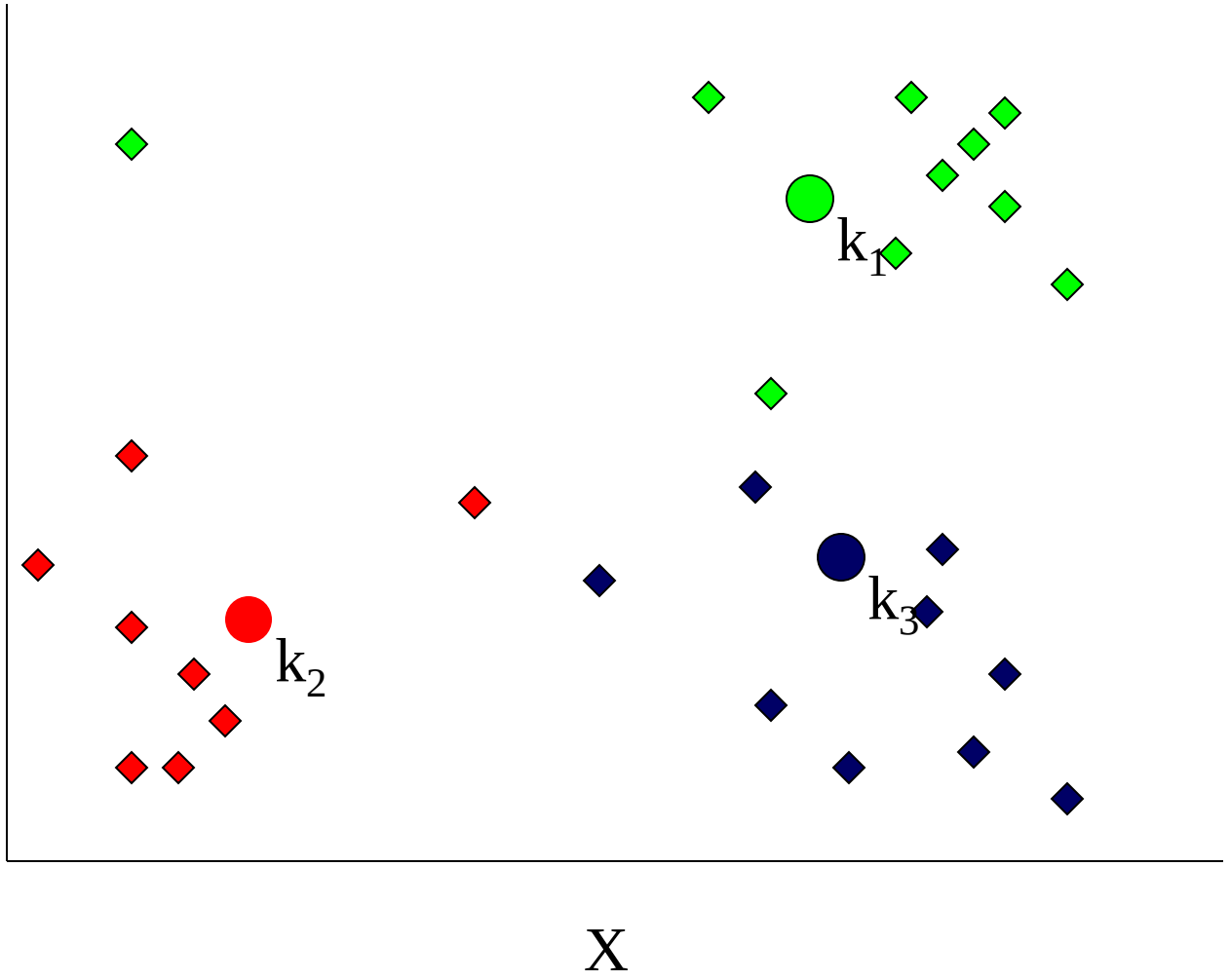
Mueve cada centro de *cluster* a la media de cada *cluster*



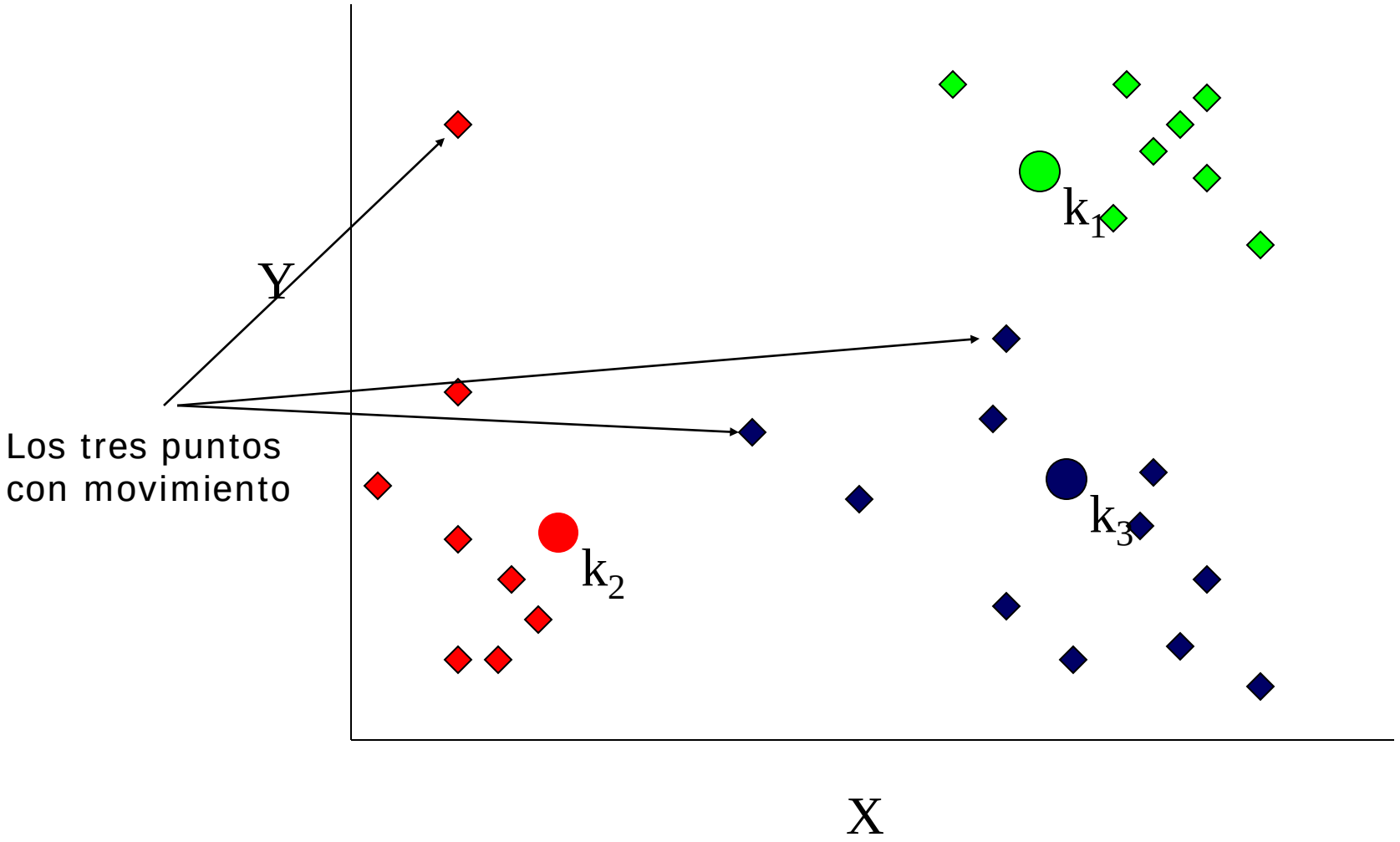
4. Métodos basados en particionamiento.
Ejemplo de *K-means*

Reasigna los puntos más cercanos a diferentes centros de *clusters*

¿Qué puntos se reasignan?

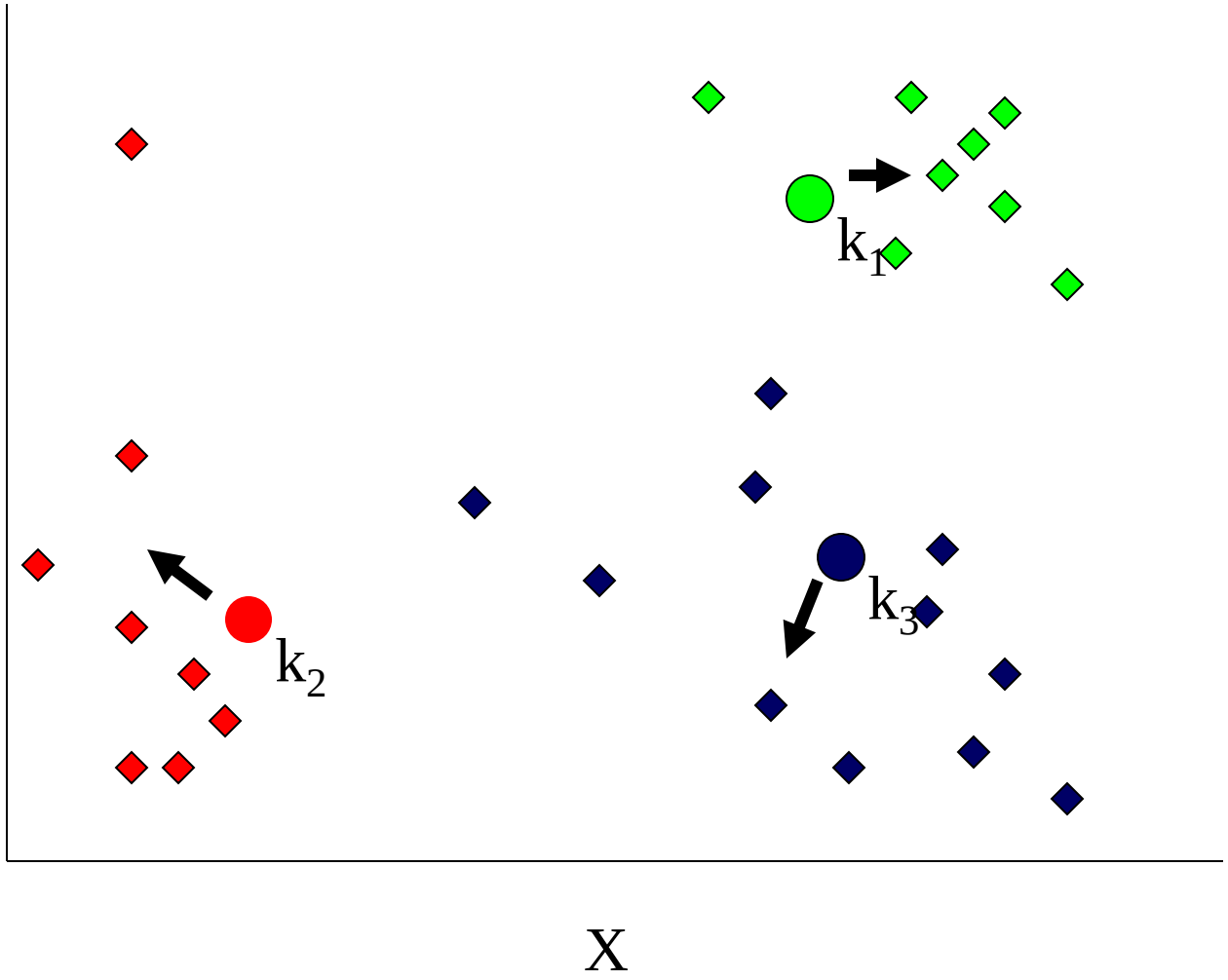


4. Métodos basados en particionamiento.
Ejemplo de *K-means*



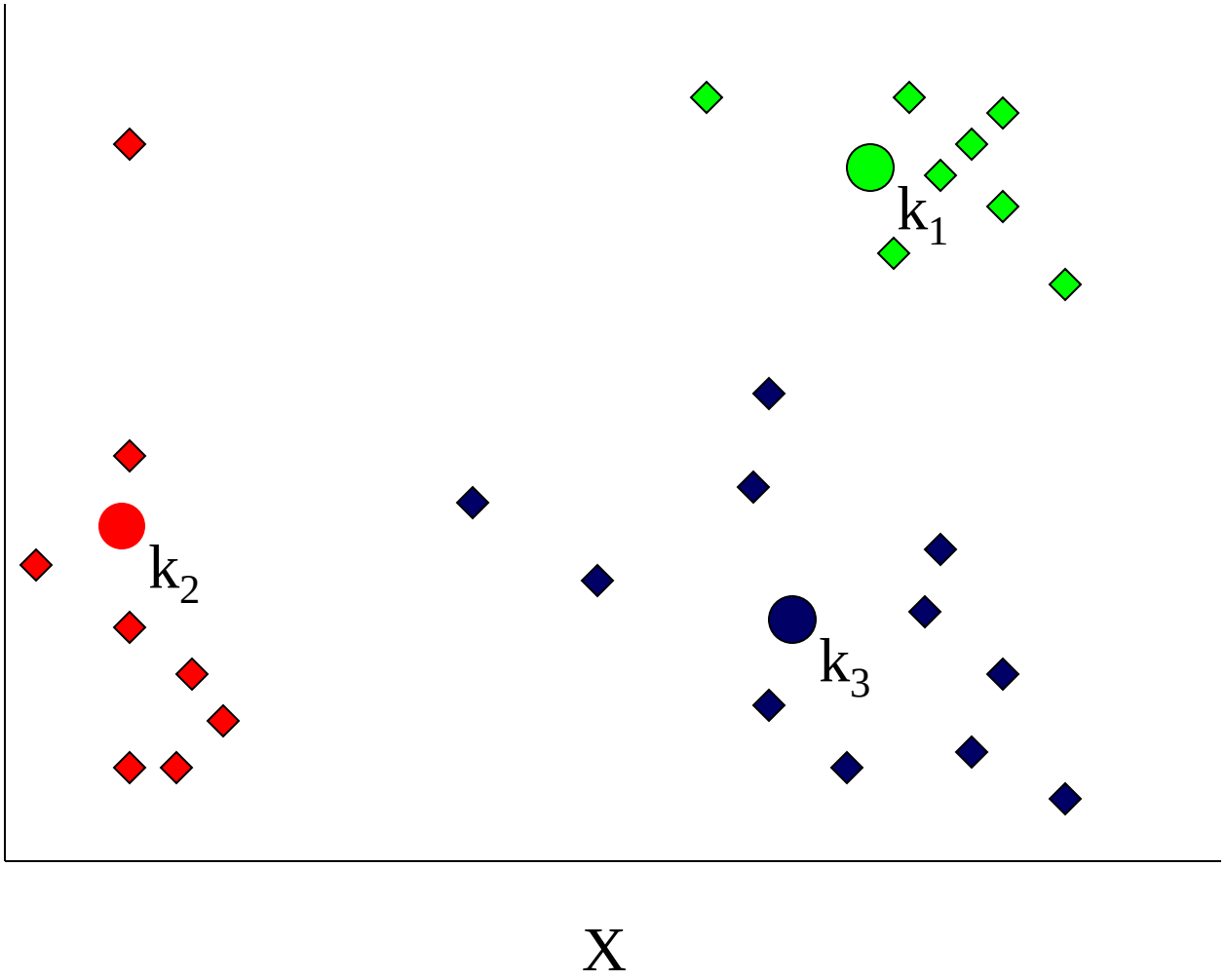
4. Métodos basados en particionamiento.
Ejemplo de *K-means*

Re-cálculo de los centros de *clusters*



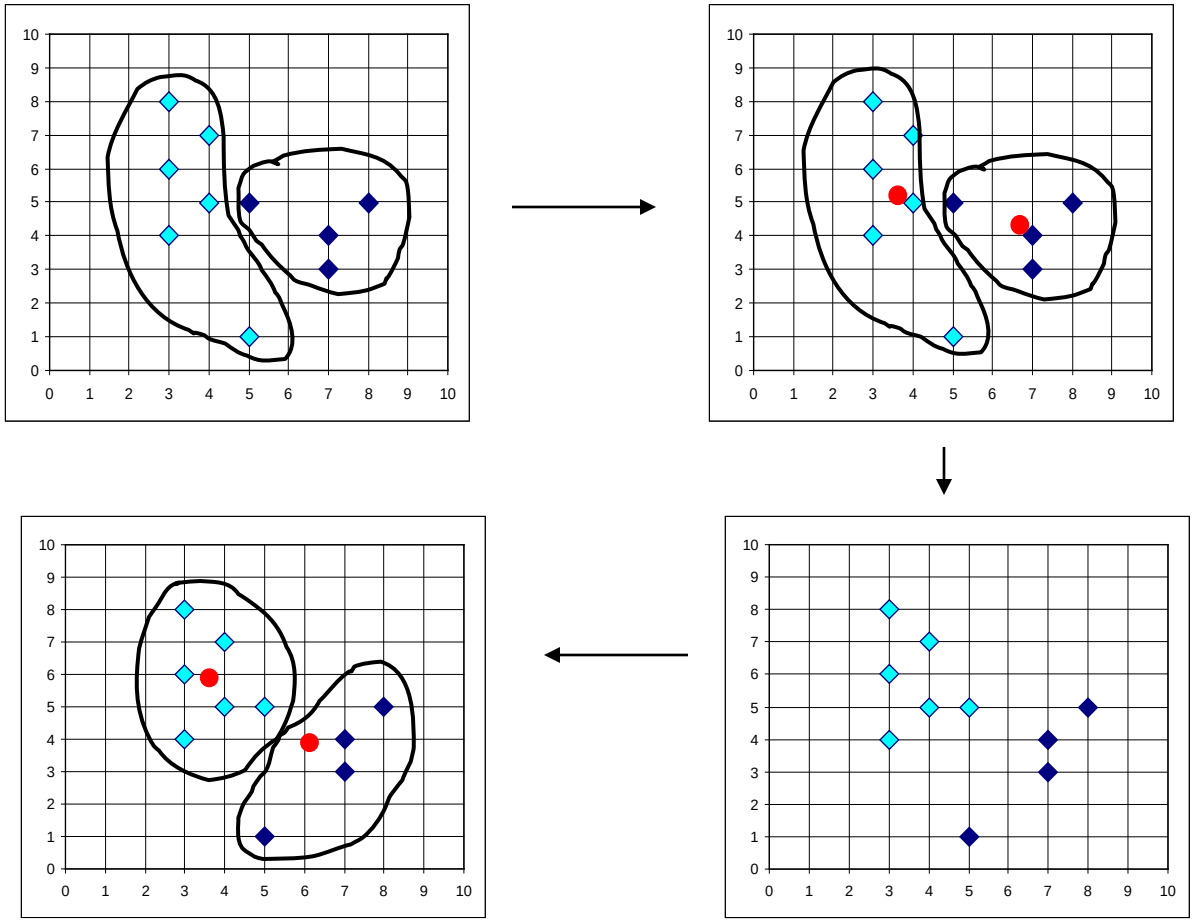
4. Métodos basados en particionamiento.
Ejemplo de *K-means*

Mover los centros de *cluster* a las medias de los *clusters*



4. Métodos basados en particionamiento. Ejemplo de *K-means*

- Ejemplo:



4. Métodos basados en particionamiento.
Ejemplo de *K-means*

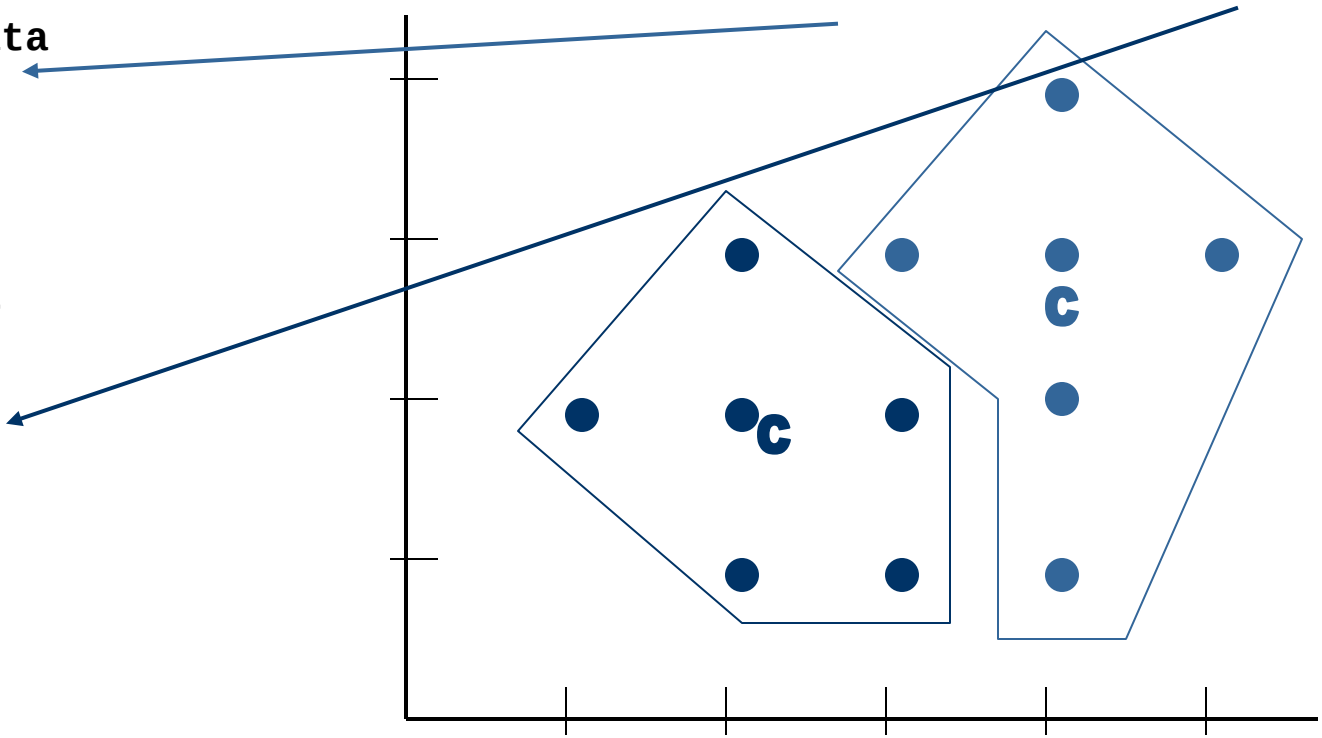
Ejemplo 12.real.arff

```
@attribute x real
@attribute y real

@data
5,3
1,2
2,3
4,1
4,4
2,1
3,2
3,1
4,3
2,2
3,3
4,2
```

Tomando K=2,

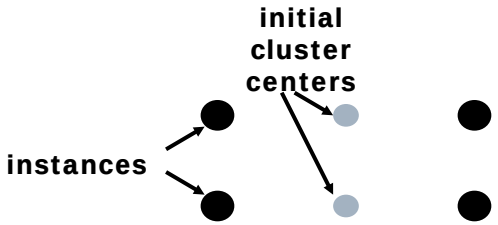
r[1]=(5,3), r[2]=(3,2)



4. Métodos basados en particionamiento. Algunos comentarios sobre k-means

Ventajas

- *Relativamente eficiente*: $O(tkn)$, donde n es # objetos, k es # *clusters*, y t es # iteraciones. Normalmente, k, $t << n$.
- Con frecuencia finaliza en un óptimo local, dependiendo de la elección inicial de los centros de *clusters*.



  - Reinicializar las semillas
  - Utilizar técnicas de búsqueda más potentes como algoritmos genéticos o enfriamiento estocástico.

Desventajas

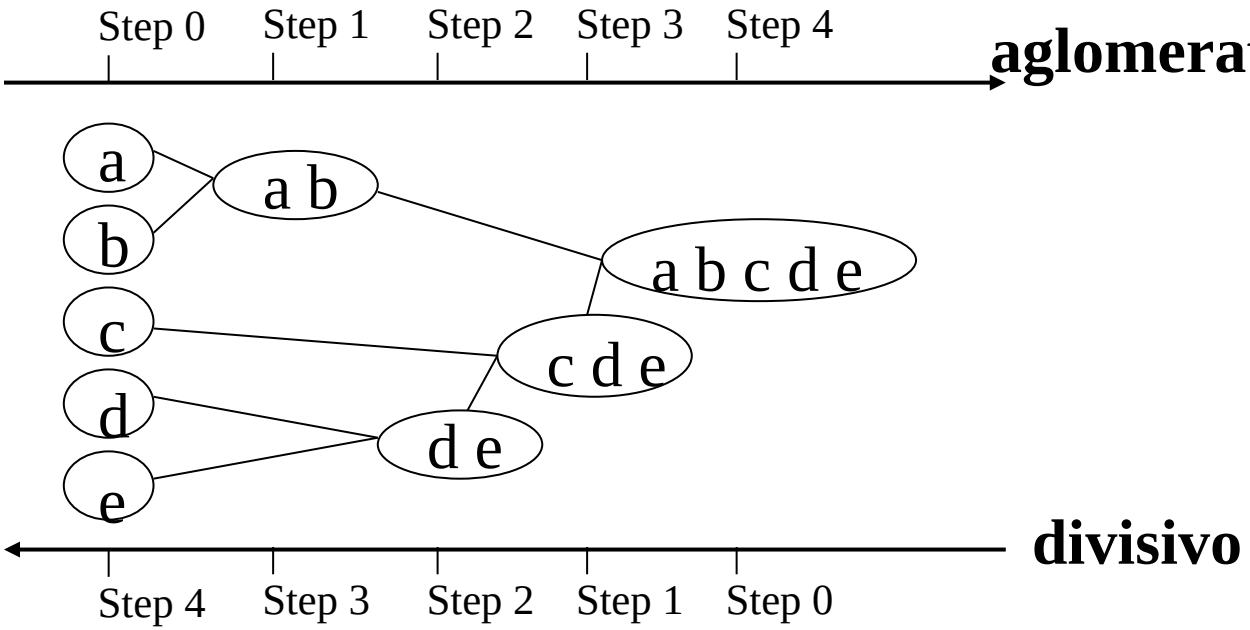
- Sólo es aplicable cuando el concepto de media es definible. ¿qué hacer con datos nominales?
- Necesidad de fijar anticipadamente el número de *clusters* (k)
- Débil ante datos ruidosos y/o con outliers
- Sólo indicado para *clusters* convexos (esféricos…)

Tema 11. Agrupamiento

1. *Clustering*/agrupamiento/segmentación
2. Medidas de distancia y similaridad
3. Distintas aproximaciones al *clustering*
4. Métodos basados en particionamiento
5. Métodos jerárquicos
   - 5.1. Métodos aglomerativos
   - 5.2. Métodos divisivos

5. Métodos jerárquicos

- La salida es una jerarquía entre *clusters*
- Dependiendo del nivel de corte obtendremos un *clustering* distinto
- No requiere como parámetro el número de *clusters*



5.1. Métodos aglomerativos

- Se basan en medir la distancia entre *clusters*
- En cada paso se fusionan los dos *clusters* más cercanos
- La situación de partida suele ser un *cluster* por cada objeto en la BD $D = \{x_1,\ldots,x_n\}$

- Para $i=1,\ldots,n$ hacer $C_i=\{x_i\}$
- Mientras haya más de un *cluster*
  - Sean C_i y C_j los dos *clusters* que minimizan la distancia entre *clusters*
  - $C_i = C_i \cup C_j$
  - Eliminar el *cluster* C_j

5.1. Métodos aglomerativos

Distancias entre *clusters*

- Distancia del vecino más próximo

$$D_{vp}(C_i, C_j) = \min_{i,j}\{d(x,y) \mid x \in C_i, y \in C_j\}$$

donde *d(,)* es una distancia entre objetos

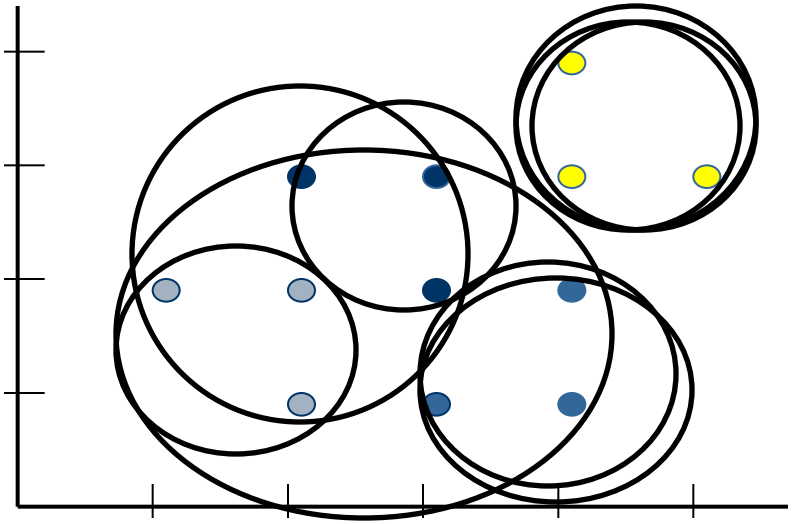
- Distancia del vecino más lejano

$$D_{vl}(C_i, C_j) = \max_{i,j}\{d(x,y) \mid x \in C_i, y \in C_j\}$$

- Distancia entre los centroides

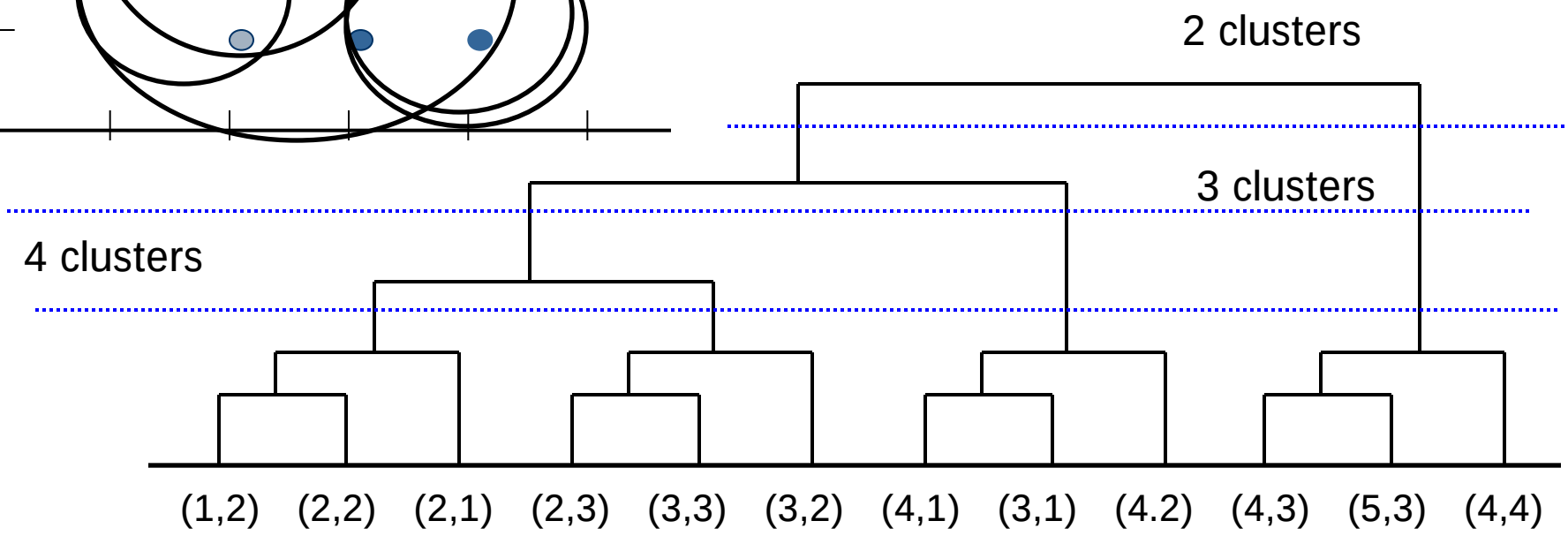
5.1. Métodos aglomerativos

Ejemplo con 12.real.arff



Clustering Aglomerativo

Distancia de Manhattan entre centroides



5.2. Métodos divisivos

- Comienzan con un único *cluster* (toda la BD) y en cada paso se selecciona un *cluster* y se subdivide
- Se debe dar una condición de parada, o en su defecto se detiene el proceso cuando cada *cluster* contiene un único objeto
- Podemos distinguir dos variantes:
  - Unidimensional (*Monothetic*). Sólo se considera una variable para hacer la partición
  - Multidimensional (*Polythetic*). Todas las variables se consideran para hacer la partición
    - Se usa una distancia entre *clusters* para medir
- Mucho menos utilizados que los métodos aglomerativos con un bajo número de vecinos “cercanos”