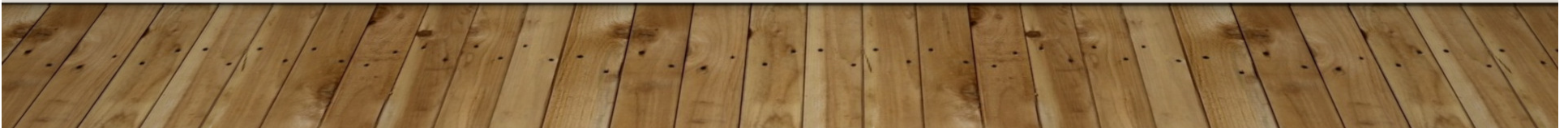


ALGORITMOS DE CLUSTERING O AGRUPAMIENTO (APRENDIZAJE NO SUPERVISADO O AUTÓNOMO)

Prof. Gustavo Ceballos Benavides

INTRODUCCIÓN

- Algoritmo de Agrupamiento o Clustering es una técnica o procedimiento de agrupación (No supervisado) de una serie de vectores de acuerdo con un criterio. Esos criterios son por lo general distancia o similitud. La cercanía se define en términos de una determinada función de distancia, como la euclídeana por ejemplo.
- Típicamente, los algoritmos no supervisados hacen inferencias de conjuntos de datos usando solo vectores de entrada sin referirse a resultados conocidos o etiquetados.
- El algoritmo de agrupamiento **k-means** (ó k-medias en español) es uno de los mas utilizados (por su simpleza y efectividad): La idea es agrupar datos similares y descubrir patrones subyacentes.
- Para lograr este objetivo, k-means busca un número fijo (k) de clusters en un conjunto de datos.



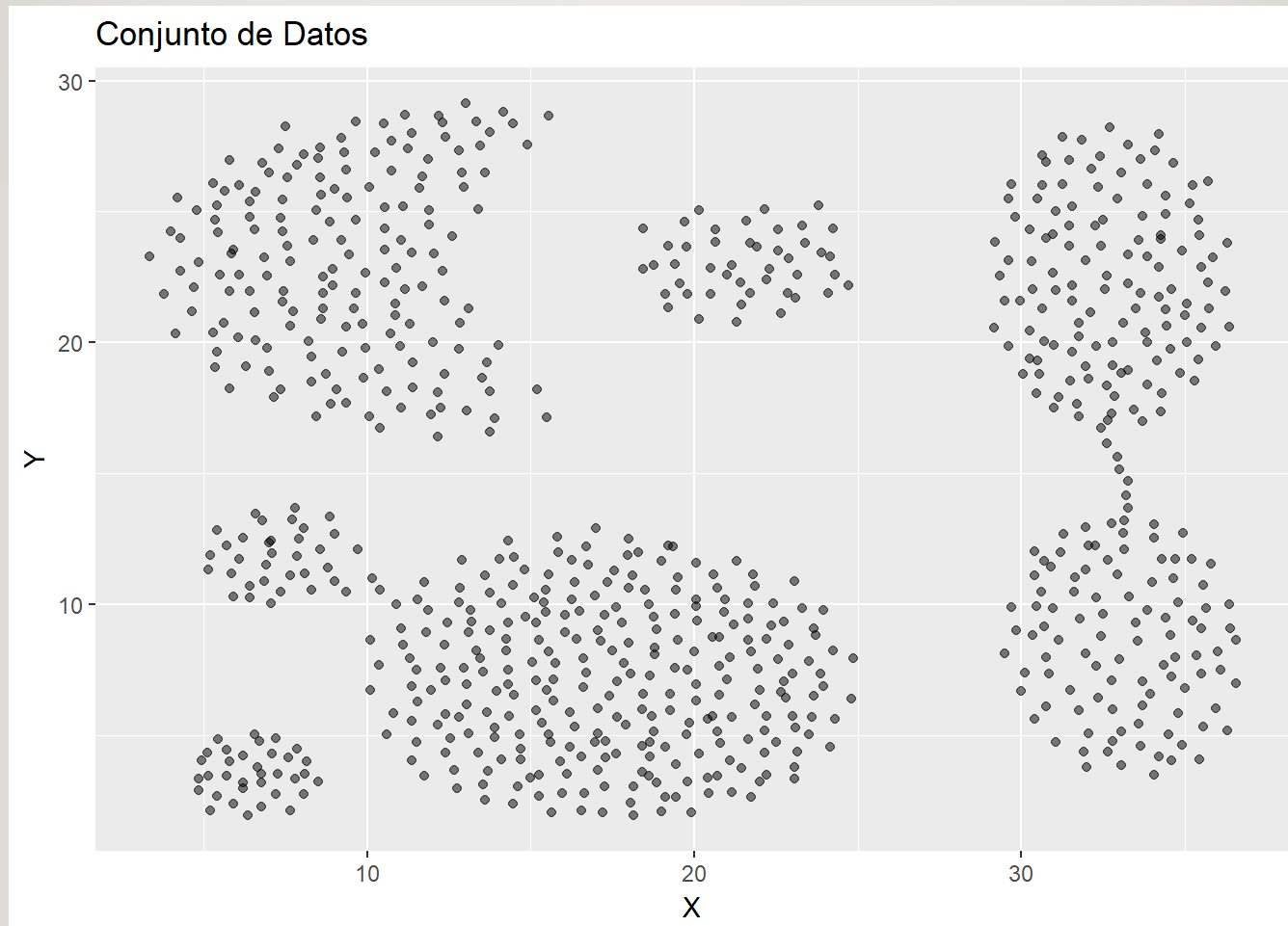
INTRODUCCIÓN

- Generalmente, los vectores de un mismo grupo (o *clústers*) comparten propiedades comunes.
- El conocimiento de los grupos puede permitir una descripción sintética de un conjunto de datos multidimensional complejo. De ahí su uso en minería de datos.
- Esta descripción sintética, se consigue sustituyendo la descripción de todos los elementos de un grupo por la de un representante característico del mismo (centroides ó centros de los cluster).

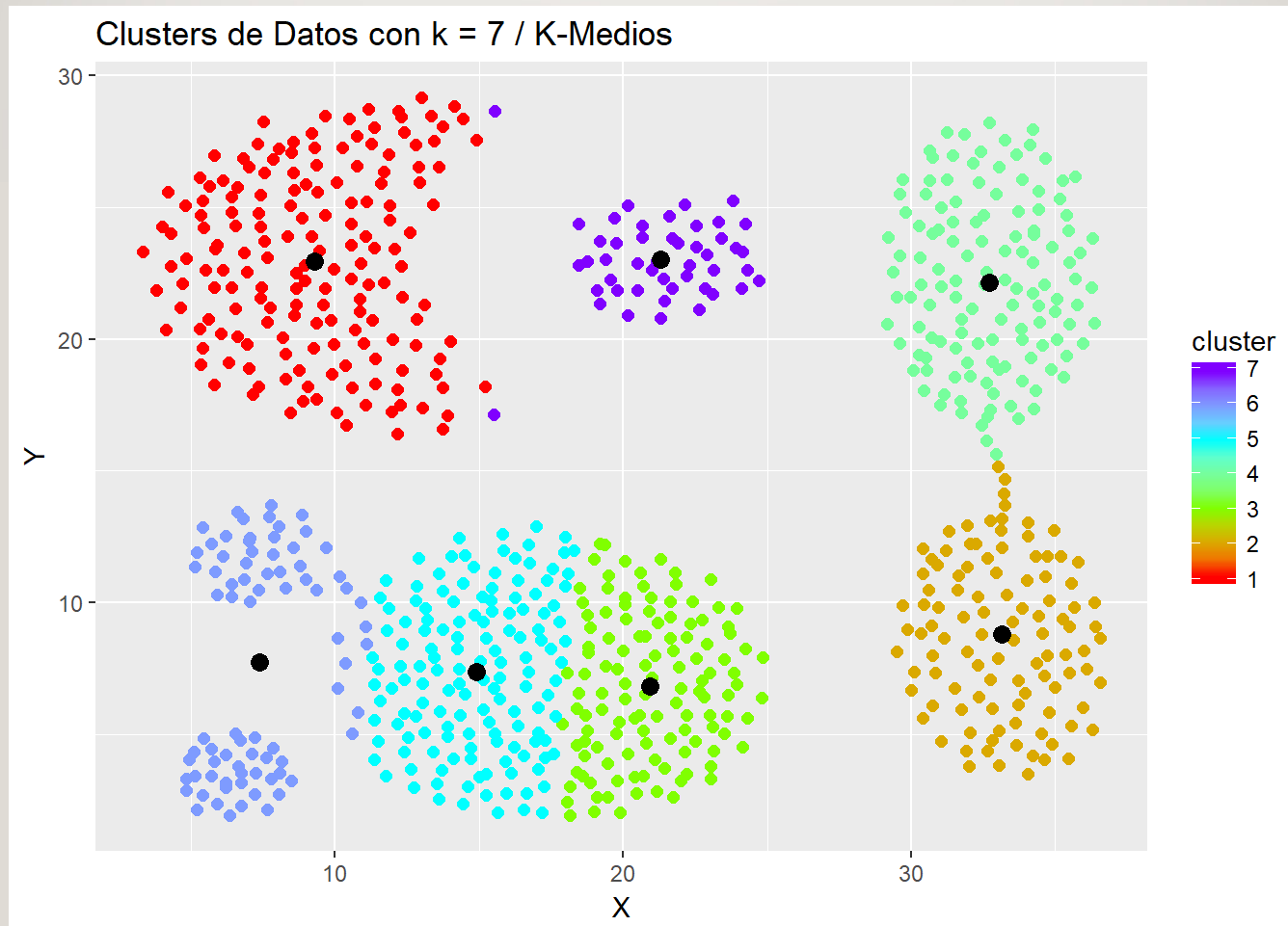
ALGORITMO K-MEANS

- 1) Se define el número de clusters y se inician centros de cluster al azar.
- 2) Se le presenta un patrón (vector de característica) de las muestras y se asocia al conjunto (grupo o cluster) más cercano en el sentido de alguna distancia (distancia euclidiana o euclidiana ponderada por ejemplo).
- 3) Se calcula el nuevo centro del cluster como el promedio entre el vector de característica presentado y el centro del cluster al cual fue asociado.
- 4) Se presenta el siguiente patrón o muestra y se vuelve al paso 2.
- 5) El algoritmo termina cuando los centros de los clusters no se modifican (o bien la modificación es menor que un cierto error prefijado por nosotros). También se puede terminar, definiendo un número máximo de iteraciones.

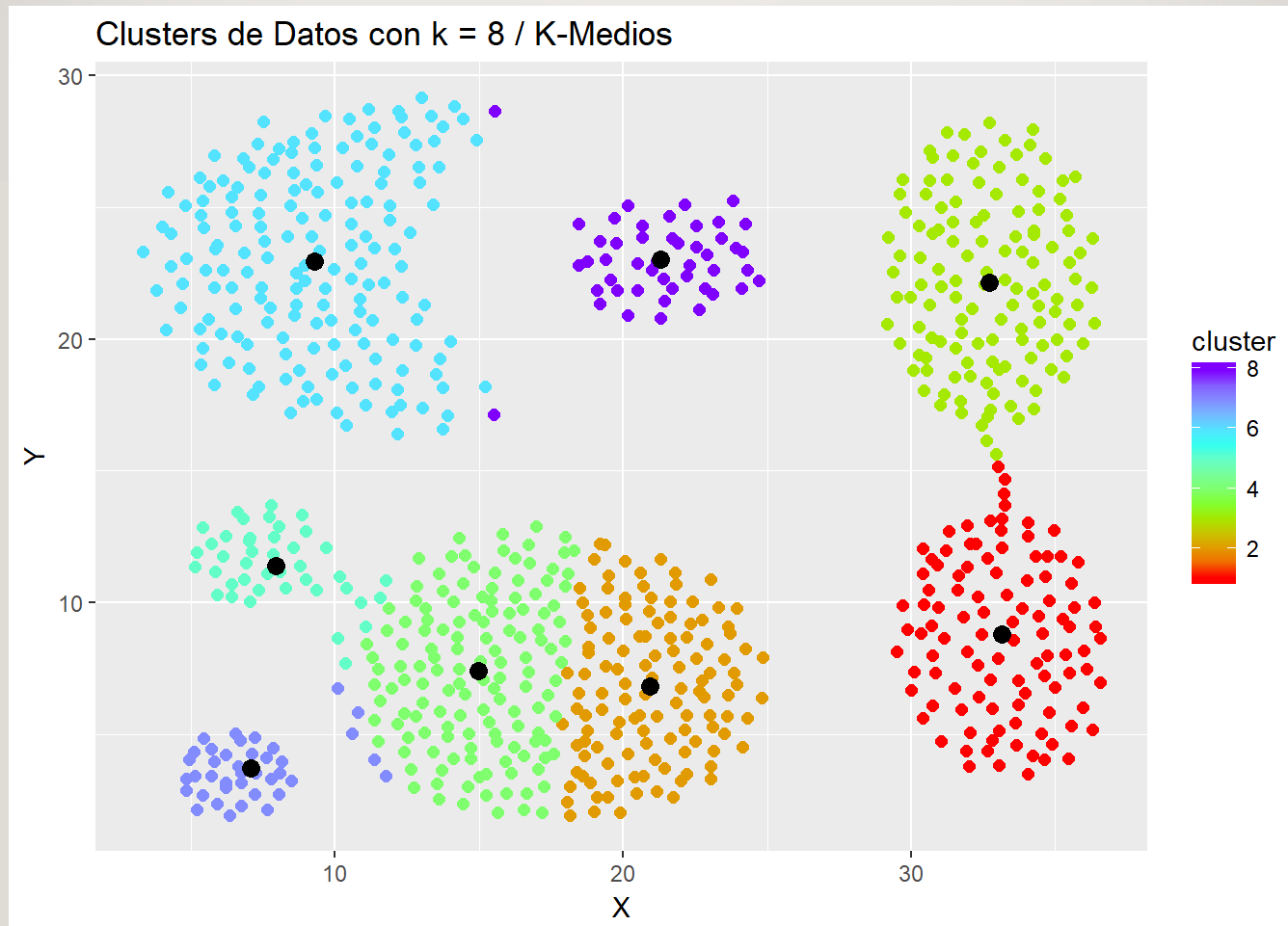
ALGORITMO K-MEANS



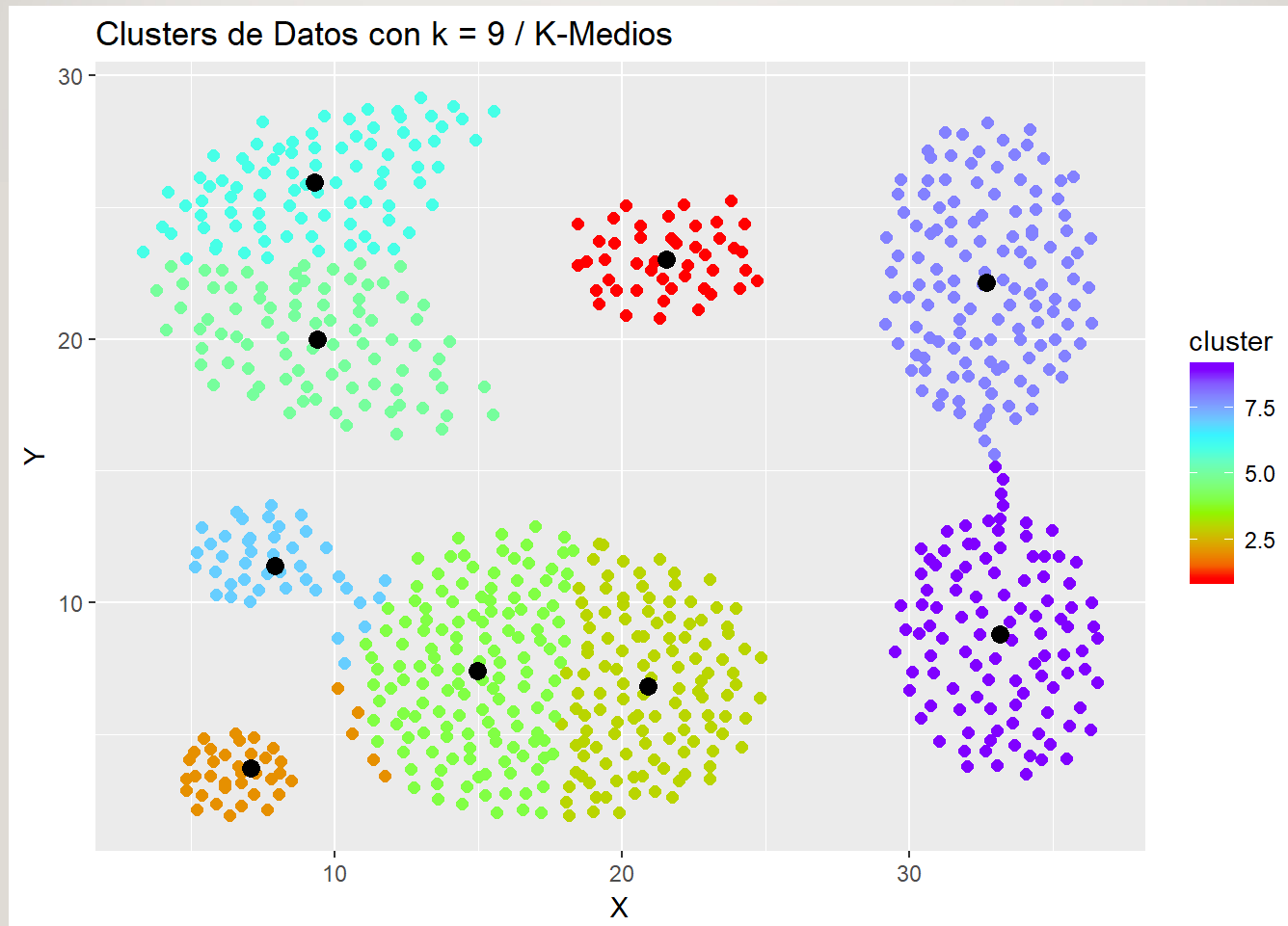
ALGORITMO K-MEANS



ALGORITMO K-MEANS



ALGORITMO K-MEANS



ALGORITMO K-MEANS CON MATLAB

- 1) se cargan las muestras, como por ejemplo:
- `load data_vinos_2clases`
- ---
- `X=data_vinos_2clases(:,1:13);`
- `%t=data_vinos_2clases(:,14);` Este dato ya no es necesario puesto que este algoritmo NO necesita entrenamiento SUPERVISADO como las Redes Neuronales feedforward multicapas vistas en la Unidad anterior del curso.
- 2) `[id,C] = kmeans(X,k,'MaxIter',1000);`
- En la matriz **C** se almacenan los centros de los Clusters e **id** indica el índice del conjunto al que fueron asociado las muestras de datos o patrones.