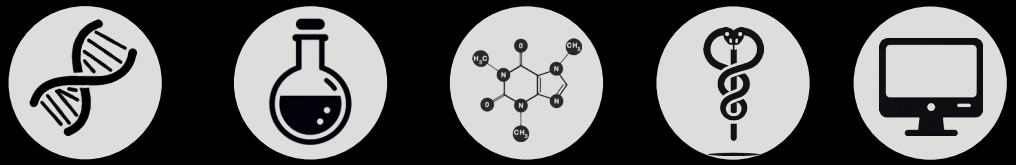
Grupo de Ciencia Computacional HIMFG



Algoritmos de Clustering (I)

Introducción al algoritmo de clustering K-means

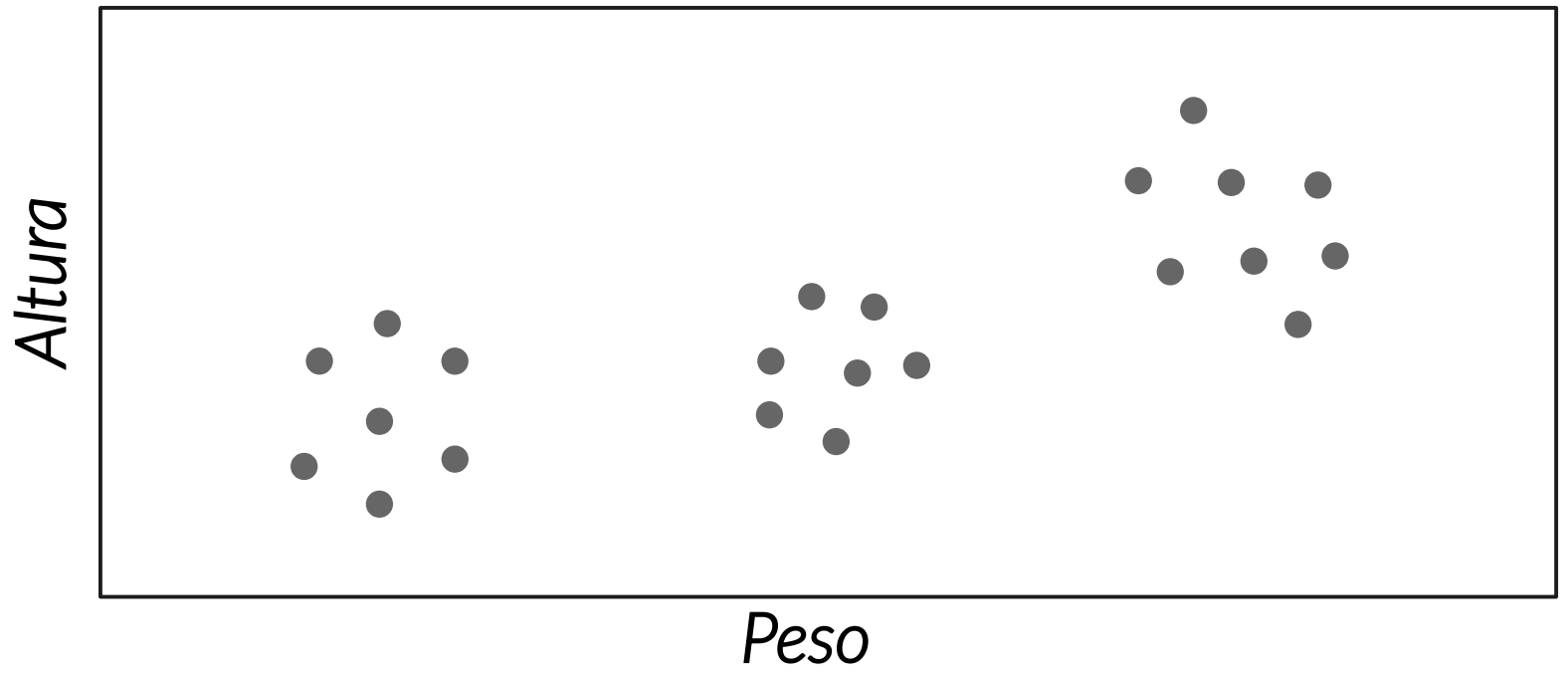
20 Septiembre 2024 || Diego Prada Gracia @ Grupo de Ciencia Computacional HIMFG


creative
commons

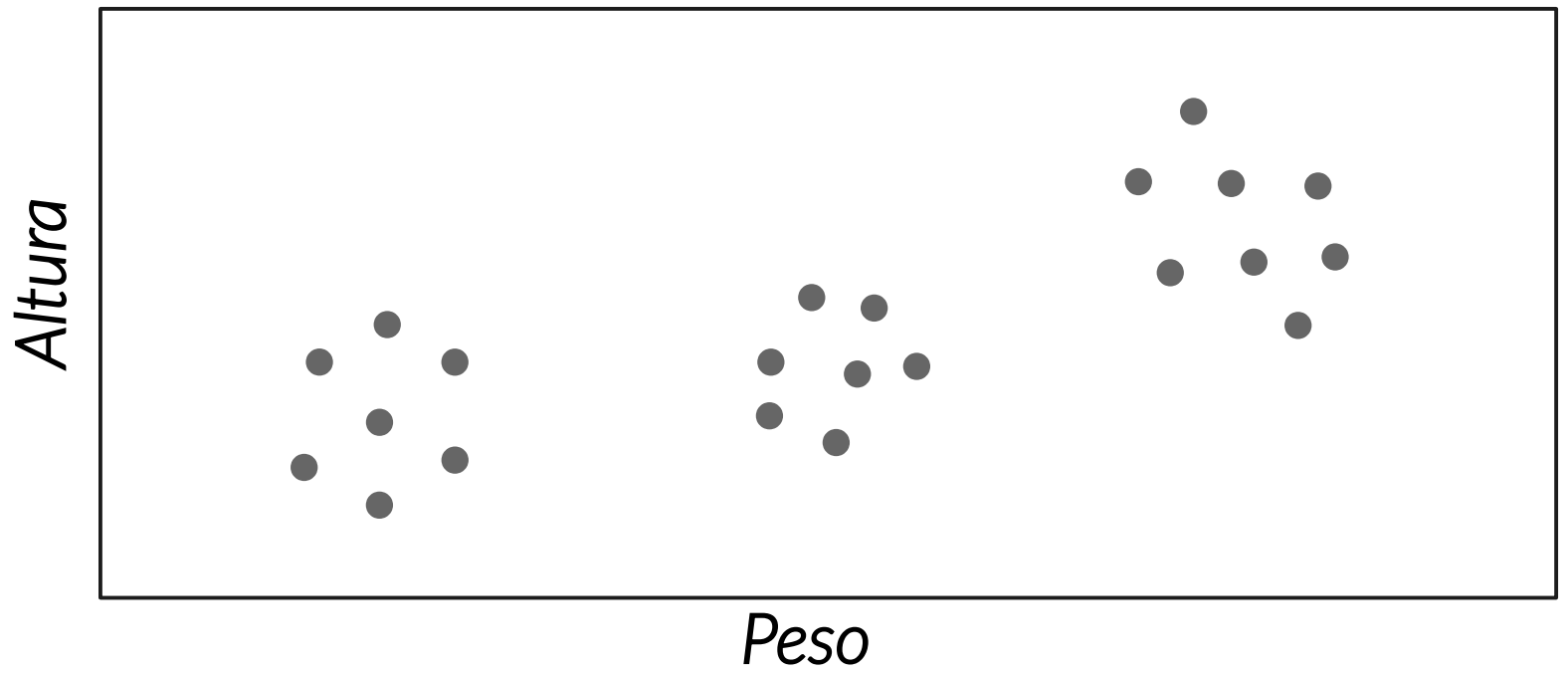

BY NC SA

¿Qué es un algoritmo de clustering?

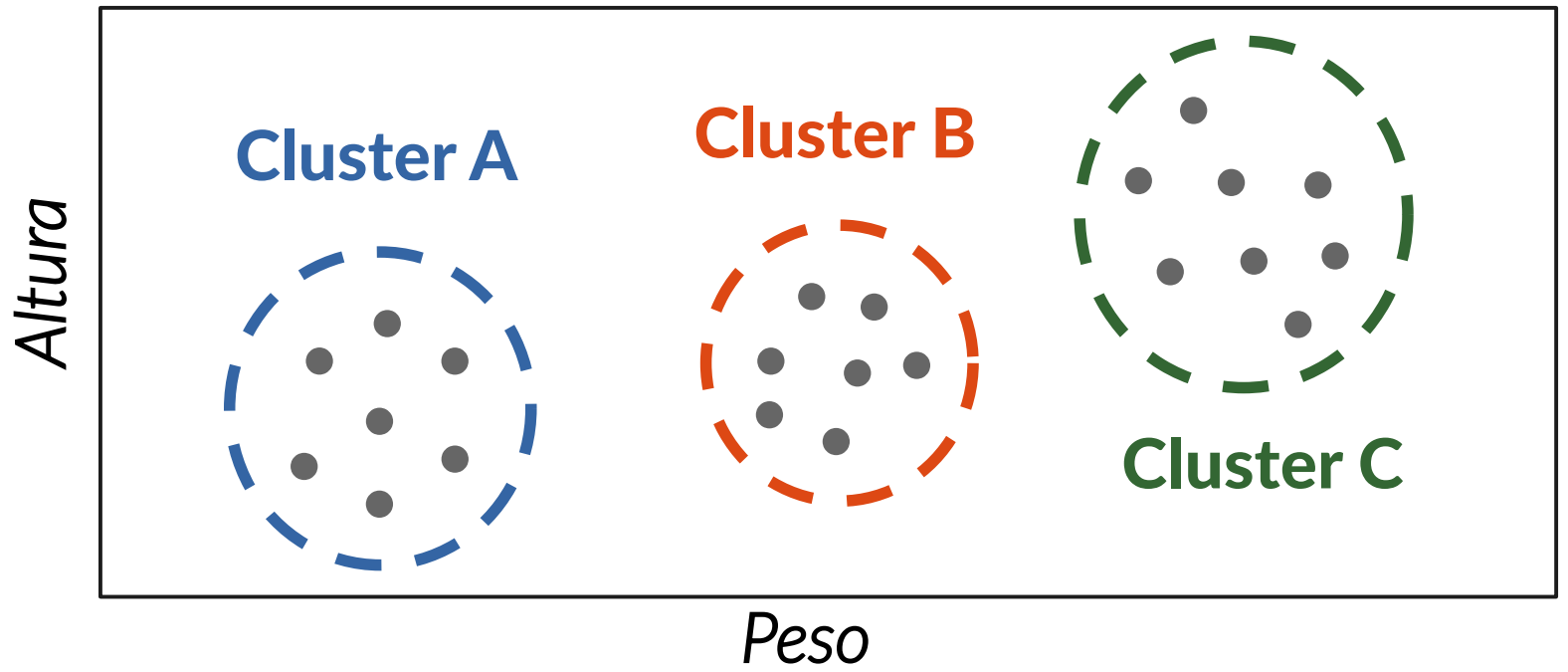
Sea una población de individuos descritos por varias características

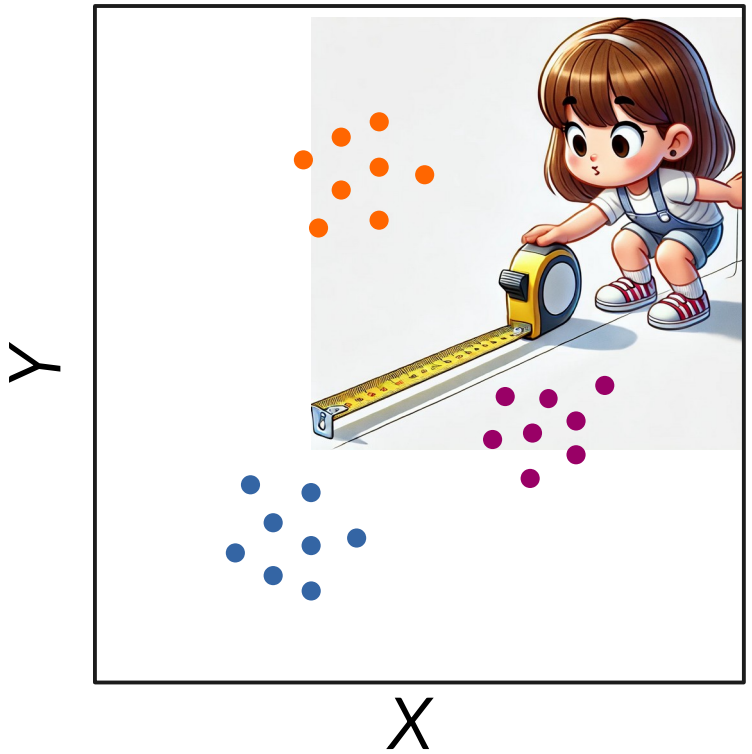


¿Ves grupos de individuos?



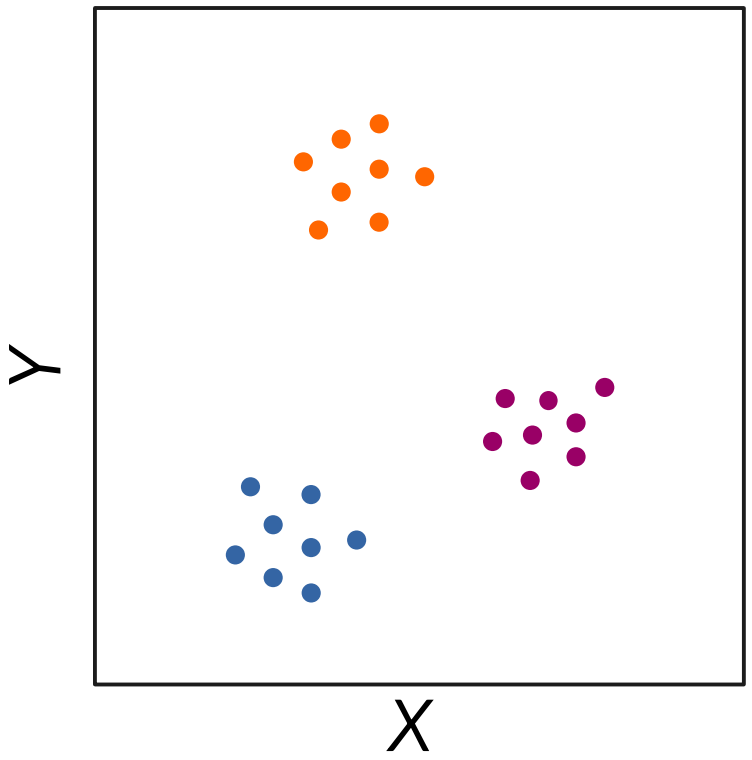
¿Ves grupos de individuos?

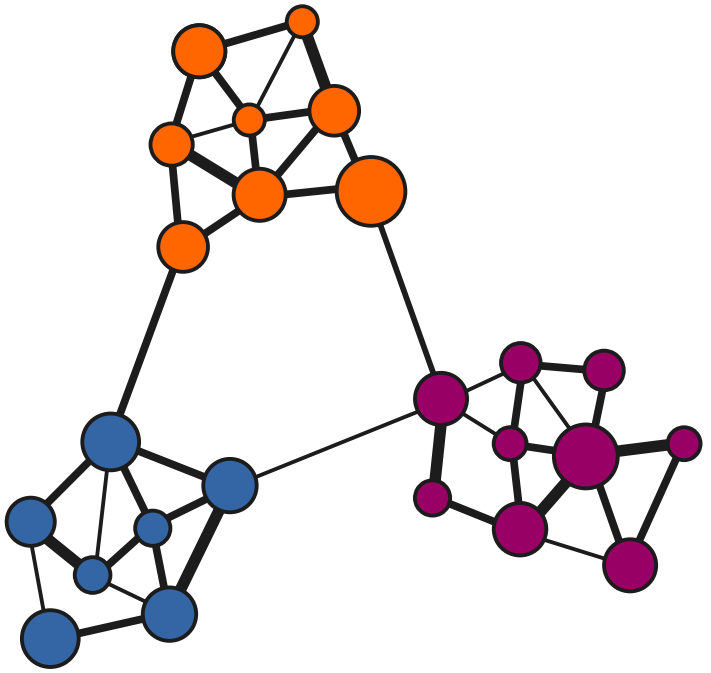




Definimos los **clusters** atendiendo a la **distancia** entre individuos

Cluster vs. Comunidad





¿Qué es la distancia?

Distancia

Una función definida entre dos puntos que cumple para todo punto A, B o C:

Distancia

Una función definida entre dos puntos que cumple para todo punto A, B o C:

1. No negatividad (Identidad de los indiscernibles)

$d(A,B) \geq 0$ *y* $d(A,B) = 0$ *si y solo si* $A = B$

Distancia

Una función definida entre dos puntos que cumple para todo punto A, B o C:

2. Simetría

$$d(A,B)=d(B,A)$$

Distancia

Una función definida entre dos puntos que cumple para todo punto A, B o C:

3. Desigualdad triangular

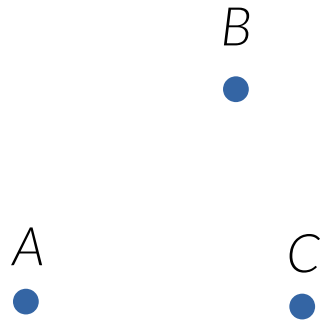
$$d(A,B) \leq d(A,C) + d(C,B)$$

Distancia

Una función definida entre dos puntos que cumple para todo punto A, B o C:

3. Desigualdad triangular

$$d(A,B) \leq d(A,C) + d(C,B)$$

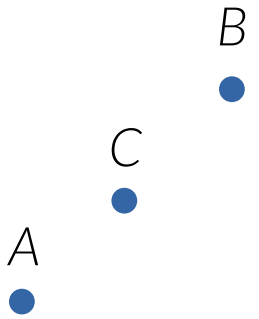


Distancia

Una función definida entre dos puntos que cumple para todo punto A, B o C:

3. Desigualdad triangular

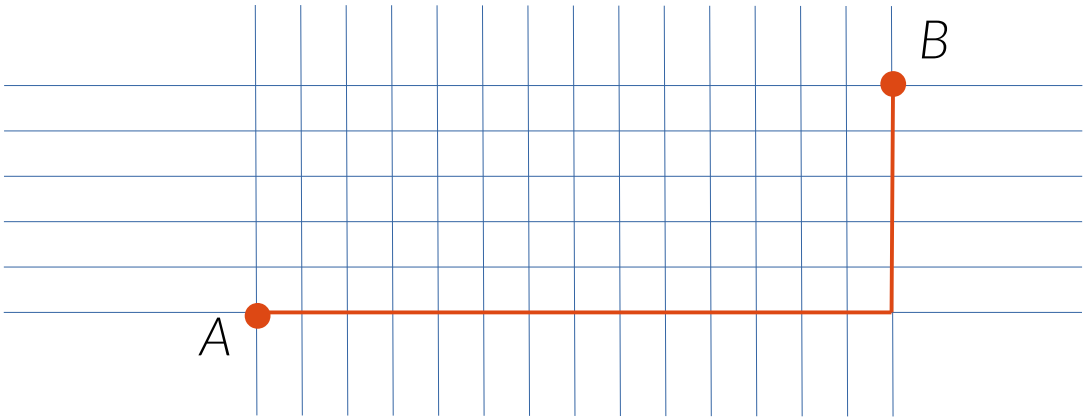
$$d(A,B) \leq d(A,C) + d(C,B)$$



Distancia

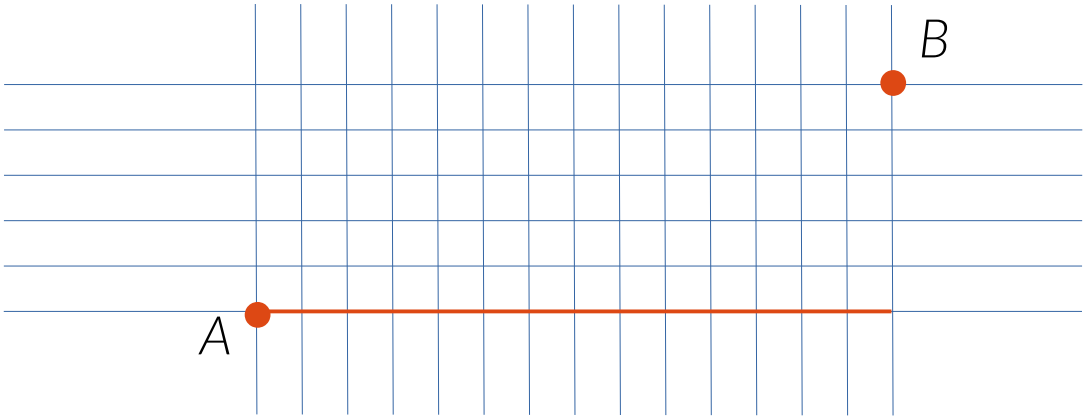
Cuantifica cuan separados están dos puntos en un espacio dado

Distancia de Manhattan



$$d(A,B)=|x_A-x_B|+|y_A-y_B|$$

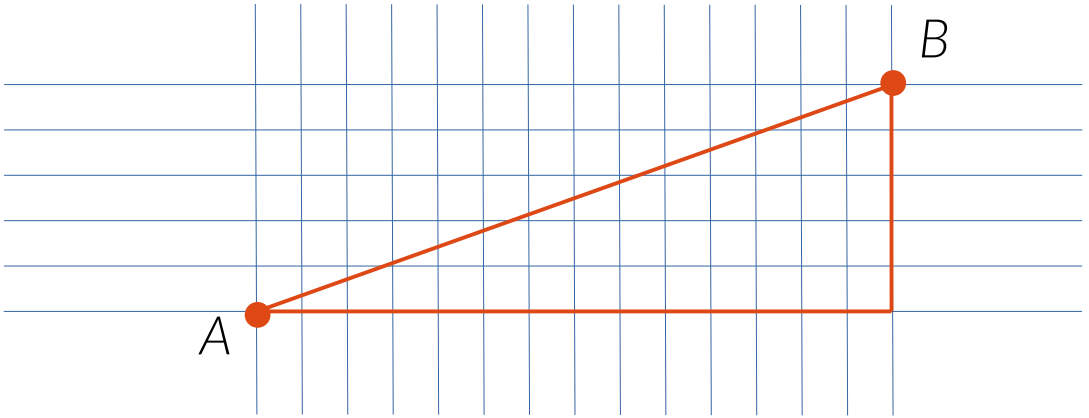
Distancia de Chebyshev



$$d(A,B)=max(|x_A-x_B|,|y_A-y_B|)$$

Distancia Euclidea 2D

Distancia de Minkowski con p=2



$$d(A,B)=\sqrt{(x_B-x_A)^2+(y_B-y_A)^2}$$

Distancia Euclidea 3D

Distancia de Minkowski con p=2

$$d(A,B)=\sqrt{(x_B-x_A)^2+(y_B-y_A)^2+(z_B-z_A)^2}$$

Distancia Euclidea 4D

Distancia de Minkowski con p=2

$$d(A,B)=\sqrt{(x_B-x_A)^2+(y_B-y_A)^2+(z_B-z_A)^2+(u_B-u_A)^2}$$

Distancia Euclidea 5D

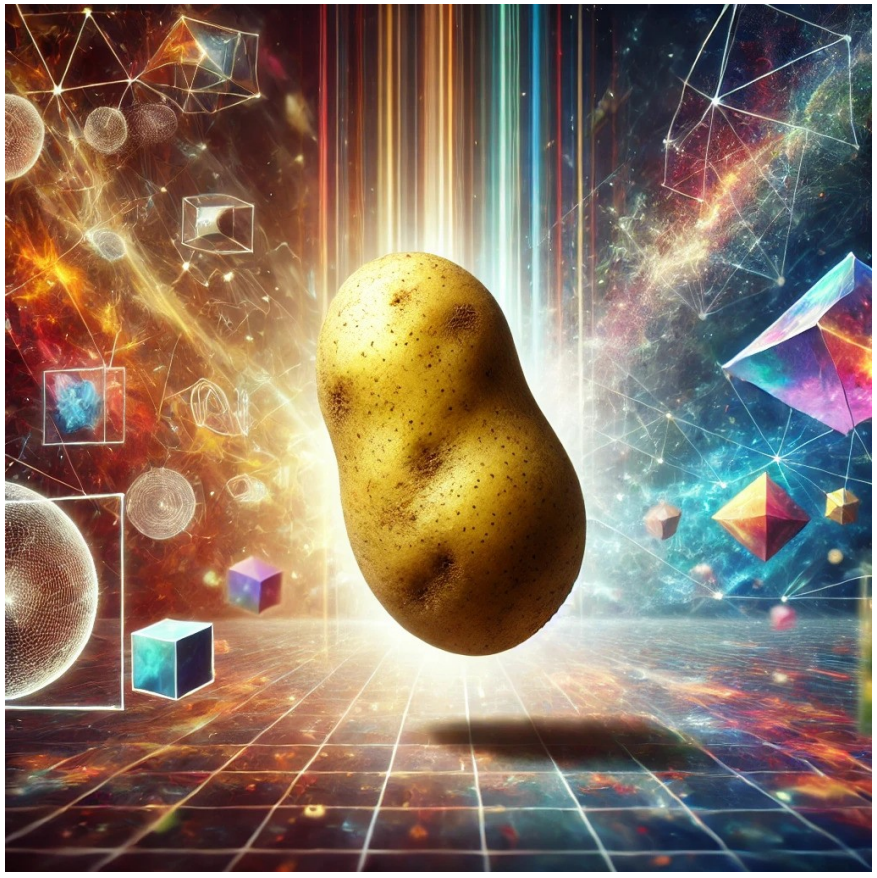
Distancia de Minkowski con p=2

$$d(A,B)=\sqrt{(x_B-x_A)^2+(y_B-y_A)^2+(z_B-z_A)^2+(u_B-u_A)^2+(v_B-v_A)^2}$$

Si tu sistema es multidimensional...

¿por qué clusterizas sólo en 2 o 3 dimensiones?

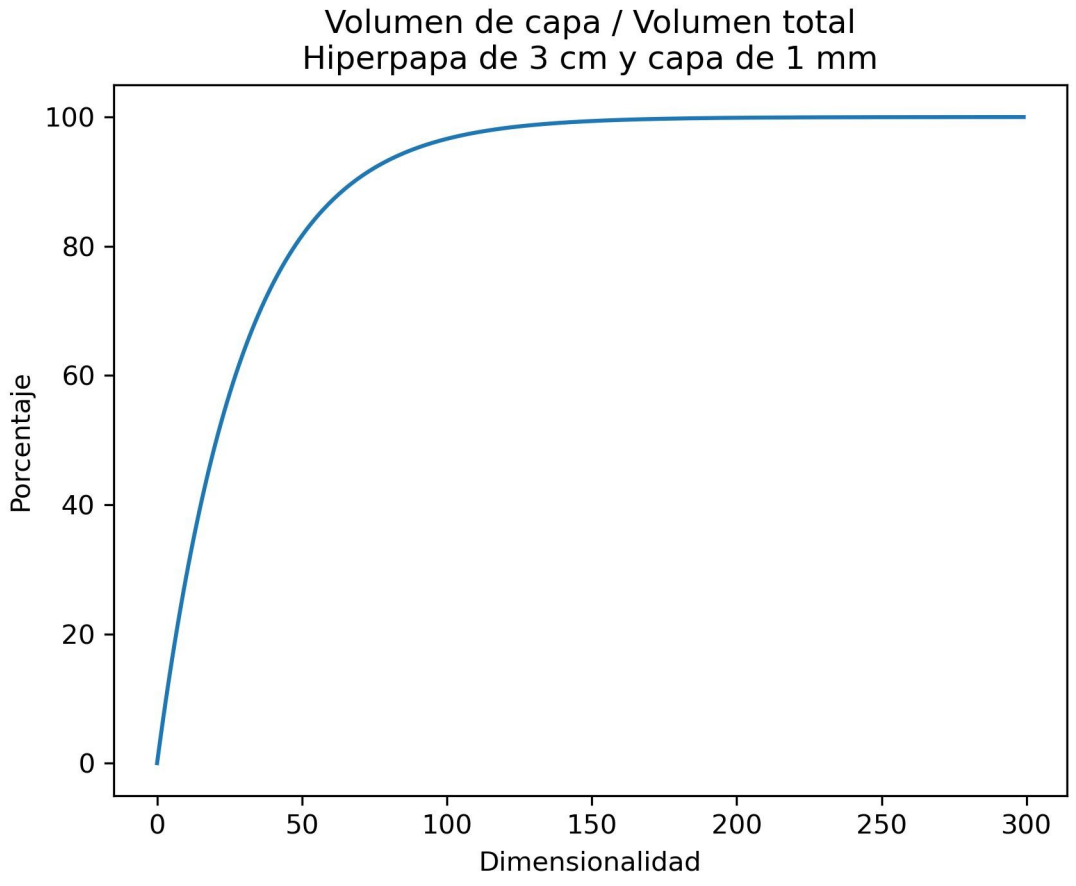

DO NOT
ENTER



¿Comerías una papa multidimensional pelada o sin pelar?

¿Qué otras "paradojas" encontrarás en un hiperespacio?

- **Fenómeno de la distancia de concentración**
- Paradoja de Borel
- Separabilidad en espacios de alta dimensionalidad
- Maldición de la dimensionalidad
- Concentración de la medida
- ...


Volumen de capa / Volumen total
Hiperpapa de 3 cm y capa de 1 mm
Porcentaje
100
80
60
40
20
0
0
50
100
150
200
250
300
Dimensionalidad

Si tu sistema es multidimensional...

Prueba a clusterizar en más dimensiones...

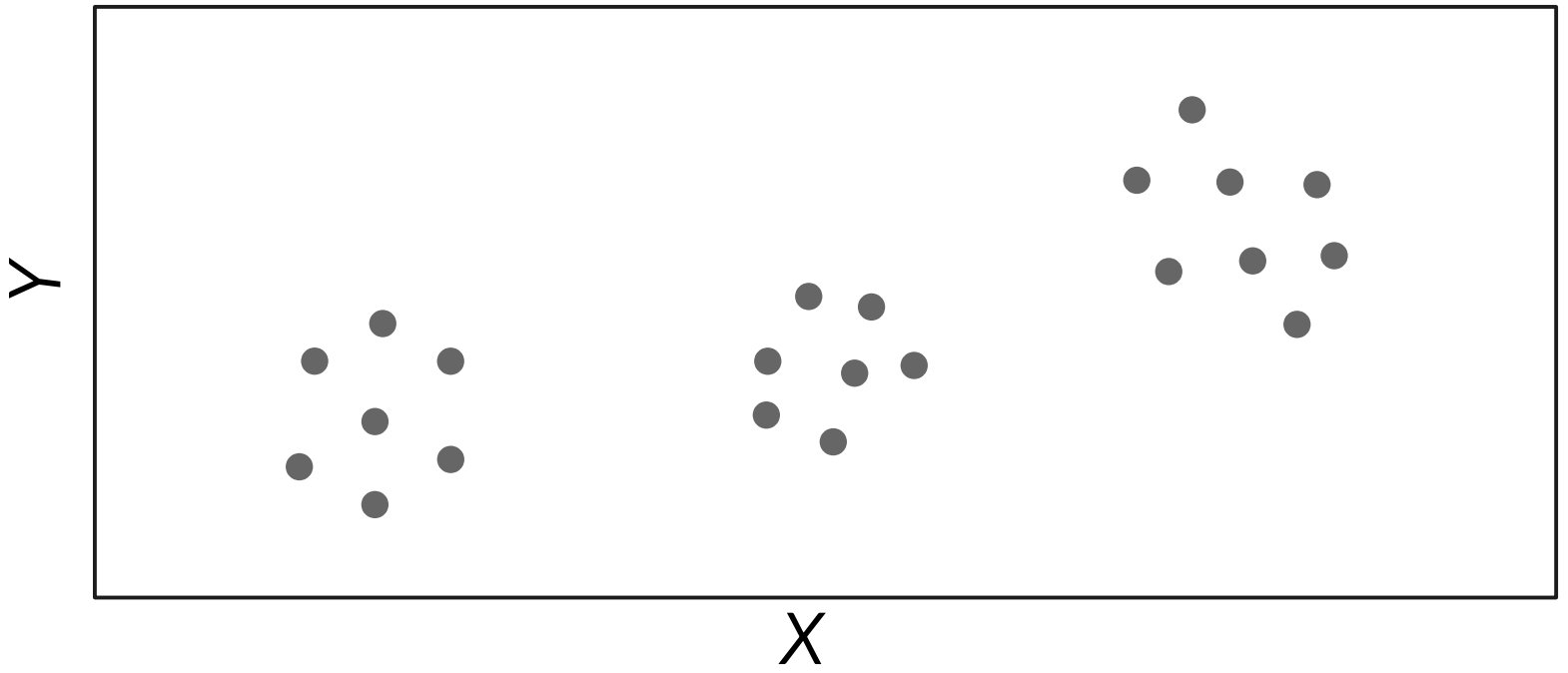
Algoritmos de clustering

Algunos algoritmos de clustering

- *K-means*
- *Hierarchical clustering*
- *DBSCAN*
- *Gaussian Mixture Models*
- *Agglomerative Clustering*
- ...

K-means

Sea una población de individuos descritos por varias características



Pseudo-código del algoritmo de K-means

1. Elije el número de centroides k.
2. Inicializa de forma aleatoria las coordenadas de los k centroides en el espacio de coordenadas de tus individuos.
3. Itera:
    4. Asigna cada individuo al centroide más cercano.
    5. Calcula las nuevas coordenadas de cada centroide como el centro geométrico del conjunto de individuos que le fueron asignados en el paso 4.
    6. Calcula la distancia que se desplazó cada centroide.
    7. Sal de la iteración si todos los desplazamientos calculados en 6 son prácticamente nulos.
8. Identifica los k clusters resultantes como los k conjuntos de individuos asignados a cada centroide.

Ningún algoritmo de clustering es perfecto

Ningún algoritmo de clustering es perfecto

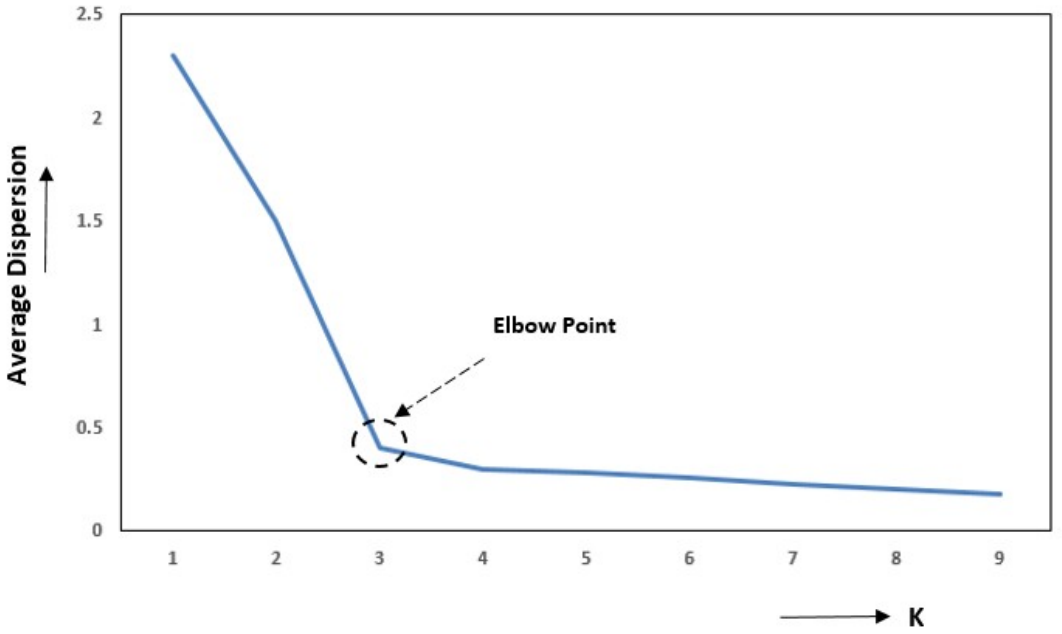
2 problemas de K-means...

1. El algoritmo no “detecta” el número de clusters

- *Elbow method*
- *Silhouette score*
- *Gap statistic*
- *Cross-validation*
- *Varianza total explicada*
- …

1. El algoritmo no “detecta” el número de clusters

· *Elbow method*



2. El algoritmo no es determinista

(los clusters pueden depender de la elección inicial de las coordenadas de los centroides)

- *Multiples ejecuciones minimizando la dispersión intra-cluster*
- *K-means++*

Otras versiones de K-means más sofisticadas

- *Mini-Batch K-means*
- *Fuzzy C-means*
- *K-medians*
- *K-medoids*
- *Weighted K-means*
- *Constrained K-means*

K-means en Python

K-means pasito a pasito

https://ciencia-computacional-himfg.github.io/Clustering/build/html/contents/k_means/k_means_paso_a_paso.html

K-means con Scikit-learn

https://ciencia-computacional-himfg.github.io/Clustering/build/html/contents/k_means/k_means_scikit_learn.html

¿Te vas a atrever...?

Más documentación y foro técnico de soporte en:
github.com/Ciencia-Computacional-HIMFG/Clustering