Inferencia Estadística y Reconocimiento de Patrones

UNaB, Lic. Cs. de Datos, 2021 cuat. 2
Sebastián Pedersen (sebastian.pedersen (at) unab (punto) edu (punto) ar)

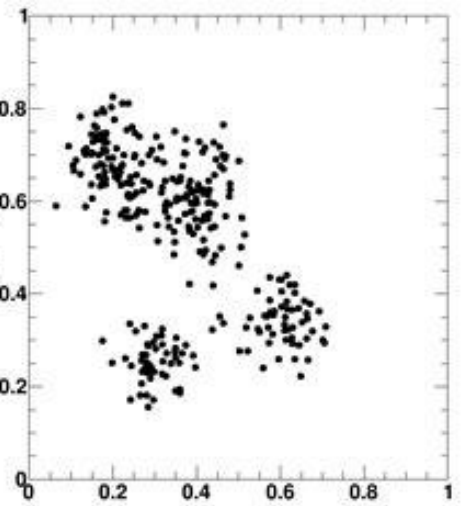
Más Clustering

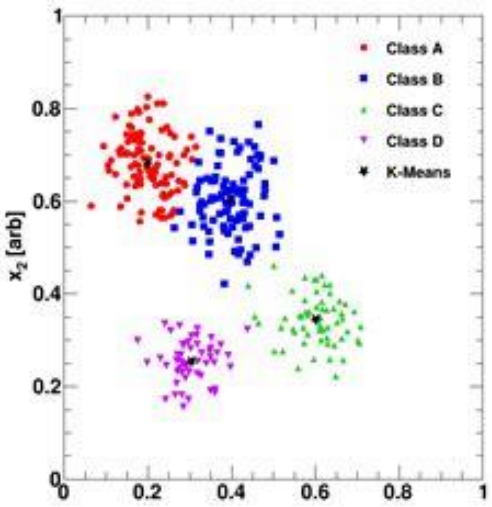
(Clasificación NO supervisada)

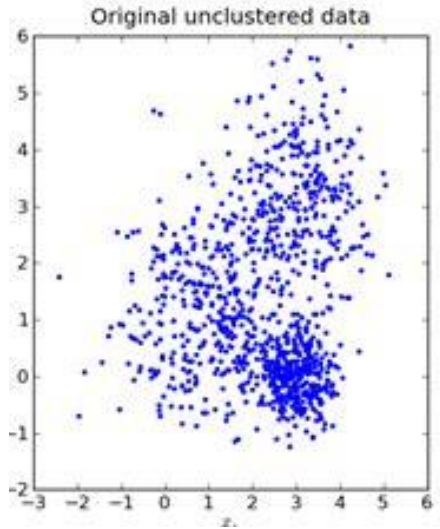
Clasificación No Supervisada

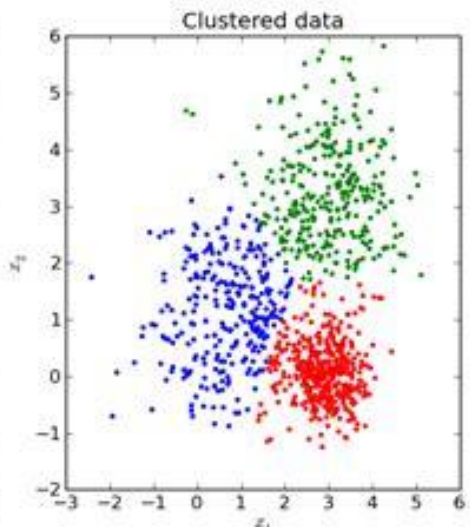
Se tiene un conjunto de datos SIN clasificar.

Objetivo: extraer agrupamientos o estructuras internas, en lo posible con interpretación.









Silhouette score

Para cada dato Xi se calcula:

$$s_i = \frac{b_i - a_i}{\text{máx}\{a_i, b_i\}}$$

- ai es el promedio de las distancias de Xi contra los datos de su propio cluster.
- bi es el mínimo promedio de las distancias de Xi contra los datos de un cluster distinto al suyo.

- Si ai es menor que bi, entonces Xi parece estar bien asignado al cluster al que pertenece, y el si vale 1 - ai/bi.
- Si ai es mayor que bi, entonces Xi parece no estar tan bien asignado al cluster al que pertenece, y el si vale -1 + bi/ai.

El Silhouette score es el promedio de todos los si.

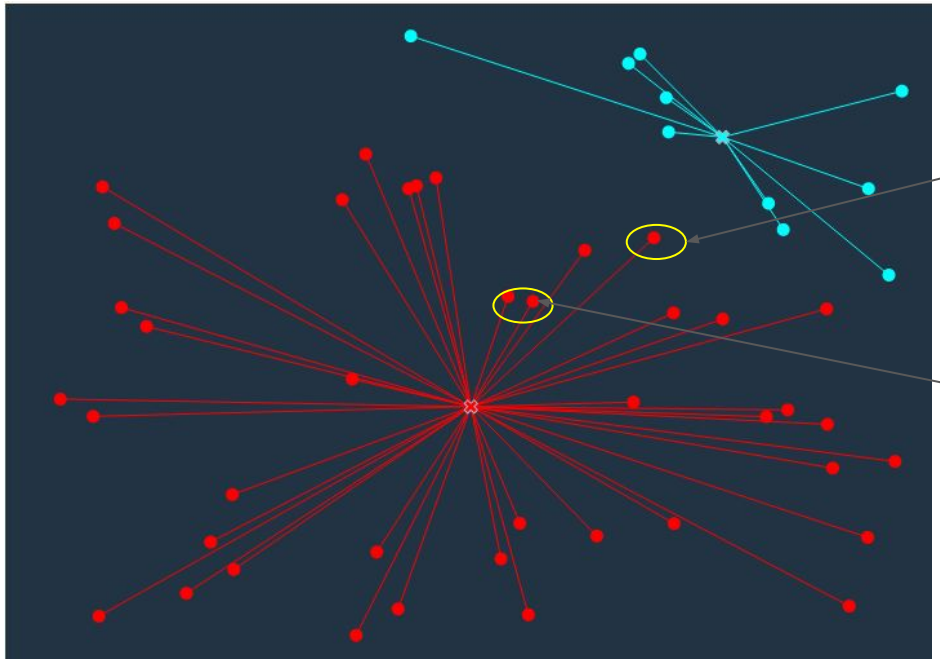
Silhouette score

Para cada dato Xi se calcula:

$$s_i = \frac{b_i - a_i}{\max\limits_{1 \leq i \leq n} \{a_i, b_i\}}$$

El Silhouette score es el promedio de todos los si.

- ai es el promedio de las distancias de Xi contra los datos de su propio cluster.
- bi es el promedio de las distancias de Xi contra los datos cluster más cercano distinto al suyo.



Este Xi parece estar en promedio más cerca a los puntos del cluster más cercano al que NO pertenece.

Este Xi parece estar en promedio más cerca a los puntos de su propio cluster.

Inercia

$$\sum_{i=0}^{n} \min_{\mu_j \in C} (||x_i - \mu_j||^2)$$

Puntos

Centroides

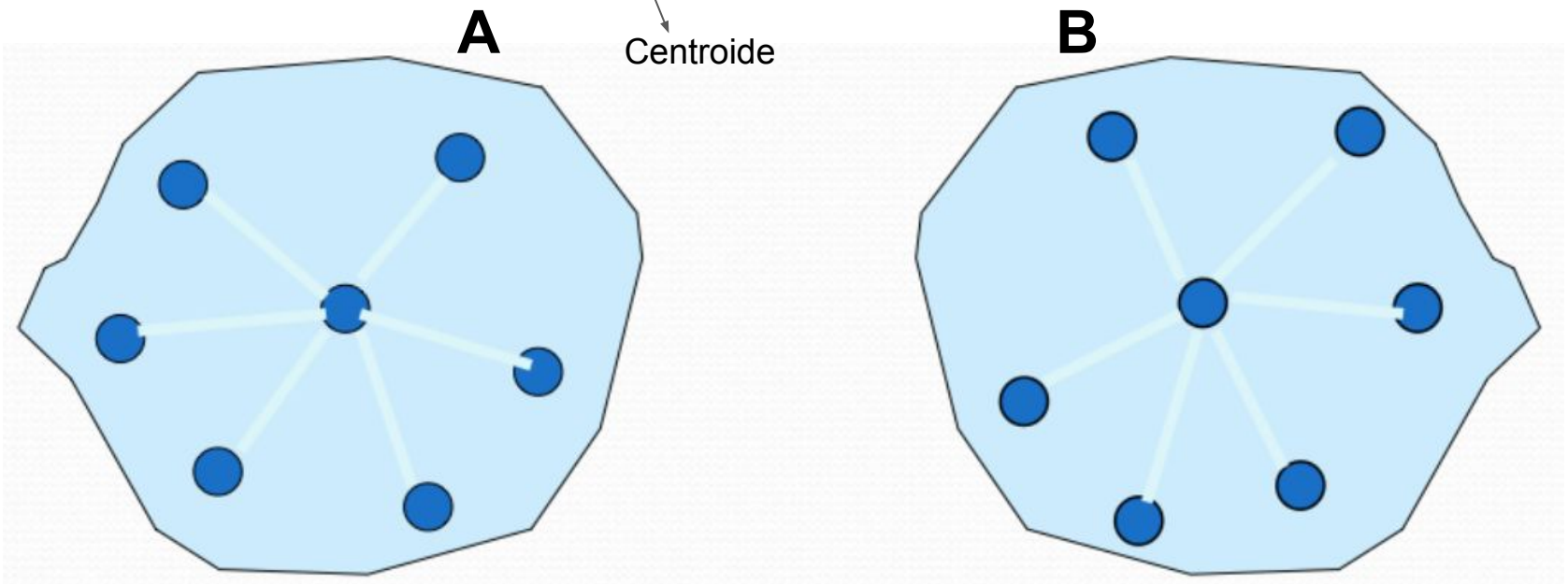
Observar que la convergencia de k-means asegura que ese mínimo se realice para el cluster al que el Xi pertenece.

Mejor clusterización cuanto menor inercia.

Ward linkage para jerárquico aglomerativo: los clusters que se unen son los que minimizan la distancia Ward.

$$\Delta(A,B) = \sum_{i \in A \bigcup B} ||\vec{x_i} - \vec{m}_{A \bigcup B}||^2 - \sum_{i \in A} ||\vec{x_i} - \vec{m}_A||^2 - \sum_{i \in B} ||\vec{x_i} - \vec{m}_B||^2$$

Centroide



Referencias

- Hastie, Tibshirani, Introduction to Statistical Learning, sección 12.4
- Chan, Análisis Inteligente de Datos, sección 10.2