K-Medias Segunda parte, Consideraciones en la implementación.

Pontificia Universidad Javeriana
Francisco Carlos Calderon Ph.D
2020

Facultad de ingeniería
Depto de electrónica

Pontificia Universidad JAVERIANA Bogotá

Objetivos

Identificar dos métricas que permitan seleccionar el número de grupos K del K-Medias con criterio.

Identificar una métrica que permita obtener de múltiples corridas de K-Medias un candidato a un mínimo Global.

Selección del K

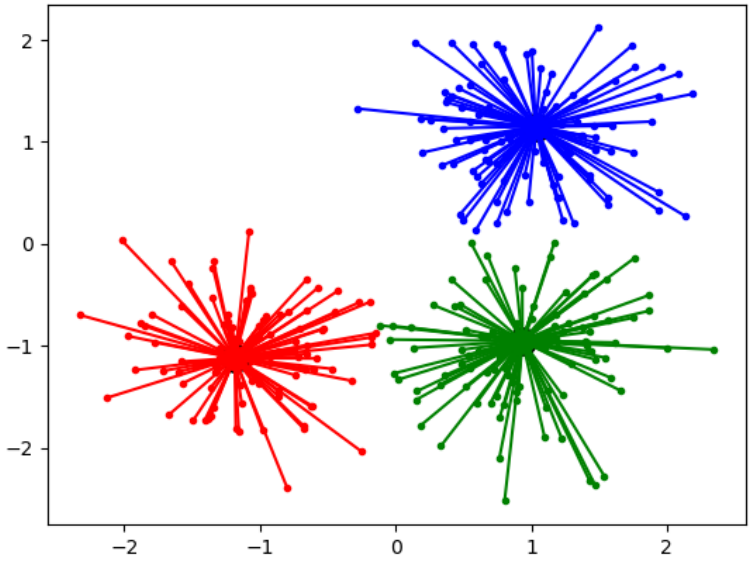
Existen dos métodos ampliamente usados para medir el desempeño de un K en específico.

- Codo de medición de inercia.
- Máximo por medición de la silueta

Inercia, o criterio de la suma de los cuadrados de distancia entre grupos

Se define la inercia, como la suma de todas las distancias que hay entre los centroides de cada grupo a cada miembro del grupo

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



Consideraciones de la Inercia

El criterio asume que la solución global del problema es :

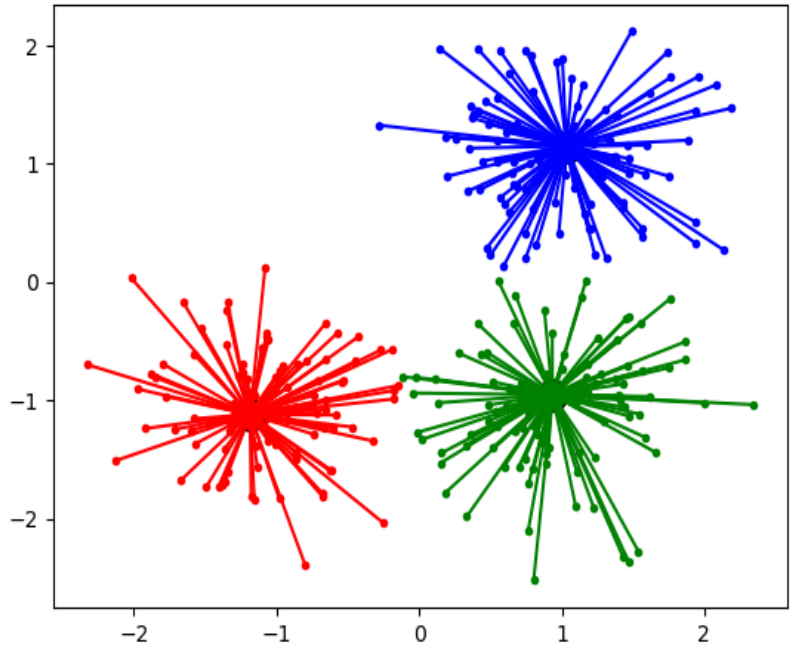
1. Convexa: un solo mínimo global! Si yo se!
2. Isotrópica: matriz de covarianza unitaria, que es decir X independientes.

Esto no siempre se cumple, i.e clusters alongados no radiales y que forman grupos diferentes a hiperesferas

Consideraciones de la Inercia

La inercia no es una métrica normalizada.

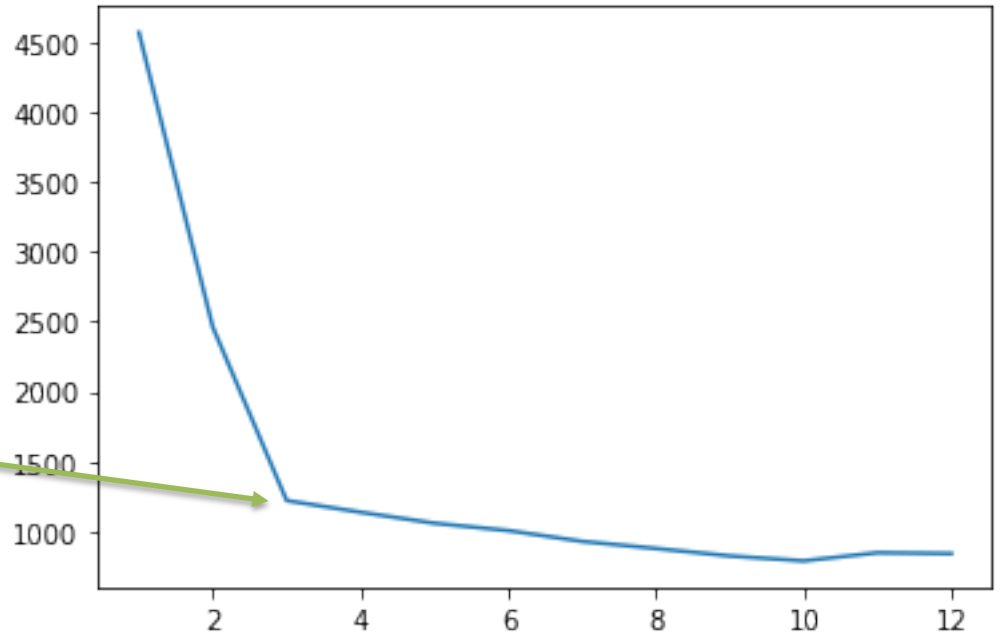
- Dependerá de los grupos y su tamaño comparativo entre ellos.
- Dependerá de la cantidad de dimensiones de X.



Codo de medición de inercia.

Se halla la métrica de inercia para cada valor de K.

Un “buen” valor de K será el dado por el codo de la curva de inercia vs K. Así:



Máximo por medición de la silueta

El valor de silueta mide qué tan similar es un punto a su propio grupo (***cohesión)***, en comparación con otros grupos (***separación)***.

Se tiene una medida de la distancia dentro del grupo (a) y Una medida de la distancia entre una muestra y la muestra más cercana del grupo del que esta muestra no es parte. (b)

Máximo por medición de la silueta

Para un punto i que pertenece al grupo C_i se define la ***cohesión*** a(i) como:

$$a(i) = \frac{1}{|C_i| - 1} \sum_{j \in C_i, i \neq j} d(i,j)$$

Donde d(i,j) es la distancia entre dos elementos del grupo C_i y $|C_i| - 1$ es el número de elementos del grupo menos 1. Luego definimos la disimilitud media del punto i con algún grupo C_k como la media de la distancia desde i a todos los puntos en C_k. La ***separación*** es entonces la mínima distancia media de i a todos los puntos en cualquier otro grupo, del cual i no es miembro

$$b(i) = \min_{k \neq i} \frac{1}{|C_k|} \sum_{j \in C_k} d(i,j)$$

La silueta para un punto i se define como:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}, \text{ if } |C_i| > 1$$

$$s(i) = 0, \text{ if } |C_i| = 1$$

$$-1 \leq s(i) \leq 1$$

$$SC = \max_k \tilde{s}\,(k)$$

Donde $\tilde{s}\,(k)$ representa la media de s(i) sobre el dataset para un numero especifico de grupos k

https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316801

Finding Groups in Data: An Introduction to Cluster Analysis, Leonard Kaufman

Consideraciones de la silueta:

S(i) será cercano a 1 si a(i) <<< b(i).

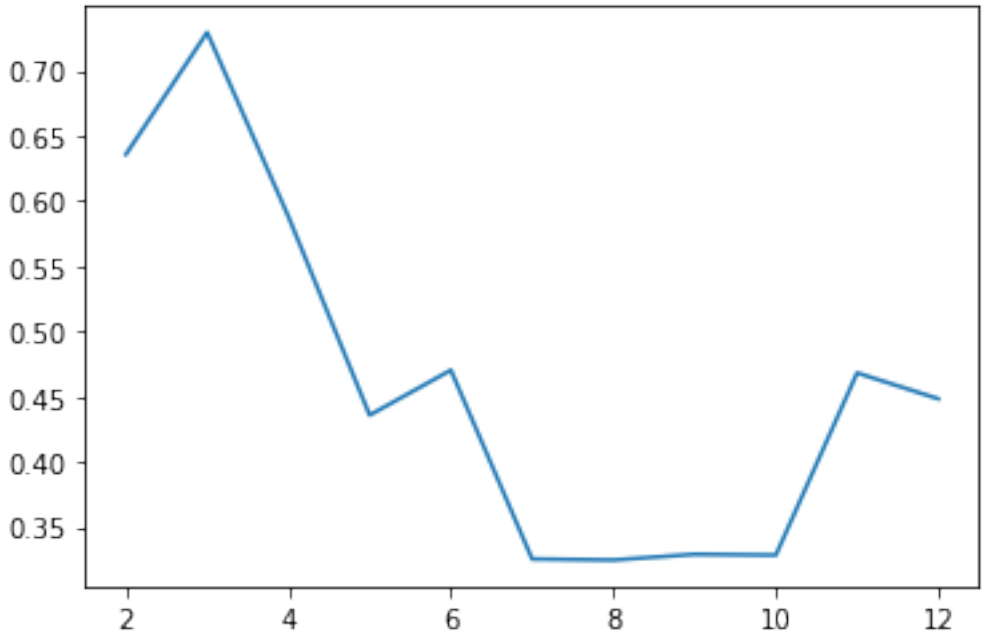
Como a(i) es una medida de lo diferente que es i con su propio grupo, un valor pequeño significa que está bien emparejado. ☺

Un b(i) grande implica que i no coincide con su grupo vecino. ☺

Máximo por medición de la silueta

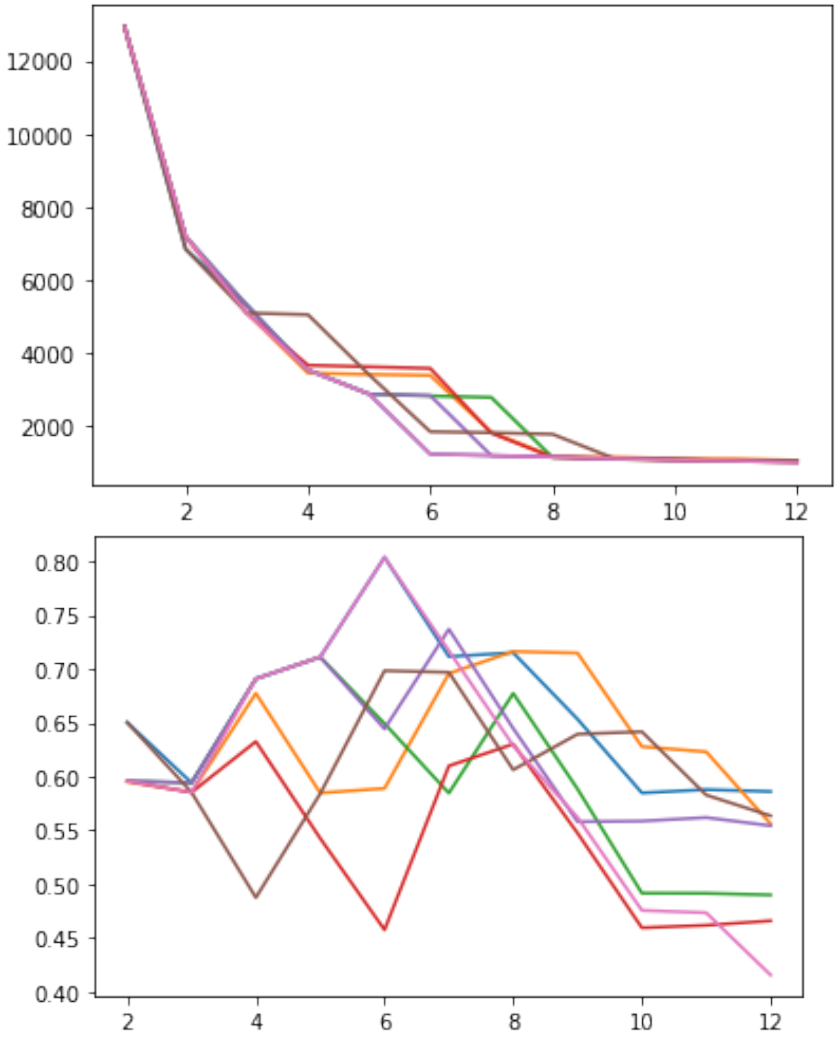
Busque el máximo valor obtenido del coeficiente de silueta para diferentes k.

Ese valor de k será el mejor de acuerdo a la métrica.



Y cómo se que llegué a un mínimo global:

1. Parta de alguna de las métricas vistas (inercia, silueta)
2. Corra varias veces el algoritmo de K-means con diferentes inicializaciones.
3. Tome el mejor resultado para cada k según su métrica.



Ejercicio en clase

- Implementar los métodos vistos en la presentación.