

Sistemas de Gestión de Datos y de la Información. Práctica 2.

K_means.

Alberto Lorente y Hristo Ivanov

17 de diciembre de 2015

Resumen

Evaluación de la calidad del *clustering* para valores de k entre 2 y 20.

1. Resultados

Para empezar hacemos una pequeña aclaración de que las líneas verticales del gráfico representan el máximo y mínimo, mientras que la línea azul respresenta la media ponderada.

Cuando el valor mínimo para una k es 0, es porque algun *cluster* tiene un solo elemento, incluso un *cluster* puede llegar a quedarse vacío.

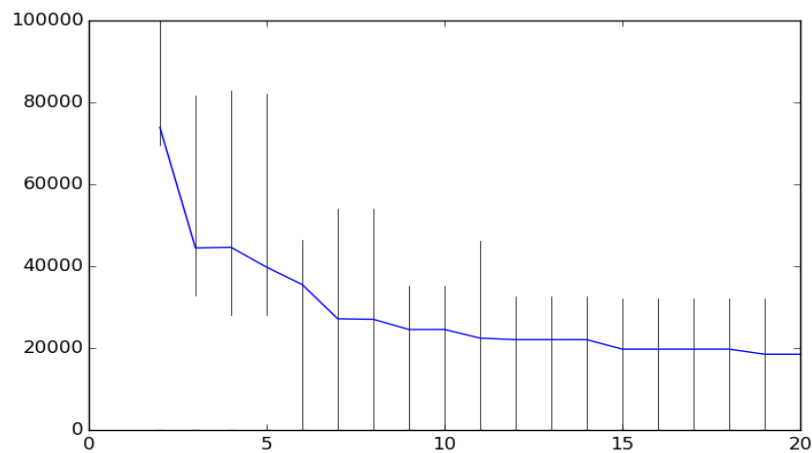


Figura 1: Radios

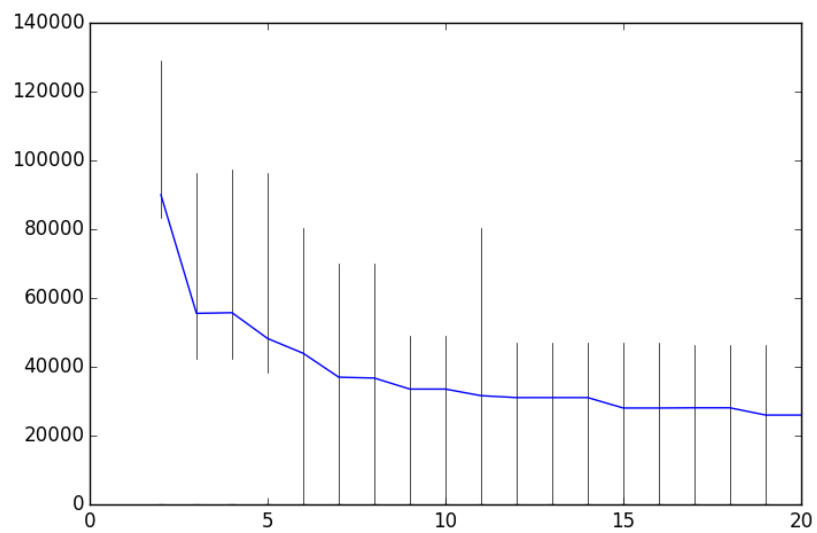


Figura 2: Diametros

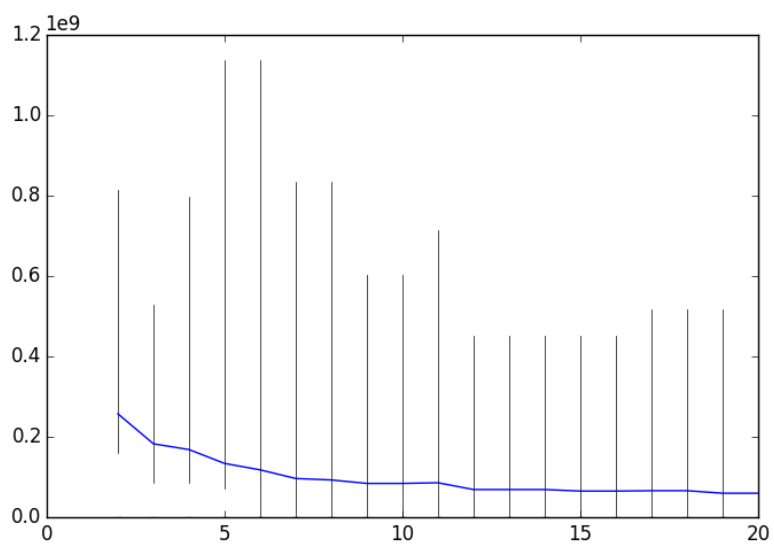


Figura 3: Distancia

2. Conclusiones

Como podemos ver en las figuras anteriores, cuanto menor es el k peor es la calidad de los *clusters* con respecto a su Radio, Distancia y Diámetro. Según aumentamos el número de *clusters* podemos observar como mejora la calidad de los mismo.

No obstante a partir de $k=7$ podemos observar que la ganancia de cohesión de los *clusters* no mejora de forma significativa. Además con $k=7$ tan solo tenemos un *cluster* con un solo elemento. Para $k>7$ el número de *clusters* que tienen un único elemento o ninguno aumenta.

Basándonos en lo anterior concluimos que mejor opción es $k=7$.