Tarea 5 :: Aglomeración

Jung Hwan **Bak** 2016193299
Daniel **Rojas Marín** 2016xxxxxx

Criterios de Evaluación

Time	Medición del tiempo de ejecución de los algoritmos utilizados.
Inertia	Suma de las distancias cuadradas de cada objeto del cluster a un centroide (punto medio de los objetos dentro del cluster).
Homo	Verifica que todos los clusters están formados sólo de los mismos de una clase.
Compl	Abreviación de completeness. Valida que todos los miembros de una clase sean asignados al mismo cluster.
V-Meas	Media armónica entre la homogeneidad y el completeness.
ARI	Adjusted Rand Index. Requiere el conocimiento de los labels originales. Calcula la similitud entre 2 clases al evaluar todos los pares de muestras y contar pares que se asignan a una misma clase o en clases diferentes según lo predicho y verdadero. Se busca un valor cercano a 0.0 para el label aleatorio independientemente del número de clases y muestras y exactamente 1.0 cuando las clases son idénticas, hasta una permutación.
AMI	Adjusted Mutual Information. Es un ajuste de la puntuación de información mutua (MI). Es más alto, generalmente, para dos clases con un mayor número de agrupaciones, independientemente de si realmente se comparte más información. La métrica es independiente de los valores absolutos de los labels. Una permutación de los valores de los labels de cluster no cambiará el valor de la puntuación de ninguna manera. Es simétrica.
Silhouette	Silhouette Coefficient. Criterio de evaluación utilizado cuando no se tienen true labels para la evaluación del modelo. Se utiliza el modelo para que realice una evaluación en sí misma.

Modos de inicialización

k-means++	Inicializa los centroides para que estén equidistantes de sus vecinos. Esto mejora el resultado en términos probabilísticos en comparación con la inicialización de valores al azar.
Random	Mejora el tiempo de ejecución en comparación al k-means++. Se escoge filas, las k observaciones, al azar como valores para los centroides iniciales.
PCA	Menor tiempo de ejecución en comparación a los otros modos de inicialización. Utiliza la descomposición del dataset multivariable con el fin de crear un nuevo dataset perfectamente ortogonal. Esto permite la observación de manera cuantificativa la máxima varianza. El método traslada a un centro a los datos, pero no los escala.