Tarea 3 Aglomeracion - Información

Abraham Arias Chinchilla, Lenin Torres Valverde.
ipseabraham@gmail.com, ttvleninn@gmail.com

I. Criterios de evaluación

¿Que significa cada criterio de evaluación?
En métodos supervisados se predice que tan preciso es gracias a que se puede saber conocer si clasifica bien o mal, pero en aglomeración no se está clasificando nada, mas bien agrupando. Las métricas surgen para poder determinar que tan bueno es el cluster mismo.

A. Adjusted Rand index (ARI)

Este criterio mide la semejanza entre los criterios de semejanza y desacuerdo, se necesita un set con agrupaciones tomadas como ciertas. No toma en cuenta permutaciones de los datos entre los dos sets y normaliza de una forma aleatoria. Con mejor puntuación 1 y peor -1. Lamentablemente se necesita conocer la agrupación verdadera.

B. Mutual Information Based Scores & Adjusted Mutual Information (AMI)

Dado un set de agrupaciones correctas se mide la concordancia entre estos sets, ignorando permutaciones. De acuerdo al tipo de normalización se pasa a la forma AMI o Normalized Mutual Information (NMI).

C. Homogeneity, completeness and V-measure

Usando análisis de entropía condicional y conociendo las agrupaciones verdaderas se pueden definir:

1) Homogeneidad: Que tanto un cluster solo contiene datos de su propia clase.
2) Completeness: Que tanto del 0 al 1, se puede decir que todos los miembros de una clase estan dentro de la clase que corresponde.
3) V-measure: Similar a NMI pero normzalizada por la suma de las entropías de cada conglomerado

D. Silhouette Coefficient

Si los conglomerados verdaderos no son conocidos, se debe de evaluar el modelo mismo. Se define por cada muestra de la forma:

La distancia media entre una muestra y todos los demás puntos en la misma clase, menos, La distancia media entre una muestra y todos los demás puntos en el siguiente grupo más cercano.

II. Modo de inicializacion

¿Que hace cada modo de inicializacion?

Este algoritmo es sensible a esta decisión. Se encargar de escoger los centroides iniciales.

1) Random: Ejecutar múltiples veces con semillas distintas, selección aleatoria de k de los datos.
2) K-means++: Solo el primer centroide es random, el que sigue se coloca lo más largo posible del actual. Con una probabilidad para los pesos para saber cúal sigue.