Tarea 5 - Aglomeración, descripciones

Esteban Agüero Pérez, Esteban Sanabria Villalobos
estape11@gmail.com, esteasvtec@gmail.com
Instituto Tecnológico de Costa Rica
Escuela de Ing. Electrónica
EL-5852 Introducción al Reconocimiento de Patrones
Profesor: Dr. Pablo Alvarado
Cartago, 23 de octubre, 2019

I. Criterios de evaluación

I-A. *Homogeneity score*

Se satisface homogeneidad si todos los clústers contienen puntos de datos que solo son miembros de una sola clase [1].

I-B. *Completeness score*

Se satisface si todos los puntos de datos de una clase dada son miembros del mismo clúster [2].

I-C. *V measure*

Es la media armónica entre *completeness* y *homogeneity* [3].

I-D. *Adjusted Rand Index*

Calcula la medida de similitud entre dos agrupaciones, considerando todos los pares de muestras y contando los pares que están asignados en el mismo o en diferentes clústers, en la predicha y la verdadera agrupación [4].

I-E. *Adjusted Mutual Information*

Parte del hecho de que el MI (*mutual information*) es generalmente más alto para dos agrupaciones con un mayor número de agrupaciones, independientemente de si realmente se comparte más información [5].

I-F. *Silhouette coefficient*

Es calculado usando la distancia media del intra-cluster y la distancia media del cluster más cercano para cada elemento [6].

II. Modo de inicialización

II-A. *k-means++*

Este busca solucionar el obstáculo de k-means relacionado con que la aproximación encontrada podría ser mala respecto a la objetiva comparada con la agrupación óptima [7]. En pocos palabras lo que realiza es tomar un centro tomado de forma aleatoria (uniformente dada) del conjunto de X, luego elige el nuevo centro con la probilidad dada por:

$$\frac{D(x)^2}{\Sigma_{x\epsilon X} D(x)^2} \tag{1}$$

Donde $D(x)$ denota la distancia más corta entre el punto y el centro más cercado que ya ha sido seleccionado. Este proceso se realiza hasta tener k centros. Después se prosigue con el algoritmo de *k-means* regular.

II-B. *Random*

Básicamente elige k observaciones de forma aleatoria como centroides iniciales [7].

II-C. *PCA-based*

Se le da de forma determinista las semillas de los centros y se ejecuta únicamente una vez el algoritmo de *k-means* [7].

Referencias

[1] Scikit Learn. $sklearn.metrics.homogeneity\_score$ — *Documentation*. Recuperado de: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.homogeneity_score.html
[2] Scikit Learn. $sklearn.metrics.completeness\_score$ — *Documentation*. Recuperado de: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.completeness_score.html
[3] Scikit Learn. $sklearn.metrics.v\_measure\_score$ — *Documentation*. Recuperado de: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.v_measure_score.html
[4] Scikit Learn. $sklearn.metrics.adjusted\_rand\_score$ — *Documentation*. Recuperado de: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_rand_score.html
[5] Scikit Learn. $sklearn.metrics.adjusted\_mutual\_info\_score$ — *Documentation*. Recuperado de: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.adjusted_mutual_info_score.html
[6] Scikit Learn. $sklearn.metrics.silhouette\_score$ — *Documentation*. Recuperado de: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.silhouette_score.html
[7] Scikit Learn. $sklearn.cluster.KMeans$ — *Documentation*. Recuperado de: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html