Data Science Project 4 Question 1 and 2: k-means and fuzzy c means clustering and their validation explanation

Yasin Boloorchi

July 28, 2020

1 k-means clustering

K-means clustering is a common algorithm for students to learn clustering the data and it's an entry-level of clustering approach. the k-means whole concept is about calculating means, K time; so the first step should be selecting the number of clusters that we want to identify in our data. (for example, K can be 3). in the next step, randomly select 3 distinct data points. next, measure the distance between the first point and the three initial clusters and then assign the point to the nearest cluster. after that all of the points are in clusters, we go on to the last step; calculate the mean of each cluster and then we repeat what we just did (measure and cluster) using the mean values until clustering do not change at all.

2 Fuzzy c means clustering

Fuzzy-c-mean algorithm is a clustring approach just like K-means (c in here is just like K in K-means), but with the difference of the clusters shape; by that i mean that unlike k-means, a record or a single data can be in multiple clusters at a time.

3 NMI validation

NMI indicates the similarity of identified clusters concerning ground truth clusters. is defined as below

$$NM(A,B) = \frac{-2\sum_{i=1}^{CA}\sum_{j=1}^{CB} Cij\ log(CijN/Ci*C*j)}{\sum_{i=1}^{CA} Ci*\ \ log(Ci*/N)\ +\sum_{i=1}^{CA} C*j\ log(C*j/N)} \quad (1)$$

where and show two cluster sets. In this paper, refers to the ground truth clusters and shows the results of a clustering algorithm. and are the number of groups in partitions and , respectively. Cij is the number of common members between $i-th$ cluster of A and $j-th$ cluster of B. $Cj* = \sum_{j=1}^{CB} Cij$ (resp. $C*j = \sum_{i=1}^{CA} Cij$) is the sum of elements of in its $i-th$ row (resp. column j), and N is the number of pints. if the algorithm generates teh clustering results as same as the real clusters,NMI takes the value of 1, and if two clusters results are completely different, the NMI will be equal to 0. NMI is a well-known and widely used clustering validation metric that estimate the quality of the clustering in comparisons with a given clustering results

4 Rand Index validatoin

Rand Index(RI) is a well-known measure which is used to compare the results of clustering methods and defined as below:

$$RandIndex(A,B) \quad = \frac{(a+b)}{\binom{n}{2}} \tag{2}$$

where n is the number of all data points, is the number of pairs in A and B belong to same clusters and b is the number of pairs that are in different groups. The value of RI is 1 while A and B are completely consisting of same groups and its value is 0 when A and B are completely different.