tut9

Monday, July 26, 2021 5:45 PM

Conceptual

1. Prove:

$$\sum_{\boldsymbol{x}_i \in C_k} d_E(\boldsymbol{x}_i, \bar{\boldsymbol{x}}_{C_k})^2 = \frac{1}{2|C_k|} \sum_{\boldsymbol{x}_i \in C_k} \sum_{\boldsymbol{x}_j \in C_k} d_E(\boldsymbol{x}_i, \boldsymbol{x}_j)^2.$$

$$\begin{aligned}
&\frac{1}{2|C_k|} \sum_{\boldsymbol{x}_i \in C_k} \sum_{\boldsymbol{x}_j \in C_k} d_E(\boldsymbol{x}_i, \boldsymbol{x}_j)^2 \\
=&\frac{1}{2|C_k|} \sum_{\boldsymbol{x}_i \in C_k} \sum_{\boldsymbol{x}_j \in C_k} \sum_{l=1}^{p} (\boldsymbol{x}_{il} - \boldsymbol{x}_{jl})^2 \\
=&\frac{1}{2|C_k|} \sum_{\boldsymbol{x}_i \in C_k} \sum_{\boldsymbol{x}_j \in C_k} \sum_{l=1}^{p} ((\boldsymbol{x}_{il} - \bar{\boldsymbol{x}}_{C_k,l}) - (\boldsymbol{x}_{jl} - \bar{\boldsymbol{x}}_{C_k,l}))^2 \\
=&\frac{1}{2|C_k|} \sum_{\boldsymbol{x}_i \in C_k} \sum_{\boldsymbol{x}_j \in C_k} \sum_{l=1}^{p} ((\boldsymbol{x}_{il} - \bar{\boldsymbol{x}}_{C_k,l})^2 + (\boldsymbol{x}_{jl} - \bar{\boldsymbol{x}}_{C_k,l})^2 - 2(\boldsymbol{x}_{il} - \bar{\boldsymbol{x}}_{C_k,l})(\boldsymbol{x}_{jl} - \bar{\boldsymbol{x}}_{C_k,l})) \\
=&\frac{1}{2|C_k|}|C_k| \sum_{\boldsymbol{x}_i \in C_k} \sum_{l=1}^{p} (\boldsymbol{x}_{il} - \bar{\boldsymbol{x}}_{C_k,l})^2 + \frac{1}{2|C_k|}|C_k| \sum_{\boldsymbol{x}_j \in C_k} \sum_{l=1}^{p} (\boldsymbol{x}_{jl} - \bar{\boldsymbol{x}}_{C_k,l})^2 \\
&- \frac{1}{|C_k|} \sum_{\boldsymbol{x}_i \in C_k} \sum_{l=1}^{p} (\boldsymbol{x}_{il} - \bar{\boldsymbol{x}}_{C_k,l}) \sum_{\boldsymbol{x}_j \in C_k} (\boldsymbol{x}_{jl} - \bar{\boldsymbol{x}}_{C_k,l}) \\
=&\frac{1}{2} \sum_{\boldsymbol{x}_i \in C_k} d_E(\boldsymbol{x}_i, \bar{\boldsymbol{x}}_{C_k})^2 + \frac{1}{2} \sum_{\boldsymbol{x}_j \in C_k} d_E(\boldsymbol{x}_j, \bar{\boldsymbol{x}}_{C_k})^2 - 0 \\
=& \sum_{\boldsymbol{x}_i \in C_k} d_E(\boldsymbol{x}_i, \bar{\boldsymbol{x}}_{C_k})^2
\end{aligned}$$

This shows that minimising the sum of squared Euclidean distances for each cluster is equivalent to minimising the within-cluster variation for each cluster.

2. Given the one-dimensional data set below, perform the K-means clustering manually. In particular, set K to be 2 and use the first two observations as the initial cluster centroids.

Obs.	X_1
1	1
2	3
3	4
4	8
5	9
6	11
7	12

. The first step of the K-means clustering algorithm (Lloyd's algorithm) is to randomly select $K = 2$ observations as initial centroids. In this case, we choose the first two observations, and let's call them cluster A (including observation 1) and cluster B (including observation 2). Next, we will iterate between assigning each observation to its closest centroid and computing new centroids as cluster means. Therefore, in order to assign the observation, we will need to first compute the squared Euclidean distance between all observations and the current centroids; the result is shown in the table below.

	cluster A	cluster B
1	0	4
2	4	0
3	9	1
4	49	25
5	64	36
6	100	64
7	121	81

From the table, we see that observation 1 will remain in cluster A and all other observations belong to cluster B. The new centroids will then be 1 for cluster A and $(3+4+8+9+11+12)/6 \approx 7.83$ for cluster B. Now we calculate the squared Euclidean distance again with respect to the new centroids.

	cluster A	cluster B
1	0	46.65
2	4	23.33
3	9	14.67
4	49	0.03
5	64	1.37
6	100	10.05
7	121	17.39

Observations 1-3 will be grouped into cluster A and observations 4-7 will be grouped into cluster B. The new centroids are $(1+3+4)/3 \approx 2.67$ for cluster A and $(8+9+11+12)/4 = 10$ for cluster B. Now, if we compute the squared Euclidean distance again, we see that the allocation of observations to clusters will not change. In other words, the K-means clustering algorithm converges.

	cluster A	cluster B
1	2.79	81
2	0.11	49
3	1.77	36
4	28.41	4
5	40.07	1
6	69.39	1
7	87.05	4

Applied

1. Perform hierarchical agglomerative clustering, K-means clustering and K-medoids clustering on the `iris` data set, assuming the class label `Species` is unavailable. In particular, answer the following questions:

   (a) How can you decide the optimal number of clusters for each method?

   (b) When the optimal number is chosen for each method, which method tends to generate the best clustering results?

   (c) How can you compare the clustering results with the ground-truth class labels?

```
#-----------------------------------------------------------------------------
# STATS5099 Data Mining Tutorial 9
#-----------------------------------------------------------------------------
# Applied Question 1
#-----------------------------------------------------------------------------
Iris <- iris[,-c(5)]
Iris <- scale(Iris)
###############################
# Select the number of cluster #
###############################
set.seed(1)
# Hierarchical agglomerative clustering
library(factoextra)
ggplot_fviz <-fviz_nbclust(USArrests,FUN=hcut,method="silhouette")
ggplot_fviz
#If the average silhouette width is used as the evaluation criterion,
#then we will select two clusters.
ggplot_fviz <-fviz_nbclust(USArrests,FUN=hcut,method="gap_stat")
ggplot_fviz
#If the Gap statistic is used as the evaluation criterion,
#then we will select five clusters.
# K-means clustering
ggplot_fviz <-fviz_nbclust(USArrests,FUN=kmeans,method="silhouette")
ggplot_fviz #2 clusters are suggested
ggplot_fviz <-fviz_nbclust(USArrests,FUN=kmeans,method="gap_stat")
ggplot_fviz #3 clusters are suggested
# K-medoids clustering
ggplot_fviz <-fviz_nbclust(USArrests,FUN=cluster::pam,method="silhouette")
ggplot_fviz #2 clusters are suggested
ggplot_fviz <-fviz_nbclust(USArrests,FUN=cluster::pam,method="gap_stat")
ggplot_fviz #6 clusters are suggested
###################################
# Evaluate clustering performances #
###################################
set.seed(1)
#Suppose we select 2 clusters for all three methods and
#compare the clustering results using the silhouette width
Iris.HAC <- hclust(dist(Iris))
Iris.HAC.clus <- cutree(Iris.HAC,k=2)
Iris.km <- kmeans(Iris,centers=2,nstart=50)
Iris.km.clus <- Iris.km$cluster
library(cluster)
Iris.PAM <- pam(Iris,k=2,nstart=50)
Iris.PAM.clus <- Iris.PAM$clustering
Iris.HAC.si <- silhouette(Iris.HAC.clus, dist(Iris))
Iris.km.si  <- silhouette(Iris.km.clus, dist(Iris))
Iris.PAM.si <- silhouette(Iris.PAM.clus, dist(Iris))
windows()
par(mfrow=c(1,3))
plot(Iris.HAC.si)
plot(Iris.km.si)
plot(Iris.PAM.si)
#K-means and K-medoids generate same clustering results;
#their average silhouette width is larger than that of HAC.
##############################################
# Compare clustering results with class labels #
##############################################
#For more details, see supplementary material on Week 9 P17
library(fpc)
HAC_baseline <- cluster.stats(
  dist(Iris),as.numeric(iris$Species),Iris.HAC.clus)
HAC_baseline
km_baseline <- cluster.stats(
  dist(Iris),as.numeric(iris$Species),Iris.km.clus)
km_baseline
PAM_baseline <- cluster.stats(
  dist(Iris),as.numeric(iris$Species),Iris.PAM.clus)
PAM_baseline
print(c(HAC_baseline$corrected.rand,
        km_baseline$corrected.rand,
        PAM_baseline$corrected.rand))
#For example, we could compare the cluster allocations with
#class labels based on adjusted Rand index. Larger values
#suggest higher consistency between the pair of points
#from clustering and the pair from class labels.
```