Homework 1

16.9.21, Chuan Lu

Problem 1. Rewrite the target function of K-means in matrix form.

Solution. $f(x) = argmin_{\mathcal{S}} \sum_{i=1}^{k} \sum_{x_j \in S_j} \sum_{1 \le m \le p} |x_{jl} - x_{jm}|^2$. □

Problem 2. Cluster $Data_i.csv$ with K-means, judge the number of clusters, and compare differences between different evaluating methods.

Solution. The code of K-means, evaluating and test scripts can be found from attachments(kmeans.R, evaluate.R, hw1.2.R). □

Result. For Data1.csv, we cluster with k = 3. The result is shown as below.

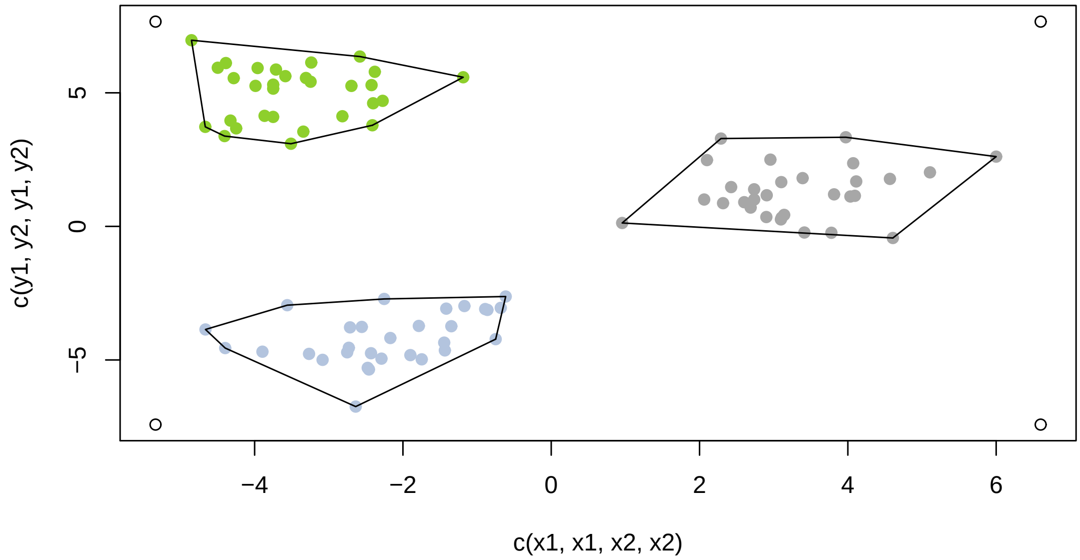
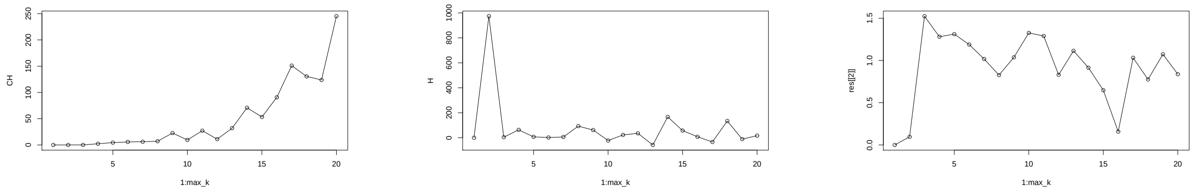


Figure 1: The clusters of Data1.csv with k = 3

For choosing k, the result of Data1.csv is shown below; The first is the result of Calinski-Harabasz method, the second Hartigan method and the last Gap Statistic.

The result of CH method is 9, if we may add a limit that $k \le 10$. The result of H method is 3, and result of GAP statistic is 3.

For Data2.csv, since each data point is of 3 dims, we cannot plot its clustering result here.



(a) CH method

(b) H method

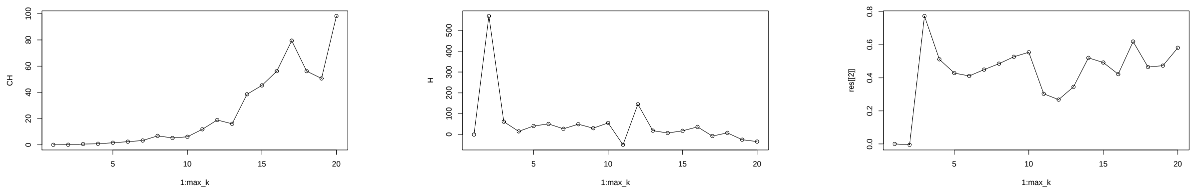
(c) GAP statistic

Figure 2: The three evaluating methods of choosing k for Data1.csv

But after deploying the evaluating methods we consider k = 3. The type of datas can be find in attachments(CLUSTER_DATA2.txt).

The result of CH method is 8, if we limit that $k \leq 10$. The result of H method and result of GAP statistic are 3.

The pics are shown below.



(a) CH method

(b) H method

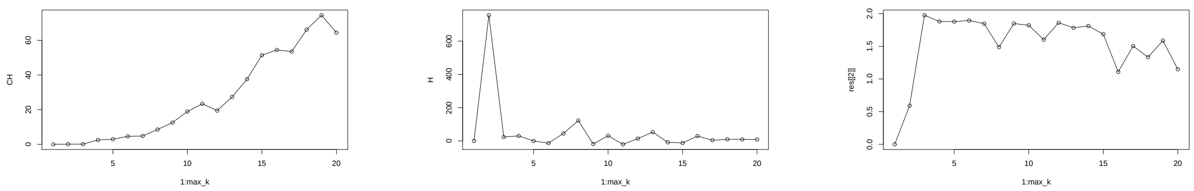
(c) GAP statistic

Figure 3: The three evaluating methods of choosing k for Data2.csv

For Data3.csv, we fix k = 3 after evaluating k with H method and GAP statistic; The result of clustering can be found in attachments(CLUSTER_DATA3.txt).

The result of CH method is 9, if we limit that $k \leq 10$. The result of H method and result of GAP statistic are 3.

The pics are shown below.



(a) CH method

(b) H method

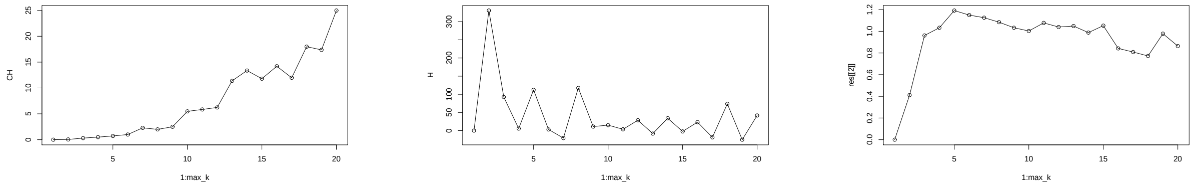
(c) GAP statistic

Figure 4: The three evaluating methods of choosing k for Data3.csv

For Data4.csv, we found k = 3 after evaluating k with H method and GAP statistic; The result of clustering can be found in attachments(CLUSTER_DATA4.txt).

The result of CH method is 7, if we limit that $k \leq 10$. The result of H method and result of GAP statistic are 3.

The pics are shown below.



(a) CH method

(b) H method

(c) GAP statistic

Figure 5: The three evaluating methods of choosing k for Data4.csv

□

Analysis. It is suprising that result of Calinski-Harabasz method does not agree with the other methods. I consider it as the result that I misunderstood the meaning of $W(k), B(k)$. I calculate the former as the target function of cluster; the latter as $\sum_{c_i,c_j \in C} \|c_i - c_j\|^2$, in which C is the set of cluster centers. □

Problem 3. Use hierarchical clustering methods to cluster $Data_i.csv$

Solution. The code of hierarchical clustering and the test script can be found from attachments(hierarchical_clustering.R, hw1.3.R). □

Result. It is somehow embarrassing that I still did't know how to draw trees in R without $hcluster()$. The result can be shown is the attachments, (HC_DATA1.txt and HC_DATA2.txt).

Each result is a (n-1) * narg matrix, in which n is the number of data points and narg is the number of dims in each data point. Each row in the matrix(for example, A[1,]) means a step when two points merged with each other; A[1, 1] and A[1, 2] merged into a new A[1, 1], and the point represented by A[1, 2] is abandoned.

It is possible to draw the tree and do Tree-Cut with the results. □