Finding Optimal Number of Clusters

TOP: Data Clustering 076/091

Instructor: Sayan Bandyapadhyay

Portland State University

Outline

Number of Clusters

Needed for common clustering models such as k-means/median/center

Number of Clusters

Needed for common clustering models such as k-means/median/center

- Elbow method
- Calinski-Harabasz index
- Silhouette method
- Gap statistics

Number of Clusters

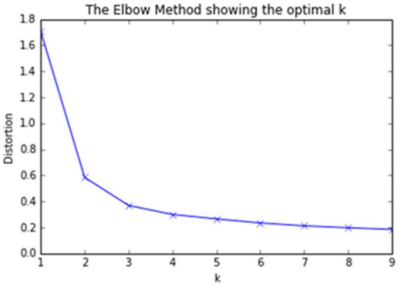
Needed for common clustering models such as k-means/median/center

- Elbow method
- Calinski-Harabasz index
- Silhouette method
- Gap statistics

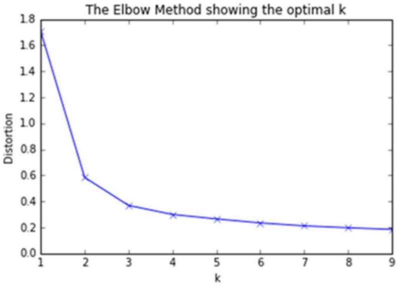
Psuedo-science alert!

Outline

Finding Elbow of a Graph



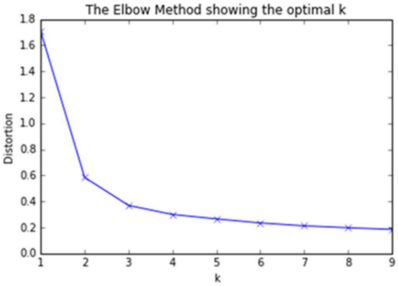
Finding Elbow of a Graph



Advantages

- Simplicity of computation

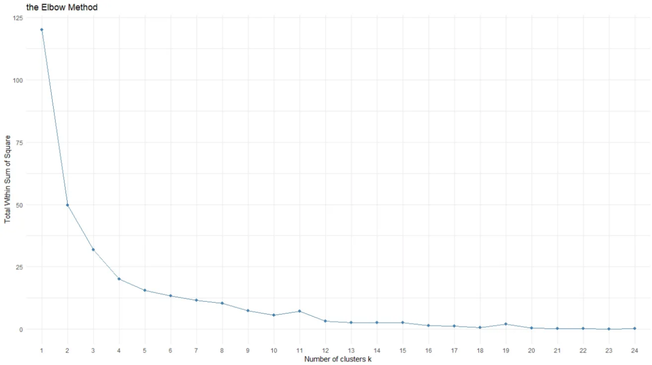
Finding Elbow of a Graph



Advantages

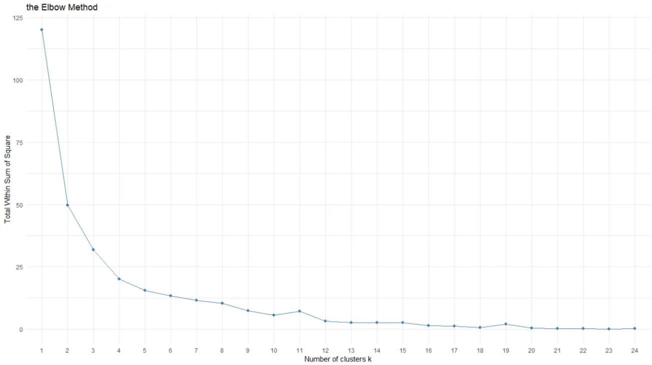
- Simplicity of computation
- Can work for any clustering model

Disadvantages



- Subjectivity in interpretation

Disadvantages



- Subjectivity in interpretation
- Can have multiple elbows: ambiguity with complex datasets

Outline

CH Index

$$CH(k) = \frac{\mathrm{BCSS}(k)}{k-1} \cdot \frac{n-k}{\mathrm{WCSS}(k)}$$

CH Index

$$CH(k) = \frac{\mathrm{BCSS}(k)}{k-1} \cdot \frac{n-k}{\mathrm{WCSS}(k)}$$

$$\mathrm{BCSS}(k) = \sum_{i=1}^{k} n_i \cdot ||c_i - c||^2$$

$C = \{c_1, c_2, \ldots, c_k\}$ are centers, c is the global center, n_i is i-th cluster-size

CH Index

$$CH(k) = \frac{\mathrm{BCSS}(k)}{k-1} \cdot \frac{n-k}{\mathrm{WCSS}(k)}$$

$$\mathrm{BCSS}(k) = \sum_{i=1}^{k} n_i \cdot ||c_i - c||^2$$

$C = \{c_1, c_2, \ldots, c_k\}$ are centers, c is the global center, n_i is i-th cluster-size

$$\mathrm{WCSS}(k) = \text{k-means-cost}(C, X)$$

CH Index

$$CH(k) = \frac{\mathrm{BCSS}(k)}{k-1} \cdot \frac{n-k}{\mathrm{WCSS}(k)}$$

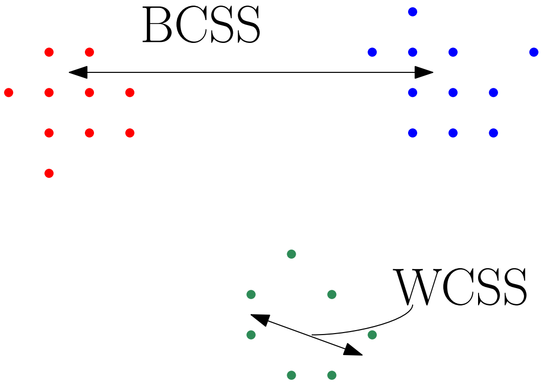
$$\mathrm{BCSS}(k) = \sum_{i=1}^{k} n_i \cdot ||c_i - c||^2$$

$C = \{c_1, c_2, \ldots, c_k\}$ are centers, c is the global center, n_i is i-th cluster-size

$$\mathrm{WCSS}(k) = \text{k-means-cost}(C, X)$$

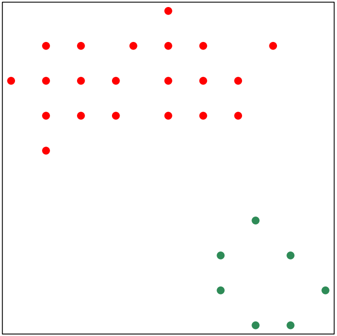
The higher is the better

Well-separated Clusters

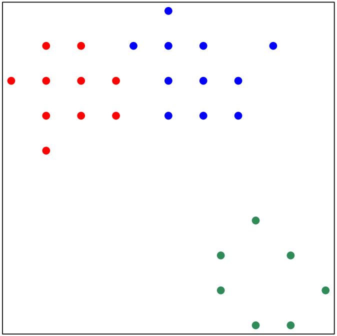


BCSS should be large and WCSS small

Why BCSS?

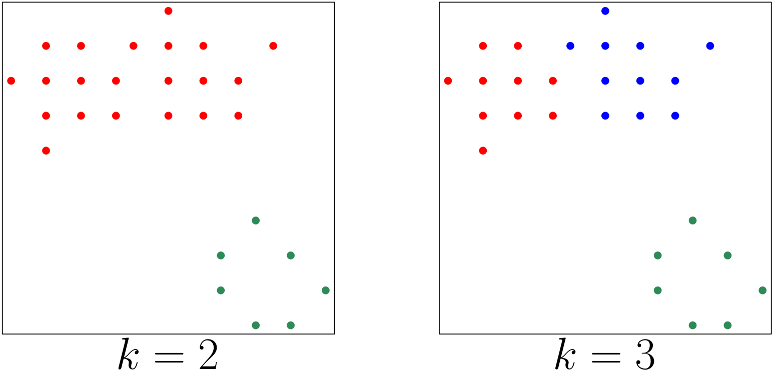


$k = 2$



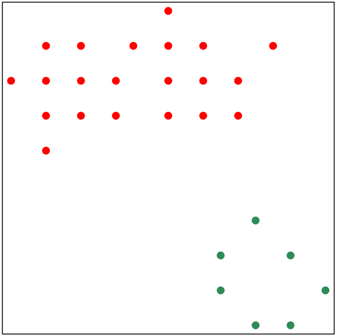
$k = 3$

Why BCSS?

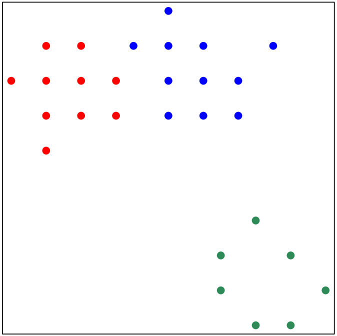


CH-index might stop unnecessary splitting of clusters

Why BCSS?



$k = 2$



$k = 3$

CH-index might stop unnecessary splitting of clusters—>
assumes clusters are well-separated

Properties

Pros

- Objective measure
- Fast computation

Properties

Pros

- Objective measure
- Fast computation

Cons

- Sensitivity to cluster shape
- Limited interpretability

Outline

Silhouette Coefficient

For any point x_i,

$$S(x_i) = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

a_i is proxy for WCSS; b_i is proxy for BCSS

Silhouette Coefficient

For any point x_i,

$$S(x_i) = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

a_i is proxy for WCSS; b_i is proxy for BCSS

a_i is the average distance of x_i from all other points in its cluster

Silhouette Coefficient

For any point x_i,

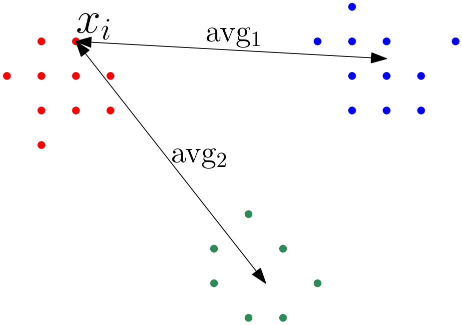
$$S(x_i) = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

a_i is proxy for WCSS; b_i is proxy for BCSS

a_i is the average distance of x_i from all other points in its cluster

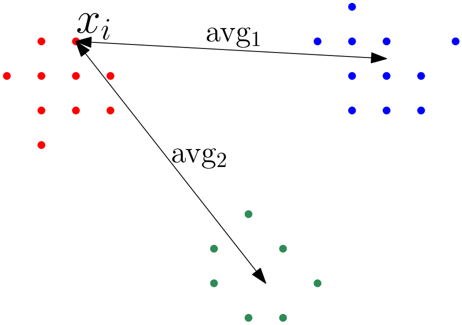
b_i is the smallest average distance of x_i to all points in any other cluster

Silhouette Coefficient



$$b_i = \min\{\text{avg}_1, \text{avg}_2\}$$

Silhouette Coefficient



$$b_i = \min\{\text{avg}_1, \text{avg}_2\}$$

b_i is the minimum average separation from x_i

Silhouette Coefficient

For any point x_i,

$$S(x_i) = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

a_i is proxy for WCSS; b_i is proxy for BCSS

a_i is the average distance of x_i from all other points in its cluster

b_i is the smallest average distance of x_i to all points in any other cluster

Silhouette Coefficient

For any point x_i,

$$S(x_i) = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

a_i is proxy for WCSS; b_i is proxy for BCSS

a_i is the average distance of x_i from all other points in its cluster

b_i is the smallest average distance of x_i to all points in any other cluster

Silhouette Coefficient —> average of the $S(x_i)$ over all points

Silhouette Coefficient

For any point x_i,

$$S(x_i) = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

a_i is proxy for WCSS; b_i is proxy for BCSS

a_i is the average distance of x_i from all other points in its cluster

b_i is the smallest average distance of x_i to all points in any other cluster

Silhouette Coefficient —> average of the $S(x_i)$ over all points —> maximized over all k

Silhouette Coefficient

For any point x_i,

$$S(x_i) = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

a_i is proxy for WCSS; b_i is proxy for BCSS

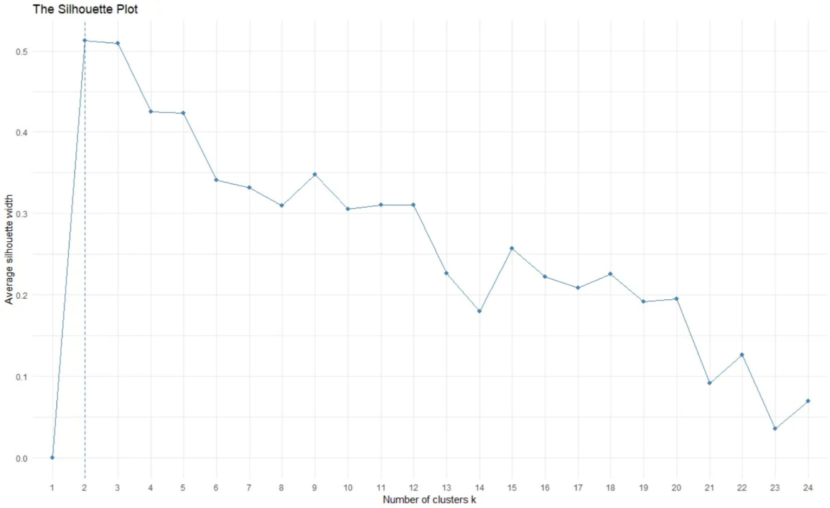
a_i is the average distance of x_i from all other points in its cluster

b_i is the smallest average distance of x_i to all points in any other cluster

Silhouette Coefficient —> average of the $S(x_i)$ over all points —> maximized over all k

Value ranges between -1 and 1: 1 -> well-separated clusters, 0 -> overlapping or ambiguous clusters, negative -> poorly separated data points

Silhouette Plot



Properties

Pros

- Objective measure
- Individual data point assessment
- Intuitive interpretation
- Works with different clustering models

Properties

Pros

- Objective measure
- Individual data point assessment
- Intuitive interpretation
- Works with different clustering models

Cons

- Difficulty with overlapping clusters
- Computational complexity

Outline

Motivation

$$Gap_n(k) = E_n^*\{WCSS(k)\} - WCSS(k)$$

Motivation

$$Gap_n(k) = E_n^*\{WCSS(k)\} - WCSS(k)$$

- The expectation is taken over B samples of size n from a reference null distribution

Motivation

$$Gap_n(k) = E_n^*\{WCSS(k)\} - WCSS(k)$$

- The expectation is taken over B samples of size n from a reference null distribution
- Null $->$ a distribution with no obvious clustering/clusters are not well-separated

Motivation

$$Gap_n(k) = E_n^*\{WCSS(k)\} - WCSS(k)$$

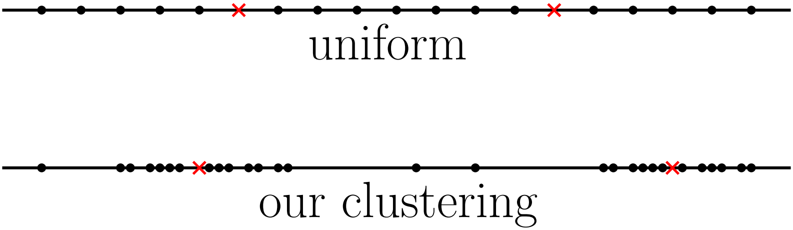
- The expectation is taken over B samples of size n from a reference null distribution
- Null $->$ a distribution with no obvious clustering/clusters are not well-separated
- Uniform distribution is an example $->$ for each feature, pick a value within the range of observed values

Motivation

$$Gap_n(k) = E_n^*\{WCSS(k)\} - WCSS(k)$$

- The expectation is taken over B samples of size n from a reference null distribution
- Null $->$ a distribution with no obvious clustering/clusters are not well-separated
- Uniform distribution is an example $->$ for each feature, pick a value within the range of observed values
- Clustering cost of reference data is expected to be large

Uniform vs Regular Clusters



Motivation

$$Gap_n(k) = E_n^*\{WCSS(k)\} - WCSS(k)$$

- The expectation/mean is taken over B samples of size n from a reference null distribution
- Null $->$ a distribution with no obvious clustering/clusters are not well-separated
- Uniform distribution is an example $->$ for each feature, pick a value within the range of observed values
- Clustering cost of reference data is expected to be large

For optimal k, the gap is expected to be maximized/falls the farthest below the reference curve

The Actual Definition

$$Gap_n(k) = E_n^*\{\log WCSS(k)\} - \log WCSS(k)$$

The Actual Definition

$$Gap_n(k) = E_n^*\{\log WCSS(k)\} - \log WCSS(k)$$

$W_1^*, \ldots, W_B^*$ are sample WCSS

The Actual Definition

$$Gap_n(k) = E_n^*\{\log WCSS(k)\} - \log WCSS(k)$$

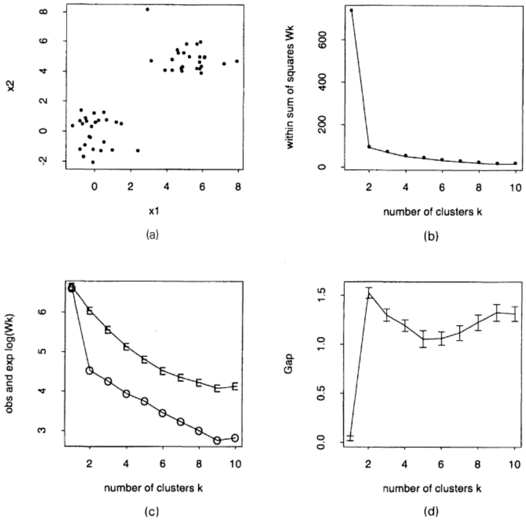
$W_1^*, \ldots, W_B^*$ are sample WCSS

$$E_n^*\{\log WCSS(k)\} = \frac{1}{B}\sum_i \log W_i^* = \frac{1}{B}\log(\Pi_i W_i^*)$$

$$= \log(\Pi_i W_i^*)^{1/B}$$

$$Gap_n(k) = \log \frac{(\Pi_i W_i^*)^{1/B}}{WCSS(k)}$$

An Example



x2
x1
(a)
within sum of squares Wk
number of clusters k
(b)
obs and exp log(Wk)
number of clusters k
(c)
Gap
number of clusters k
(d)

Properties

Pros

- Objective measure
- Statistically grounded assessment of clustering quality
- Relatively robust to noise and outliers

Properties

Pros

- Objective measure
- Statistically grounded assessment of clustering quality
- Relatively robust to noise and outliers

Cons

- Computationally intensive
- Limited applicability to certain datasets
- Lack of consensus on reference distribution

Take Home Message

- No method is perfect

Take Home Message

- No method is perfect
- Try multiple methods; compare and contrast

Take Home Message

- No method is perfect
- Try multiple methods; compare and contrast
- Always remember clustering is an exploratory technique

Links

[Implementation in R]

[Implementation in Python]

[Gap statistics paper]

[CH-index paper]