Machine Learning with Python

Session 12: Cluster Analysis

Arghya Ray

Rational for Measuring Cluster Goodness

- What are the optimal number of cluster to identify?
- How do I measure weather one set of clusters is preferable to another?
- The **silhouette** method and the **pseudo-F statistic** will help us address these questions by measuring cluster goodness.

Concepts Measures Should Address

- Cluster separation represents how distant the clusters are from each other
- Cluster cohesion refers to how tightly related the records within the individual clusters are
- Good measures should incorporate both as do the silhouette and pseudo-F statistic
- However, the sum of squares error (SSE) only accounts for cluster cohesion and is monotonically decreasing with increasing numbers of clusters.

Measuring Cluster Goodness: **The Silhouette Method**

For each data value *i* the silhouette is used to gauge how good the cluster assignment is for that point:

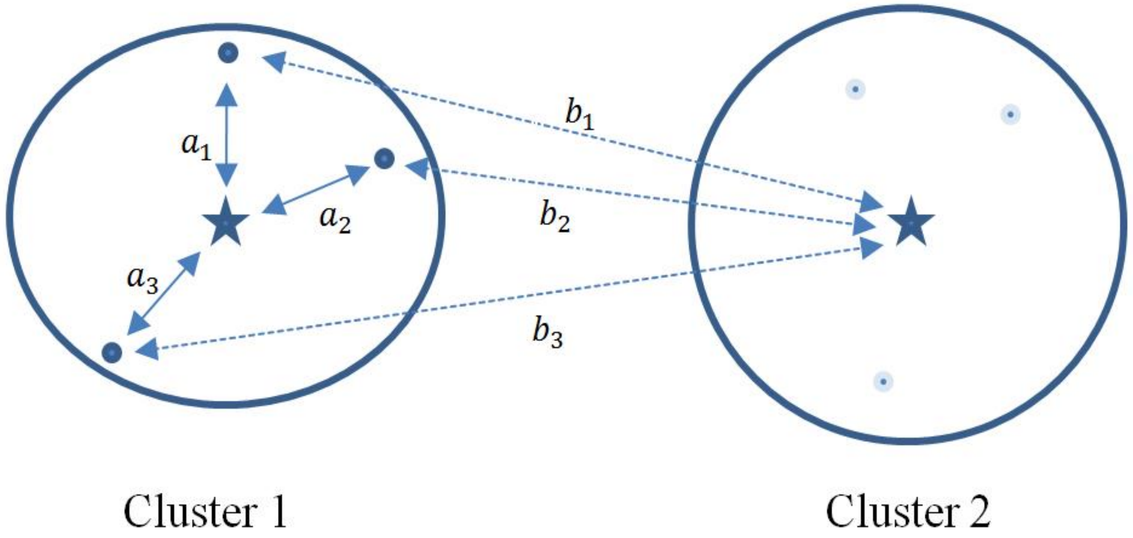
$$\boldsymbol{Silhouette_i = s_i = \frac{b_i - a_i}{\max(b_i, a_i)}}$$

where a_i is the distance between the data value and its cluster center and represents *cohesion*

and b_i is the distance between the data value and the next closest cluster center and represents *separation*

Silhouette Accounts for Separation & Cohesion

Each data value in Cluster 1 has its values of ai and bi represented by solid and dotted lines, respectively



bi > ai for each data value, thus each data value's silhouette is positive, indicating the data are not misclassified

Measuring Cluster Goodness: **The Silhouette Method contd…**

- A positive value indicates that the assignment is good, with higher values better than lower values.
- A value close to zero is considered to be weak since the observation could have been assigned to the next cluster with little negative consequence.
- A negative value is considered to be misclassified since assignment to the next closest cluster would have been better.

The Average Silhouette Value

The average silhouette value over all records yields a measure of how well the cluster solution fits. A thumbnail interpretation, meant as a guide only:

- 0.5 or better provides good evidence of the reality of the clusters in the data
- 0.25 – 0.5 provides some evidence of the reality of the clusters in the data.
- Less than 0.25 provides scant evidence of cluster reality

Silhouette Example (cont.)

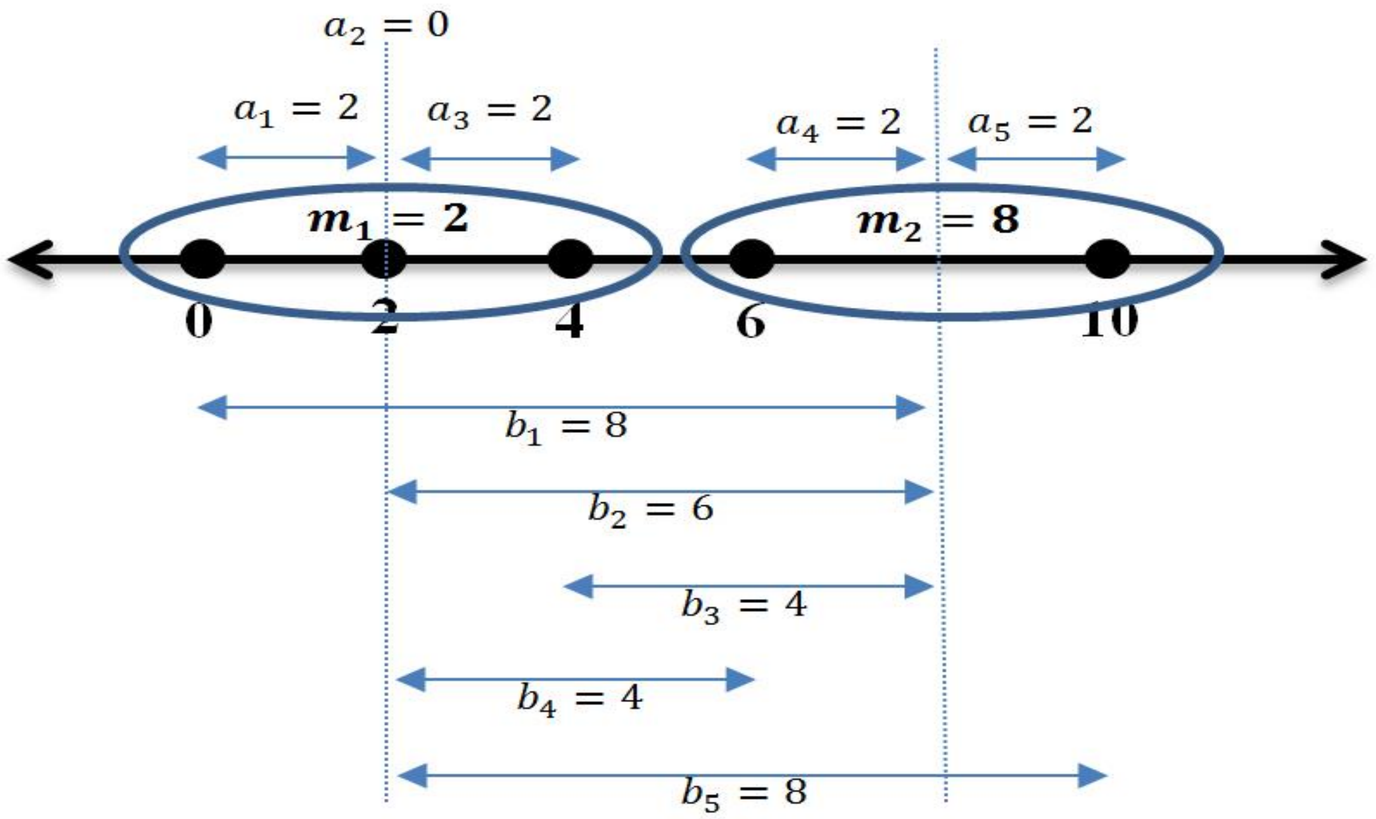
- Apply *k*-means clustering to the following data set:

$x_1 = 0 \quad x_2 = 2 \quad x_3 = 4 \quad x_4 = 6 \quad x_5 = 10$

- The first three data values are assigned to Cluster 1 and the last two to Cluster 2
- Center for Cluster 1 is $m_1 = 2$ and for Cluster 2 is $m_2 = 8$
- Values for a_i are distance between x_i and its cluster center; values for b_i are distance between x_i and the other cluster center

Silhouette Example (cont.)

Distances between the data values and cluster centers:



Silhouette Example (cont.)

Calculations for individual data values:

x_i	a_i	b_i	$Max(a_i, b_i)$	$\text{Silhouette}_i = s_i = \frac{b_i - a_i}{\text{Max}(b_i, a_i)}$
0	2	8	8	$\frac{8-2}{8} = 0.75$
2	0	6	6	$\frac{6-0}{6} = 1.00$
4	2	4	4	$\frac{4-2}{4} = 0.50$
6	2	4	4	$\frac{4-2}{4} = 0.50$
10	2	8	8	$\frac{8-2}{8} = 0.75$
				Mean Silhouette = 0.7

The pseudo-*F* Statistic

Let:

k be number of clusters

$\sum n_i = N$ betotal sample size

x_{ij} refer to the j^{th} data value in the i^{th} cluster

m_i refer to cluster center (centroid) of the i^{th} cluster

M represent the grand mean of all the data

and $Distance(a, b) = \sqrt{\sum (a_i - b_i)^2}$

The pseudo-*F* Statistic (cont.)

Then the *sum of squares between* the clusters is:

$$SSB = \sum_{i=1}^{k} n_i \cdot Distance^2(m_i, M)$$

And the *sum of squares error*, or the *sum of squares within* the clusters is:

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_j} Distance^2(x_{ij}, m_i)$$

And the *pseudo-F statistic* is:

$$F = \frac{MSB}{MSE} = \frac{SSB/k-1}{SSE/N-k}$$

The pseudo-*F* Statistic (cont.)

- The hypotheses being tested are:

$$H_0: \text{There are no clusters in the data.}$$
$$H_a: \text{There are } k \text{ clusters in the data.}$$

- Reject H_o for sufficiently small p-value where:

$$p\text{-}value = P(F_{k-1,n-k}) > \text{Pseudo F value})$$

- The pseudo-F statistic rejects the null hypothesis too easily.

The pseudo-*F* Statistic (cont.)

The pseudo-F statistic should not be used to determine the presence of clusters but can be used to select the optimal number of clusters as follows:

1. Use a clustering algorithm to develop a clustering solution for a variety of values of k.

2. Calculate the pseudo-F statistic and p-value for each candidate, and select the candidate with the smallest p-value as the best clustering solution.

Pseudo-*F* Statistic Example

- Apply *k*-means clustering to the following data set:

$x_1 = 0 \quad x_2 = 2 \quad x_3 = 4 \quad x_4 = 6 \quad x_5 = 10$

- The first three data values are assigned to Cluster 1 and the last two to Cluster 2
- Center for Cluster 1 is $m_1 = 2$ and for Cluster 2 is $m_2 = 8$
- $n_1 = 3$ and $n_2 = 2$ data values, and $N = 5$, the grand mean is $M = 4.4$. And, because we are in one dimension, $Distance(m_i, M) = |m_i - M|$

Pseudo-*F* Statistic Example (cont.)

Then

$$SSB = \sum_{i=1}^{k} n_i \cdot Distance^2(m_i, M)$$
$$= 3 \cdot (2 - 4.4)^2 + 2 \cdot (8 - 4.4)^2 = 43.2$$

And

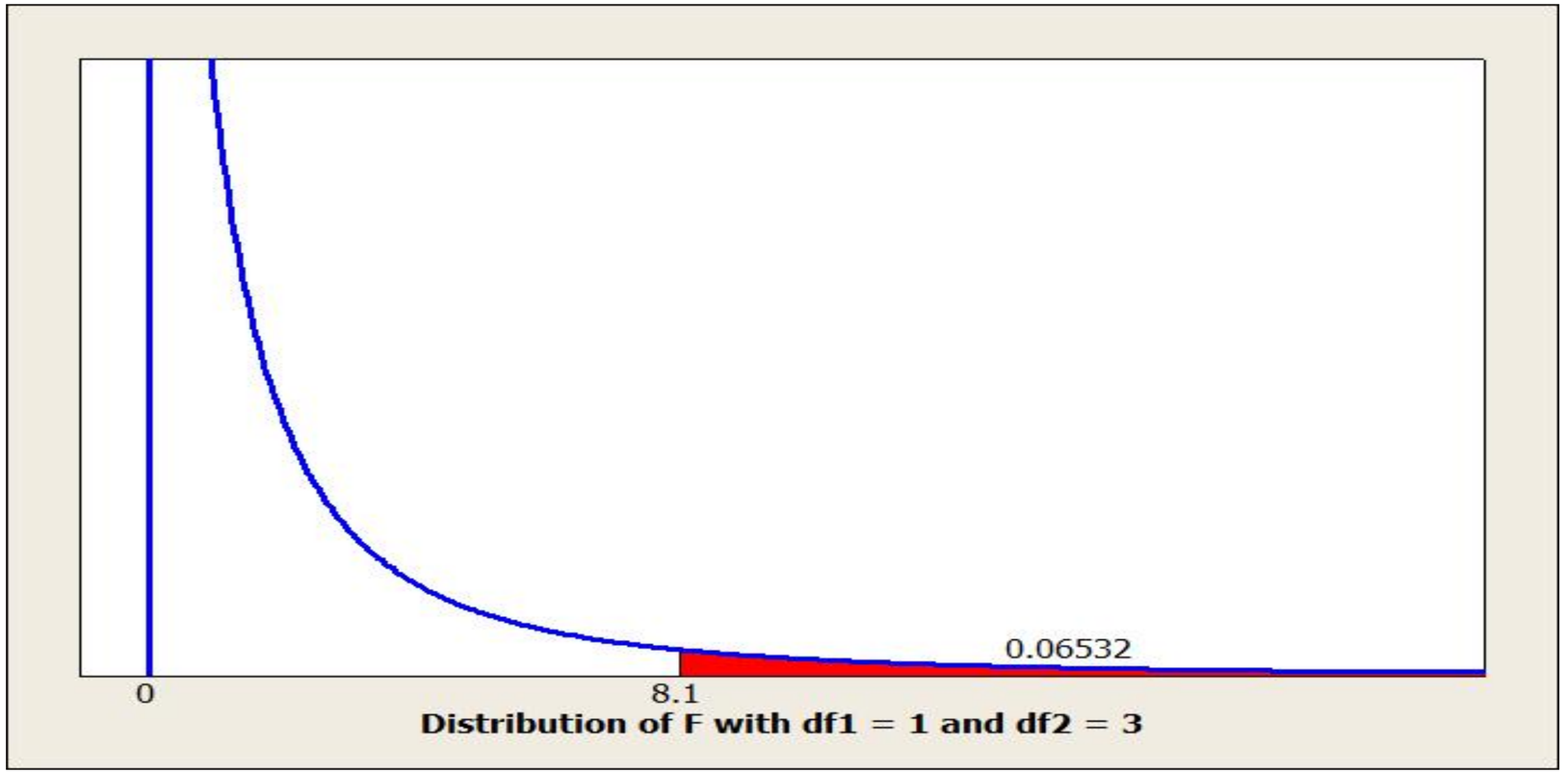
$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_j} Distance^2(x_{ij}, m_i)$$
$$= (0 - 2)^2 + (2 - 2)^2 + (4 - 2)^2 + (6 - 8)^2 + (10 - 8)^2 = 16$$

And

$$F = \frac{MSB}{MSE} = \frac{SSB/k - 1}{SSE/N - k} = \frac{43.2/1}{16/3} = \frac{43.2}{5.33} = 8.1$$

Pseudo-*F* Statistic Example (cont.)

Distribution of the F statistic shows that p-value of 0.06532 does not indicate strong evidence of clusters:



Cluster Validation

- As with any other data mining modeling technique, cluster analysis should be subject to cross-validation to ensure the clusters are real
- A simple graphical and statistical approach with the goal of confirming the clusters found in the test data match those in the training data is:

1. Apply cluster analysis to training data
2. Apply cluster analysis to test data
3. Use graphics and statistics to confirm the clusters in the training data match those in the test data

The content of the slides are prepared from different textbooks.

References:

- Data Mining and Predictive Analytics, By Daniel T. Larose. Copyright 2015 John Wiley & Sons, Inc.
- Predictive Analytics for Dummies, By Anasse Bari, Mohamed Chaouchi, & Tommy Jung, Copyright 2016, John Wiley & Sons, Inc.
- Introduction to Data Mining with Case Studies, By G.K. Gupta. Copyright 2014 by PHI Learning Private Limited.


Thank you..