Clustering Metrics

Week 07 - Day 04

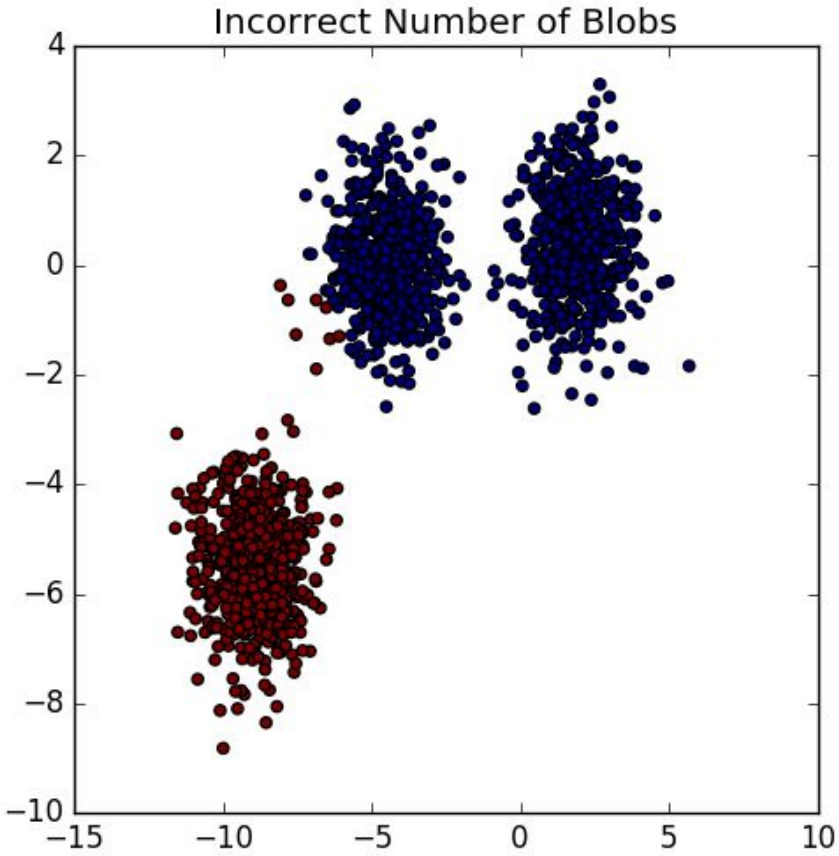
K-Means

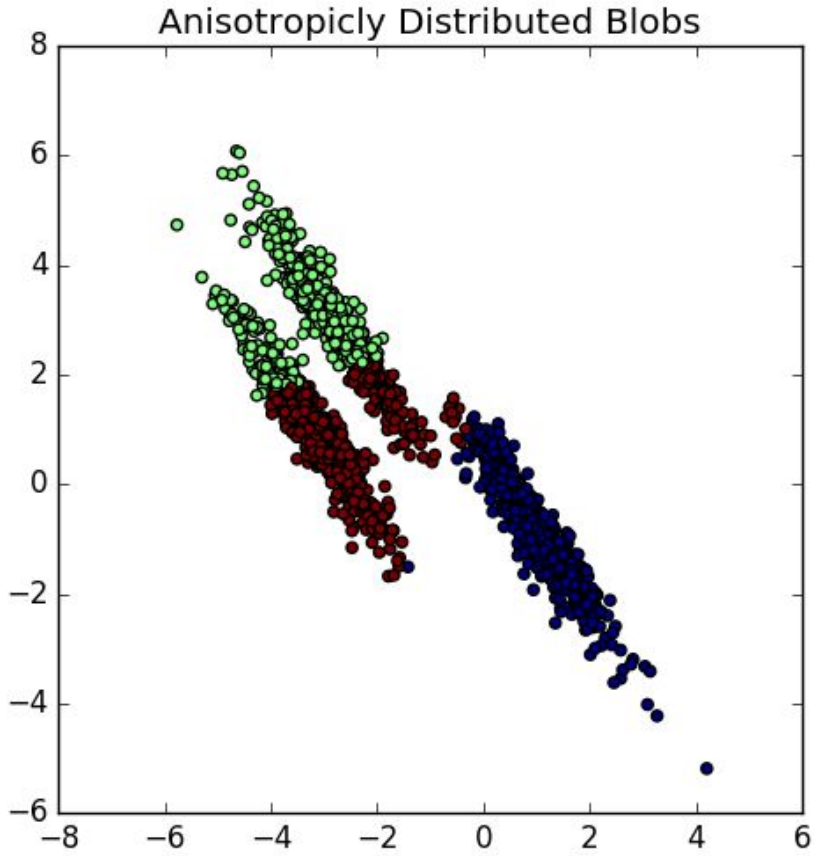
1. Based on distances
2. Uses centroids
3. Iterations
4. K is an input

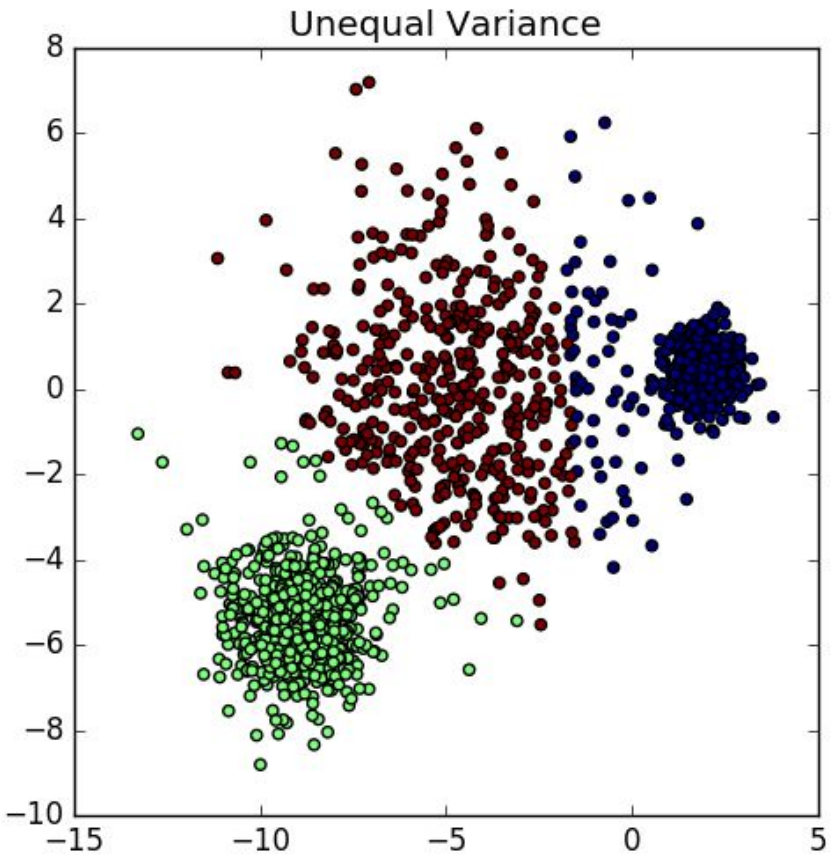
Problems

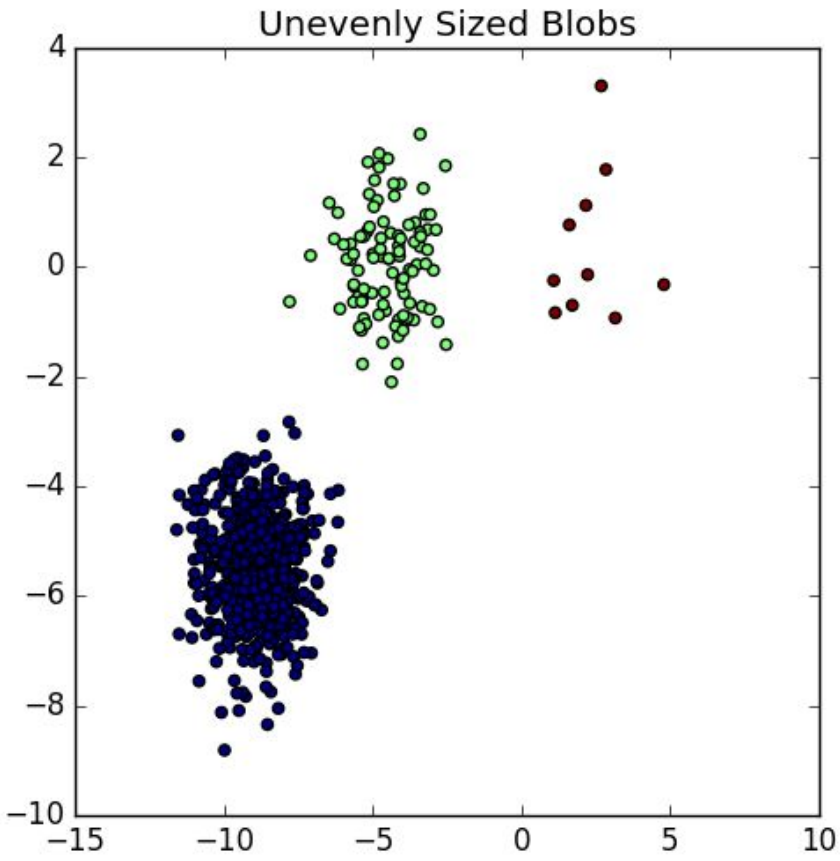
1. K is an input
2. Outliers
3. Irregular shapes

No ground truth


Incorrect Number of Blobs
4
2
0
−2
−4
−6
−8
−10
−15
−10
−5
0
5
10


Anisotropicly Distributed Blobs
8
6
4
2
0
−2
−4
−6
−8
−6
−4
−2
0
2
4
6


Unequal Variance
8
6
4
2
0
−2
−4
−6
−8
−10
−15
−10
−5
0
5


Unevenly Sized Blobs
4
2
0
−2
−4
−6
−8
−10
−15
−10
−5
0
5
10

Visual inspection is crucial!

Clustering Metrics

Inertia, Silhouette

"Many indices (more than 30) has been published in the literature for finding the right number of clusters in a dataset.”

Inertia

Average squared distance between each point and its centroid

Similar to MSE

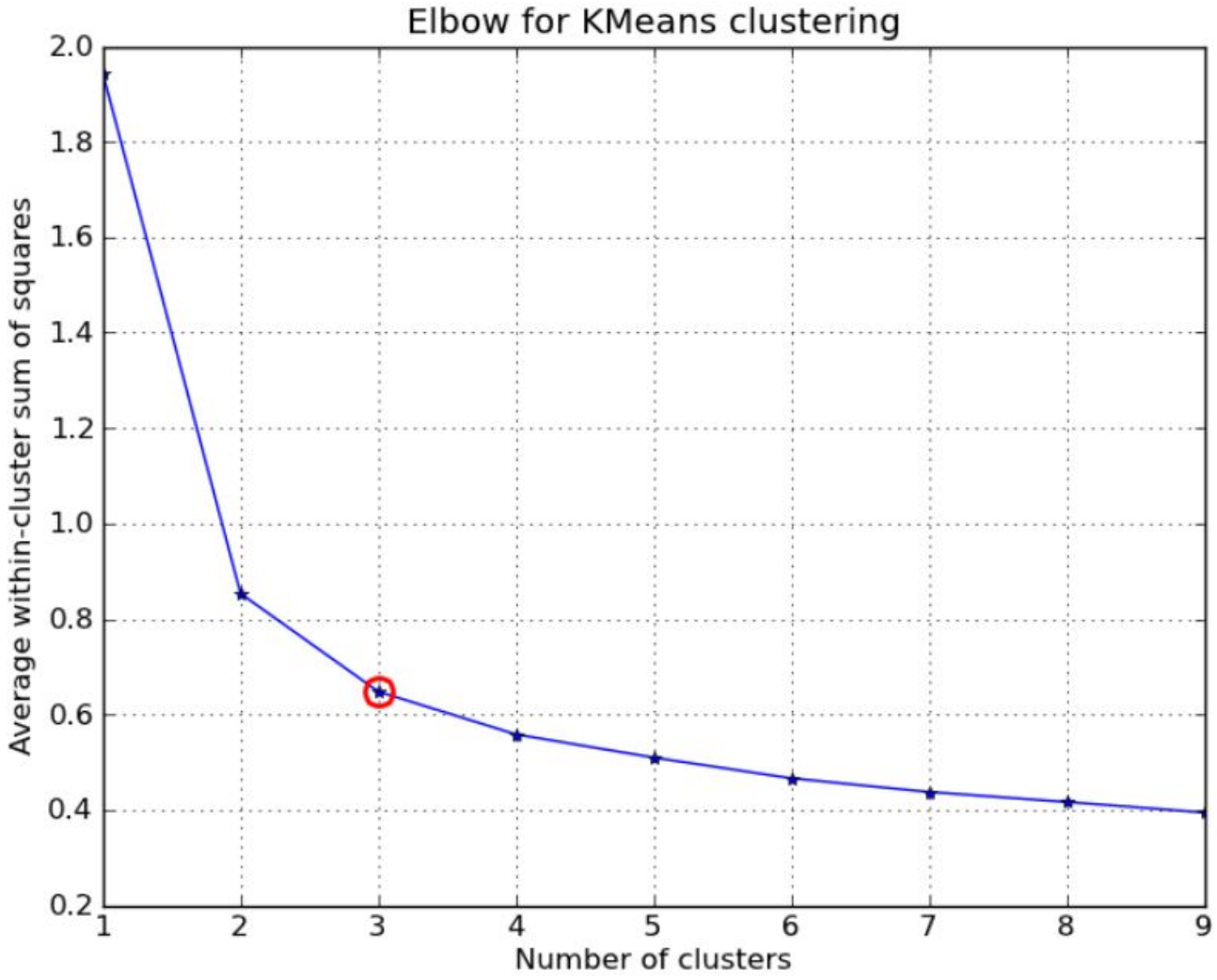
(mean squared error)

Perfect inertia

=

One cluster for each point

Inertia + elbow


Elbow for KMeans clustering
Average within-cluster sum of squares
Number of clusters
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
1
2
3
4
5
6
7
8
9

Elbow

=

Often good-enough

Silhouette Score/Coefficient

Silhouette = Cohesion + Separation

Can be calculated for every point!

Cohesion

=

“Intra-cluster distance”

Cohesion

=

Distance from point to centroid

Separation

=

“Inter-cluster distance”

Separation

=

Distance from point to closer cluster

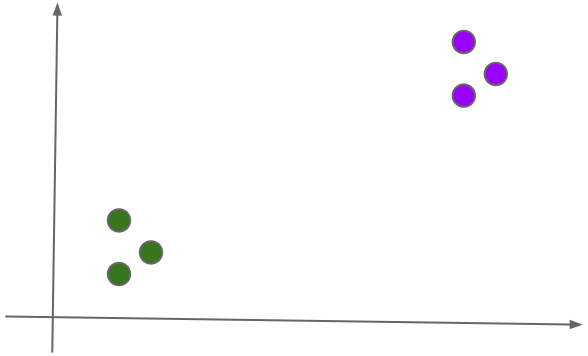
$$\frac{\text{separation - cohesion}}{\max(\text{separation},\text{cohesion})}$$

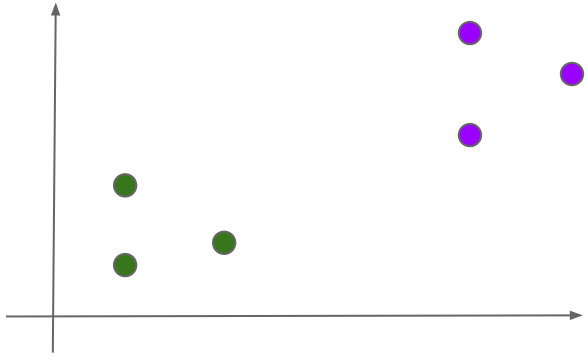
Silhouette Coefficient

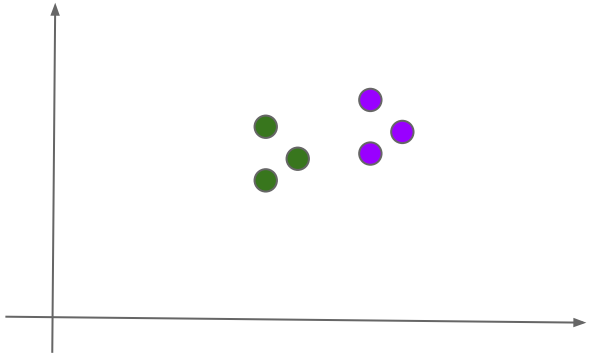
=

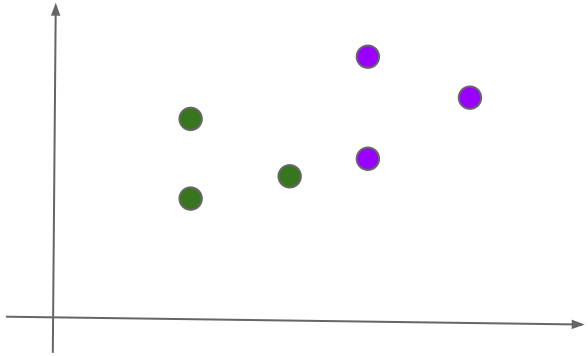
Average Silhouette

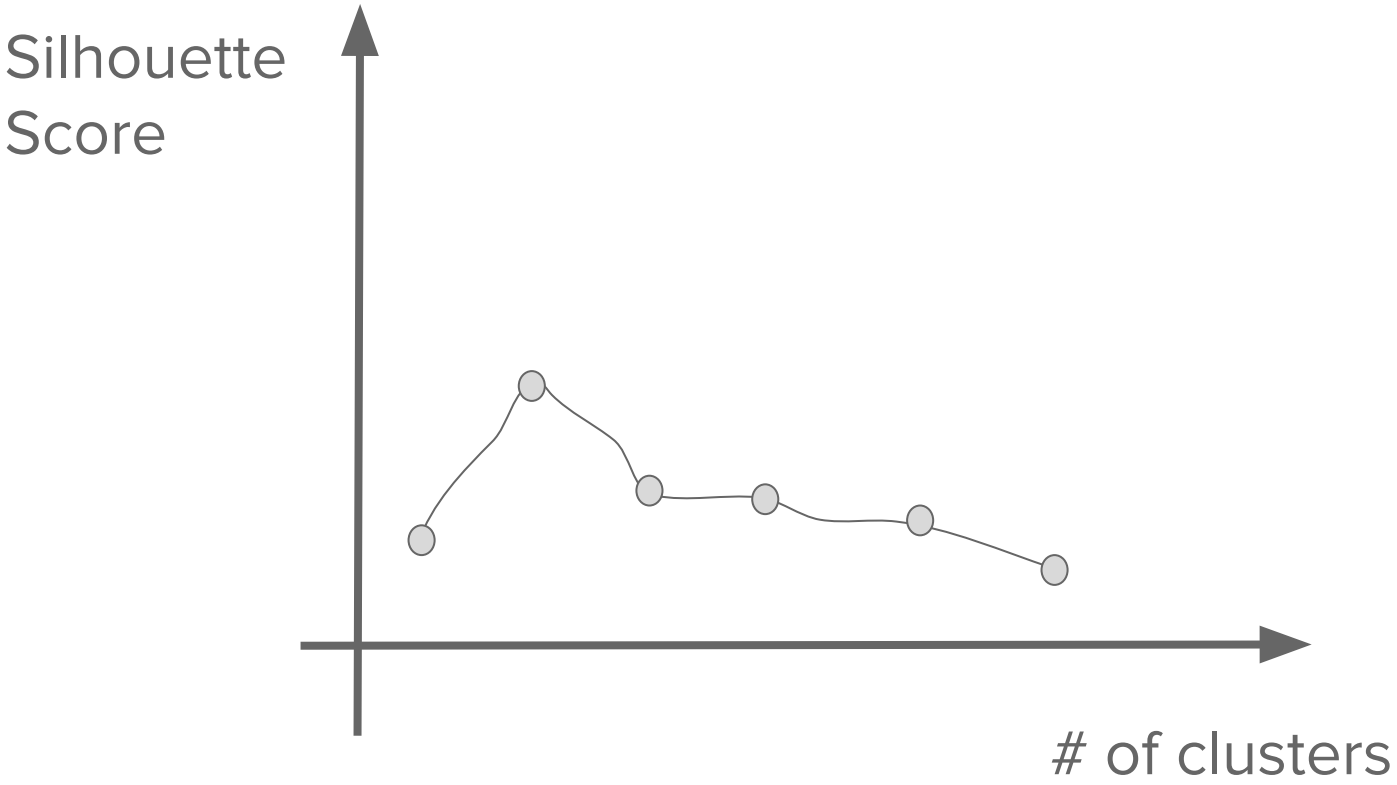
(of each point)

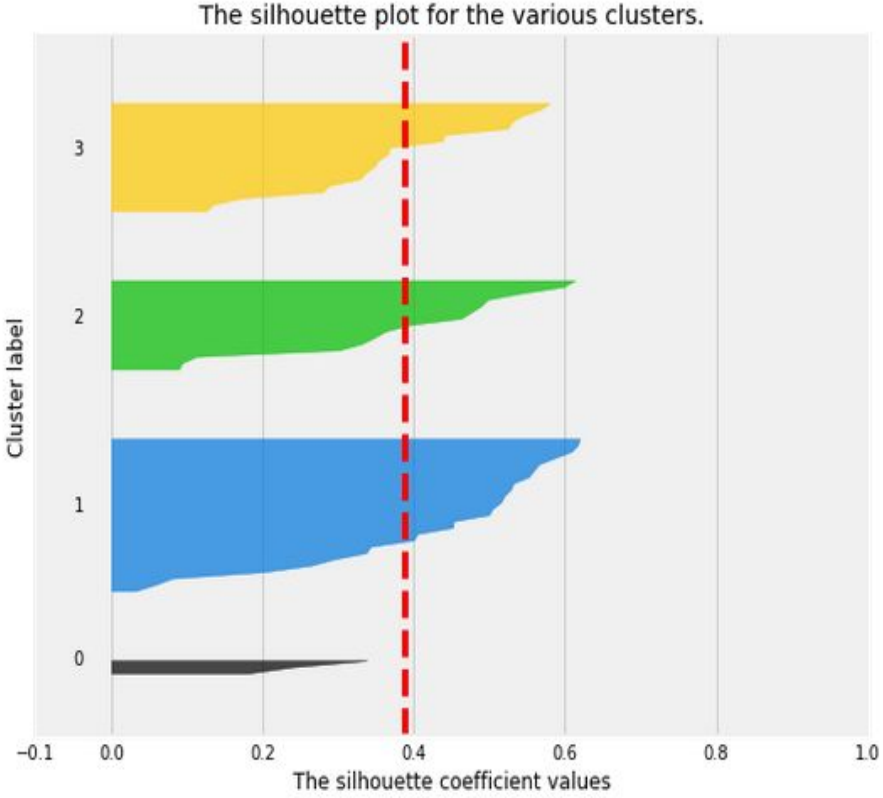


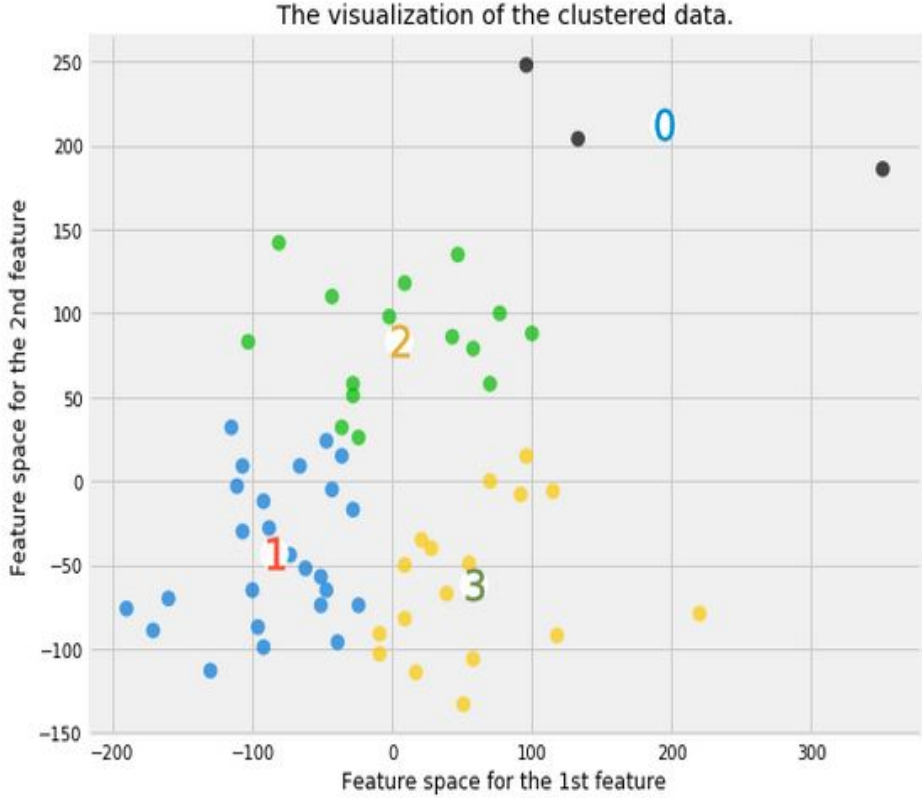







Silhouette
Score
# of clusters


The silhouette plot for the various clusters.
Cluster label
3
2
1
0
−0.1
0.0
0.2
0.4
0.6
0.8
1.0
The silhouette coefficient values


The visualization of the clustered data.
Feature space for the 2nd feature
250
200
150
100
50
0
−50
−100
−150
−200
−100
0
100
200
300
Feature space for the 1st feature
0
1
2
3

Silhouette Coefficient

is good but not perfect!

Metrics when you have the labels

Unusual...why?

Labels = Classification!

Too easy :)

Completeness

Homogeneity

V-Measure Score

Mutual Information Score

Completeness

“Indicates that all members of a given class are assigned to the same cluster.”

[0,1]

Homogeneity

“Indicates each cluster contains only members of a single class.”

[0,1]

V-Measure

$$V = \frac{2 \cdot \text{homogeneity} \cdot \text{completeness}}{\text{homogeneity} + \text{completeness}}$$

[0,1]

Similar to what?

F1 score!

Mutual Information Score

$$MI(i,j) = \sum_{a,b} P(a_i, b_j) \cdot log \left(\frac{P(a_i, b_j)}{P(a_i) \cdot P(b_j)} \right)$$