



Pattern Recognition

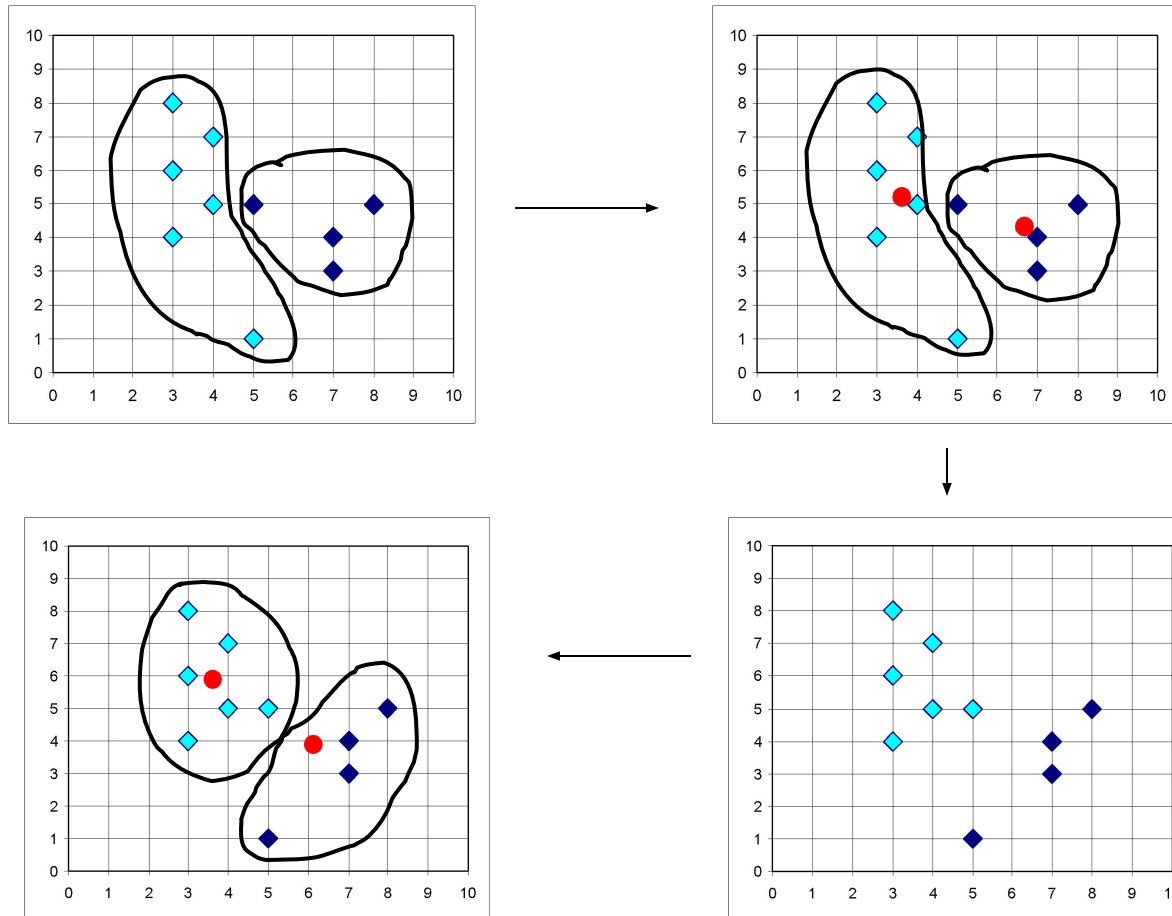
- S. S. Samant

K-means Clustering

- Given k , the *k-means* algorithm consists of four steps:
 - Select initial centroids at random.
 - Assign each object to the cluster with the nearest centroid.
 - Compute each centroid as the mean of the objects assigned to it.
 - Repeat previous 2 steps until there is no change.

K-means Clustering

- Example



K-means example (K=2)

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

K-means example (K=2)

Initialization: Randomly we choose following two centroids ($k=2$) for two clusters.

In this case the 2 centroid are: $m_1=(1.0,1.0)$ and $m_2=(5.0,7.0)$.

Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

	Individual	Mean Vector
Group 1	1	(1.0, 1.0)
Group 2	4	(5.0, 7.0)

K-means example (K=2)

- Thus, we obtain two clusters containing:

{1,2,3} and {4,5,6,7}.

- Their new centroids are:

$$m_1 = \left(\frac{1}{3}(1.0 + 1.5 + 3.0), \frac{1}{3}(1.0 + 2.0 + 4.0) \right) = (1.83, 2.33)$$

$$m_2 = \left(\frac{1}{4}(5.0 + 3.5 + 4.5 + 3.5), \frac{1}{4}(7.0 + 5.0 + 5.0 + 4.5) \right)$$

$$= (4.12, 5.38)$$

Individual	Centroid 1	Centroid 2
1	0	7.21
2 (1.5, 2.0)	1.12	6.10
3	3.61	3.61
4	7.21	0
5	4.72	2.5
6	5.31	2.06
7	4.30	2.92

$$d(m_1, 2) = \sqrt{|1.0 - 1.5|^2 + |1.0 - 2.0|^2} = 1.12$$

$$d(m_2, 2) = \sqrt{|5.0 - 1.5|^2 + |7.0 - 2.0|^2} = 6.10$$

K-means example (K=2)

- Now using these centroids we compute the Euclidean distance of each object, as shown in table.
- Therefore, the new clusters are:
 $\{1,2\}$ and $\{3,4,5,6,7\}$
- Next centroids are: $m1=(1.25,1.5)$ and $m2 = (3.9,5.1)$

Individual	Centroid 1	Centroid 2
1	1.57	5.38
2	0.47	4.28
3	2.04	1.78
4	5.64	1.84
5	3.15	0.73
6	3.78	0.54
7	2.74	1.08

K-means example (K=2)

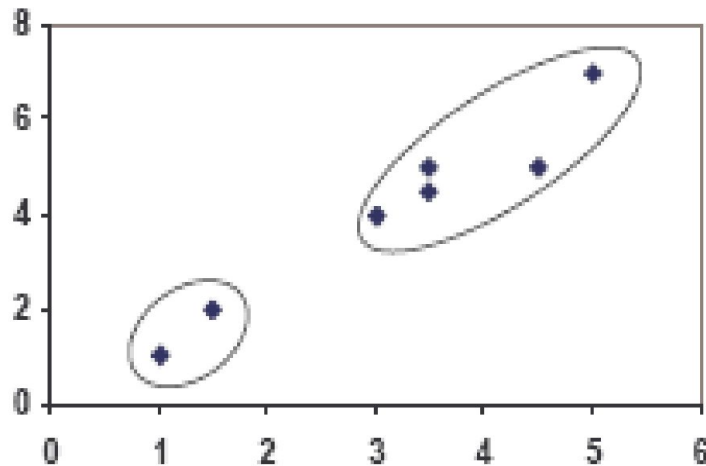
The clusters obtained are:

{1,2} and {3,4,5,6,7}

- Therefore, there is no change in the cluster.
- Thus, the algorithm comes to a halt
 - Final result consist of 2 clusters {1,2} and {3,4,5,6,7}

Individual	Centroid 1	Centroid 2
1	0.56	5.02
2	0.56	3.92
3	3.05	1.42
4	6.66	2.20
5	4.16	0.41
6	4.78	0.61
7	3.75	0.72

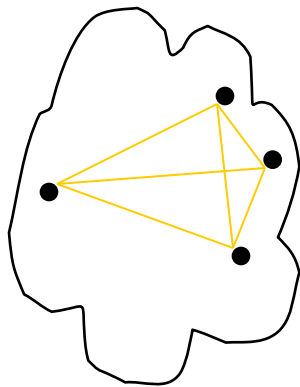
K-means example (K=2)



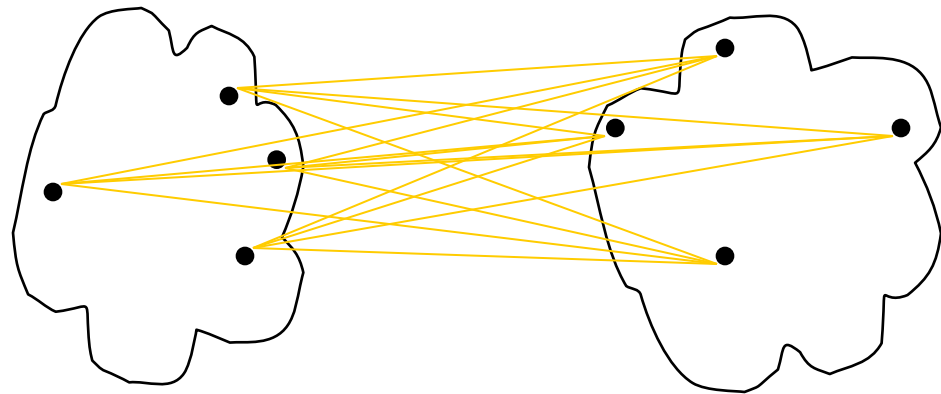
Individual	Variable 1	Variable 2
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Measuring Cluster Quality: Cohesion and Separation

- A proximity graph based approach can be used for cohesion and separation.
 - Cluster cohesion is the sum of the weight of all links within a cluster.
 - Cluster separation is the sum of the weights between nodes in the cluster and nodes outside the cluster.



cohesion



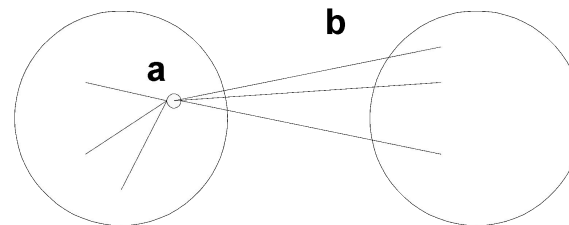
separation

Silhouette Coefficient

- Silhouette Coefficient combine ideas of both cohesion and separation, but for individual points, as well as clusters and clustering
- For an individual point i
 - Calculate a = average distance of i to the points in its cluster
 - Calculate b = min (average distance of i to points in another cluster)
 - The silhouette coefficient for a point is then given by

$$s = (b - a) / \max(a, b)$$

- Varies between -1 and 1.
- The closer to 1 the better.



- Can calculate the Average Silhouette coefficient for a cluster or a clustering

Example

Using the distance matrix in the following table, compute the **silhouette coefficient** for each point

$$s = (b - a) / \max(a, b)$$

	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

Cluster 1: {P1, P2}

Cluster 2: {P3, P4}

Example

Using the distance matrix in the following table, compute the **silhouette coefficient** for each point

	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

Cluster 1: {P1, P2}

Cluster 2: {P3, P4}

Let a indicate the average distance of a point to other points in its cluster.
Let b indicate the minimum of the average distance of a point to points in another cluster.

Point P1: $SC = 1 - a/b = 1 - 0.1/((0.65+0.55)/2) = 5/6 = 0.833$

Point P2: $SC = 1 - a/b = 1 - 0.1/((0.7+0.6)/2) = 0.846$

Point P3: $SC = 1 - a/b = 1 - 0.3/((0.65+0.7)/2) = 0.556$

Point P4: $SC = 1 - a/b = 1 - 0.3/((0.55+0.6)/2) = 0.478$

Example

Using the distance matrix in the following table, compute the **silhouette coefficient** for each point **and each of the two clusters**

	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

Cluster 1: {P1, P2}

Cluster 2: {P3, P4}

Let a indicate the average distance of a point to other points in its cluster.
Let b indicate the minimum of the average distance of a point to points in another cluster.

Point P1: $SC = 1 - a/b = 1 - 0.1/((0.65+0.55)/2) = 5/6 = 0.833$

Point P2: $SC = 1 - a/b = 1 - 0.1/((0.7+0.6)/2) = 0.846$

Point P3: $SC = 1 - a/b = 1 - 0.3/((0.65+0.7)/2) = 0.556$

Point P4: $SC = 1 - a/b = 1 - 0.3/((0.55+0.6)/2) = 0.478$

Example

Using the distance matrix in the following table, compute the **silhouette coefficient** for each point **and each of the two clusters**

	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

Cluster 1: {P1, P2}

Cluster 2: {P3, P4}

Let a indicate the average distance of a point to other points in its cluster.
Let b indicate the minimum of the average distance of a point to points in another cluster.

Point P1: $SC = 1 - a/b = 1 - 0.1/((0.65+0.55)/2) = 5/6 = 0.833$

Point P2: $SC = 1 - a/b = 1 - 0.1/((0.7+0.6)/2) = 0.846$

Point P3: $SC = 1 - a/b = 1 - 0.3/((0.65+0.7)/2) = 0.556$

Point P4: $SC = 1 - a/b = 1 - 0.3/((0.55+0.6)/2) = 0.478$

Cluster 1 Average SC = $(0.833+0.846)/2 = 0.84$

Cluster 2 Average SC = $(0.556+0.478)/2 = 0.52$

Example

Using the distance matrix in the following table, compute the **silhouette coefficient** for each point, each of the two clusters, and **overall clustering**

	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

Cluster 1: {P1, P2}

Cluster 2: {P3, P4}

Let a indicate the average distance of a point to other points in its cluster.
Let b indicate the minimum of the average distance of a point to points in another cluster.

Point P1: $SC = 1 - a/b = 1 - 0.1/((0.65+0.55)/2) = 5/6 = 0.833$

Point P2: $SC = 1 - a/b = 1 - 0.1/((0.7+0.6)/2) = 0.846$

Point P3: $SC = 1 - a/b = 1 - 0.3/((0.65+0.7)/2) = 0.556$

Point P4: $SC = 1 - a/b = 1 - 0.3/((0.55+0.6)/2) = 0.478$

Cluster 1 Average SC = $(0.833+0.846)/2 = 0.84$

Cluster 2 Average SC = $(0.556+0.478)/2 = 0.52$

Example

Using the distance matrix in the following table, compute the **silhouette coefficient** for each point, each of the two clusters, and **overall clustering**

	P1	P2	P3	P4
P1	0	0.10	0.65	0.55
P2	0.10	0	0.70	0.60
P3	0.65	0.70	0	0.30
P4	0.55	0.60	0.30	0

Cluster 1: {P1, P2}

Cluster 2: {P3, P4}

Let a indicate the average distance of a point to other points in its cluster.
Let b indicate the minimum of the average distance of a point to points in another cluster.

Point P1: $SC = 1 - a/b = 1 - 0.1/((0.65+0.55)/2) = 5/6 = 0.833$

Point P2: $SC = 1 - a/b = 1 - 0.1/((0.7+0.6)/2) = 0.846$

Point P3: $SC = 1 - a/b = 1 - 0.3/((0.65+0.7)/2) = 0.556$

Point P4: $SC = 1 - a/b = 1 - 0.3/((0.55+0.6)/2) = 0.478$

Cluster 1 Average SC = $(0.833+0.846)/2 = 0.84$

Cluster 2 Average SC = $(0.556+0.478)/2 = 0.52$

Overall Average SC = $(0.840+0.517)/2 = 0.68$

K-means with scikit-learn

```
from sklearn import datasets
import matplotlib.pyplot as plt
import pandas as pd
from sklearn.cluster import KMeans
from sklearn import metrics
```

```
iris = datasets.load_iris()
X = iris.data
y = iris.target
plt.scatter(X[:,0], X[:,2], c=y, cmap='spring', s=100)
plt.title('Actual',fontsize=25, fontweight='bold')
plt.xlabel('Sepal Length',fontsize=20)
plt.ylabel('Petal Length',fontsize=20)
plt.figure()
```

```
cls = KMeans(n_clusters = 3, random_state=42)
cls.fit(X)
print 'Final centroids:'
print cls.cluster_centers_
```

```
km_labels = cls.labels_
print metrics.silhouette_score(X, km_labels)
plt.scatter(X[:,0], X[:,2],c=km_labels, cmap='spring', s=100)
plt.xlabel('Sepal Length',fontsize=20)
plt.ylabel('Petal Length',fontsize=20)
plt.title('Predicted clusters',fontsize=25, fontweight='bold')
plt.show()
```



Iris Versicolor



Iris Setosa



Iris Virginica

Classification-oriented measures

Cluster	Enter-tainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27		
2	4	7	280	29	39	2		
3	1	1	1	7	4	671		
4	10	162	3	119	73	2		
5	331	22	5	70	13	23		
6	5	358	12	212	48	13		
Total	354	555	341	943	273	738		

Calculate Entropy, purity, precision, recall, F-score of each of the clusters above.

Classification-oriented measures

Cluster	Enter- tainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

Classification-oriented measures

Cluster	Enter- tainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

What is the precision and recall of Cluster-1 *wrt* Metro class

Classification-oriented measures

Cluster	Enter-tainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

What is the precision and recall of Cluster-1 *wrt* Metro class

Precision = 0.74

Recall = 0.53



Thank You!