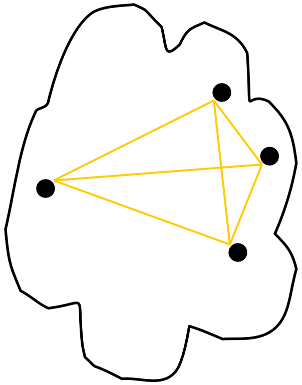
Pattern Recognition

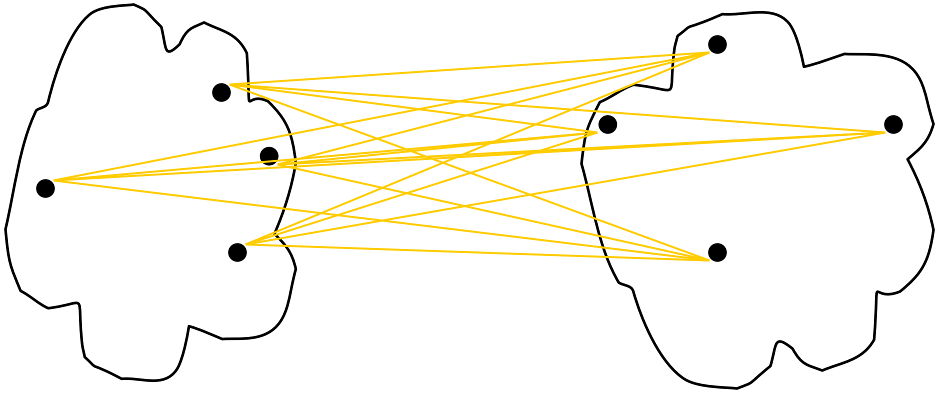
- S. S. Samant

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



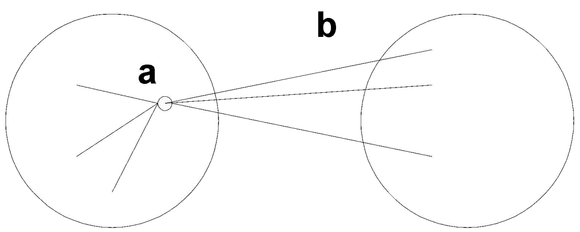
separation

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 P~~2~~3: SC = 1- a/b = 1 - 0.3/((0.65+0.7)/2) = 0.556
Point P~~2~~4: 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 P~~2~~3: SC = 1- a/b = 1 - 0.3/((0.65+0.7)/2) = 0.556
Point P~~2~~4: 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 P~~2~~3: SC = 1- a/b = 1 - 0.3/((0.65+0.7)/2) = 0.556
Point P~~2~~4: 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()
```

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

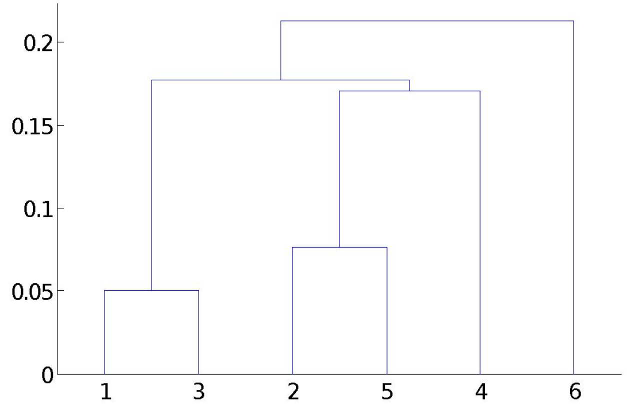
Precision = 0.74
Recall = 0.53

Agglomerative Hierarchical Clustering

- Starts with each point being a cluster, and at each step, merge the *closest* pair of clusters
- Displayed graphically using a **dendrogram – a tree like structure**

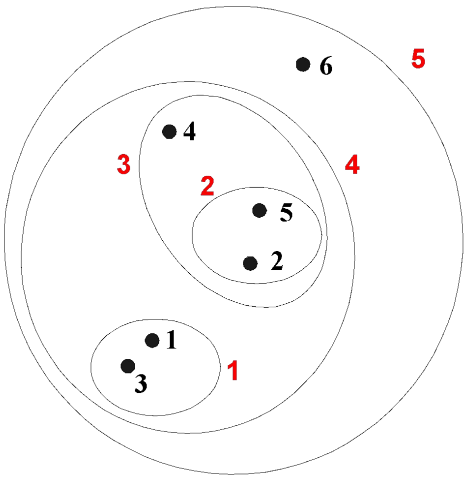
Agglomerative Hierarchical Clustering

- Starts with each point being a cluster, and at each step, merge the *closest* pair of clusters
- Displayed graphically using a **dendrogram – a tree like structure** (*dendro* "tree", *gramma* "drawing")



Agglomerative Hierarchical Clustering

- Starts with each point being a cluster, and at each step, merge the *closest* pair of clusters
- Can also be displayed graphically using a **nested cluster diagram**



Agglomerative Hierarchical Clustering

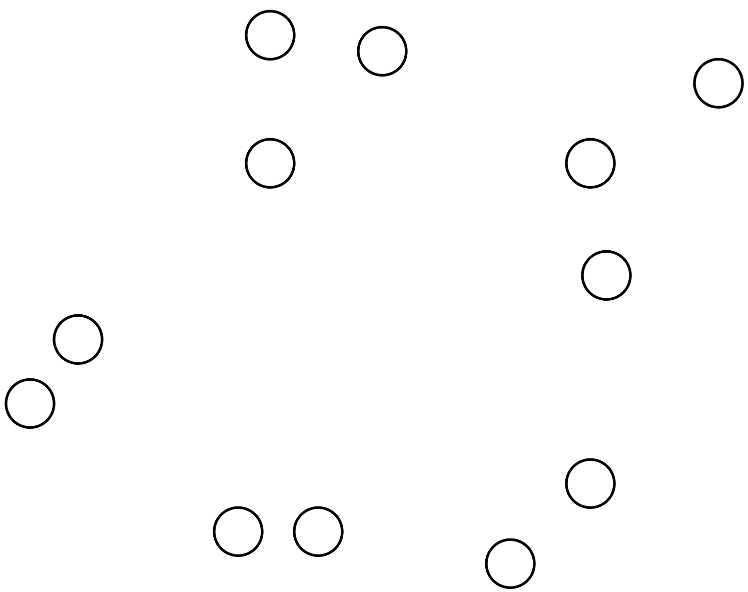
Basic algorithm

1. Compute the proximity matrix
2. Let each data point be a cluster
3. **Repeat**
4. Merge the two closest clusters
5. Update the proximity matrix
6. **Until** only a single cluster remains

- Key operation is the computation of the proximity of two clusters
  - Different approaches to defining the distance between clusters distinguish the different algorithms

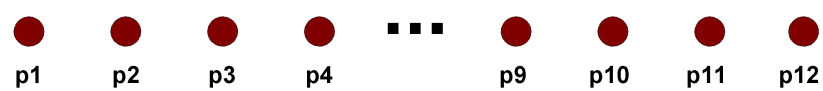
Starting Situation

- Start with clusters of individual points and a proximity matrix



	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

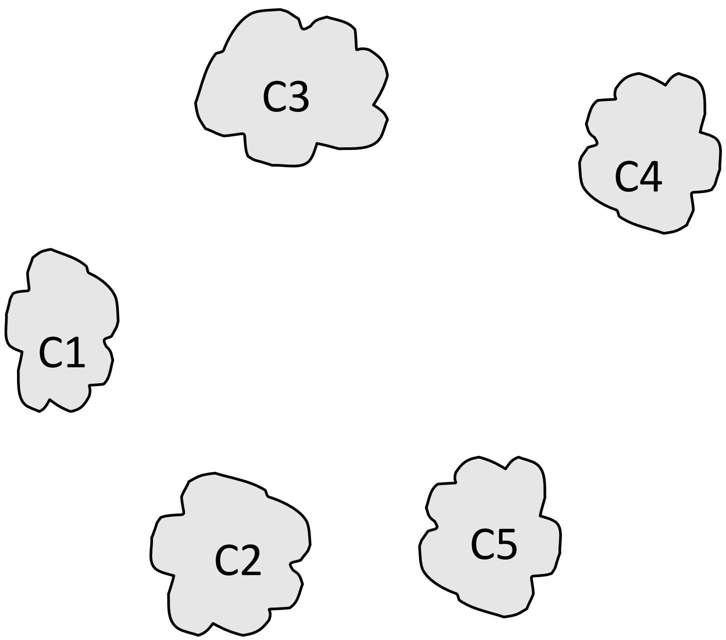


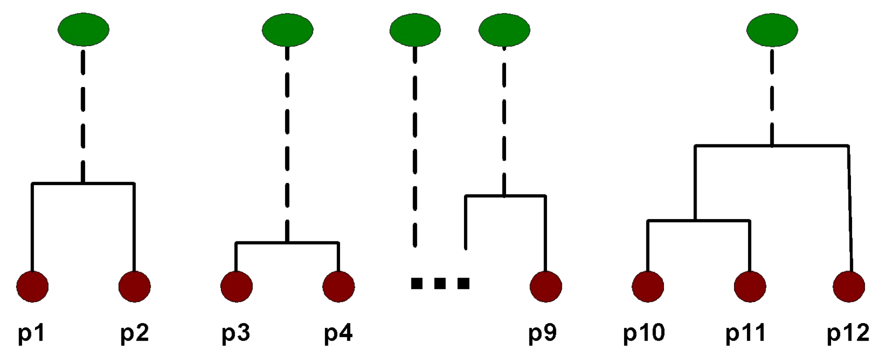
Intermediate Situation

- After some merging steps, we have some clusters

	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

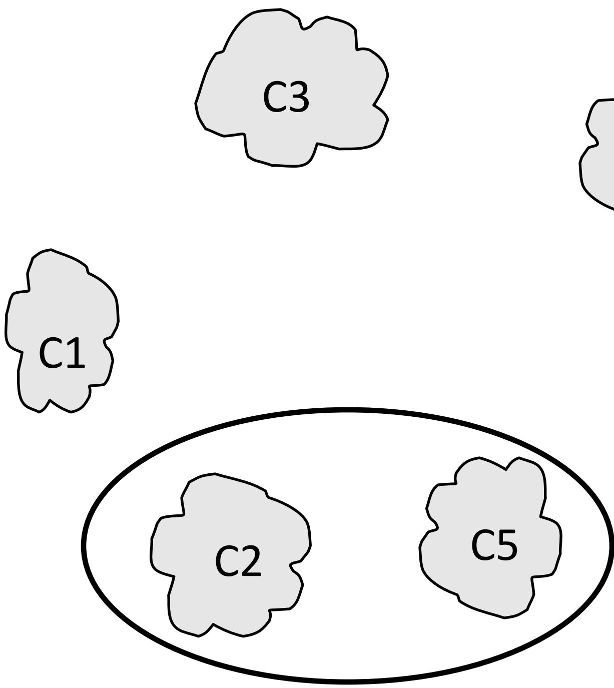
Proximity Matrix

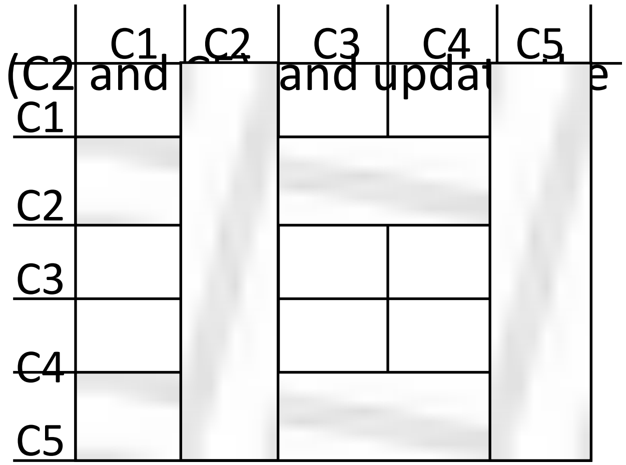




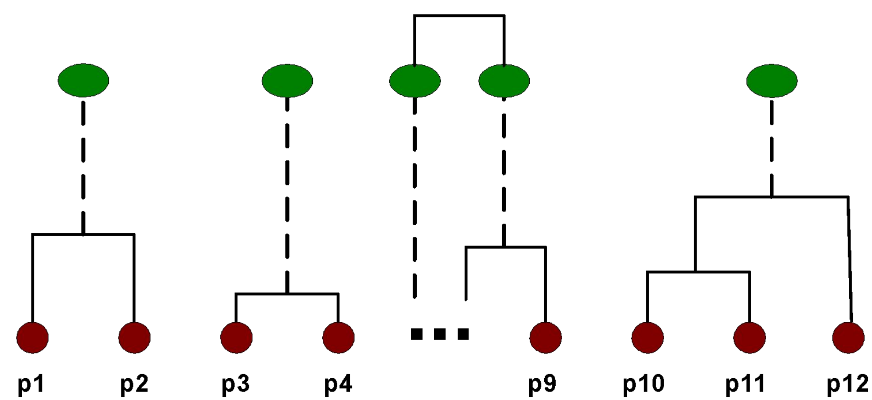
Intermediate Situation

- We want to merge the two closest clusters (C2 and C5) and update the proximity matrix.



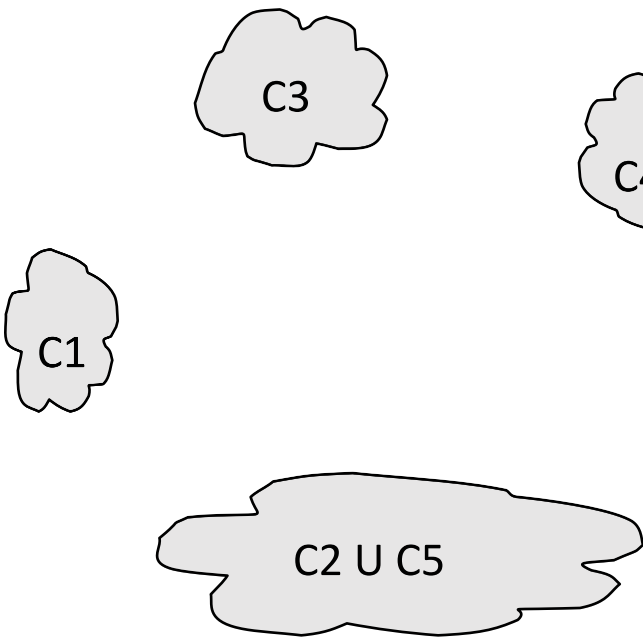


Proximity Matrix



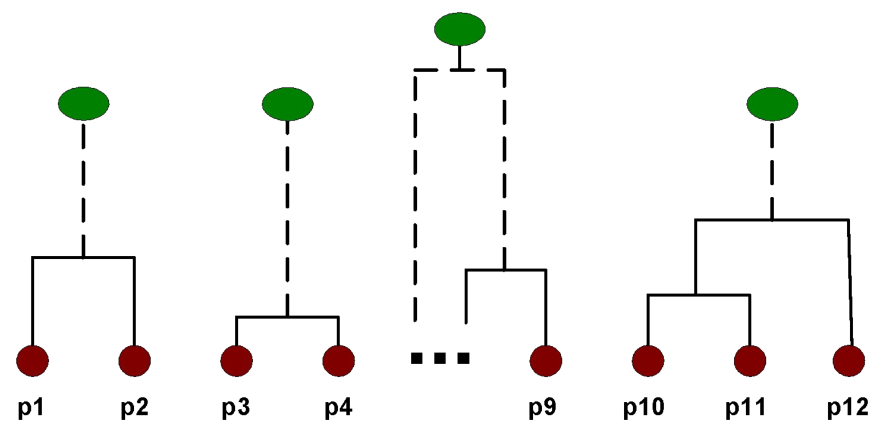
After Merging

- The question is “How do we update the proximity matrix?”

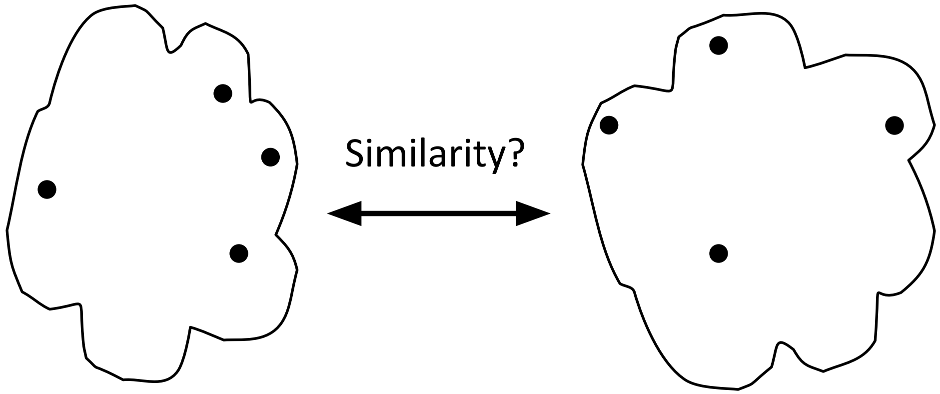


	C1	C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



How to Define Inter-Cluster Similarity

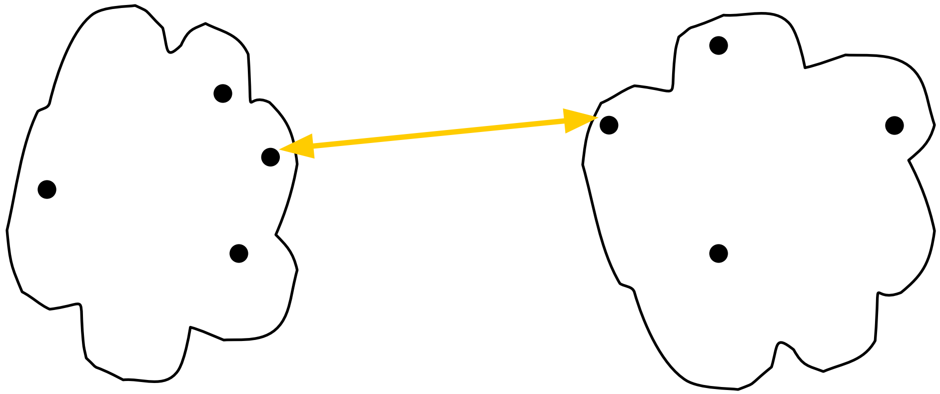


	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids

How to Define Inter-Cluster Similarity

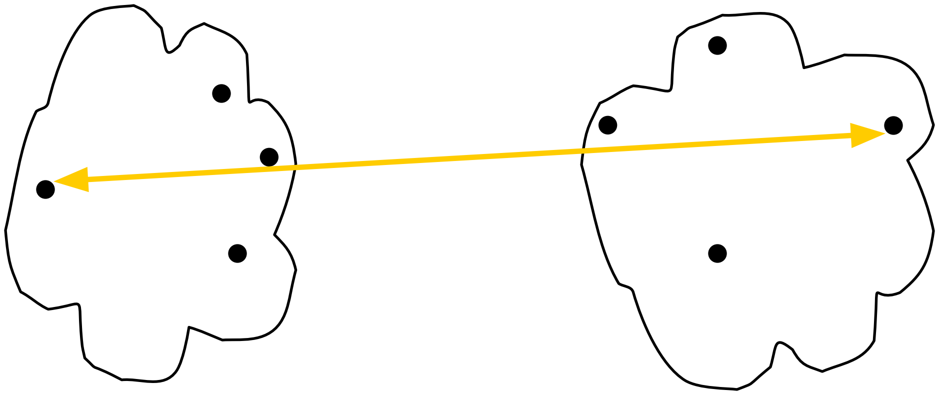


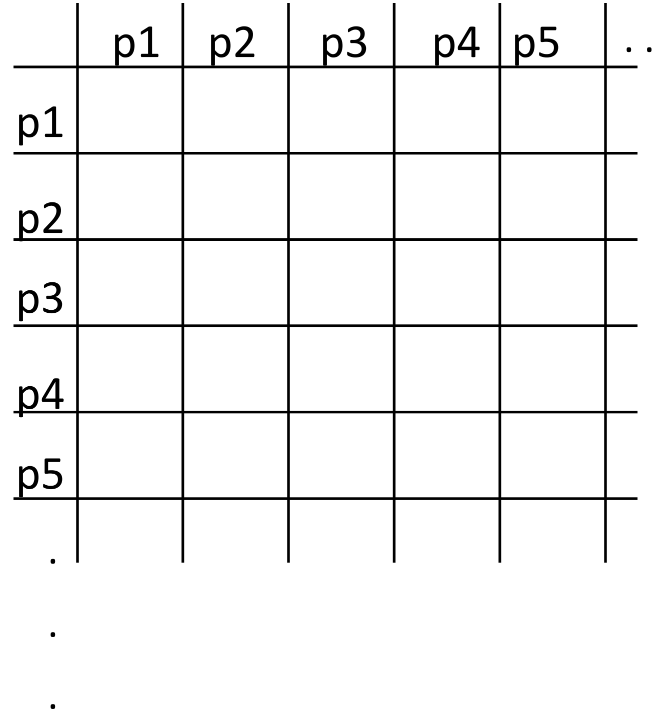
	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids

How to Define Inter-Cluster Similarity

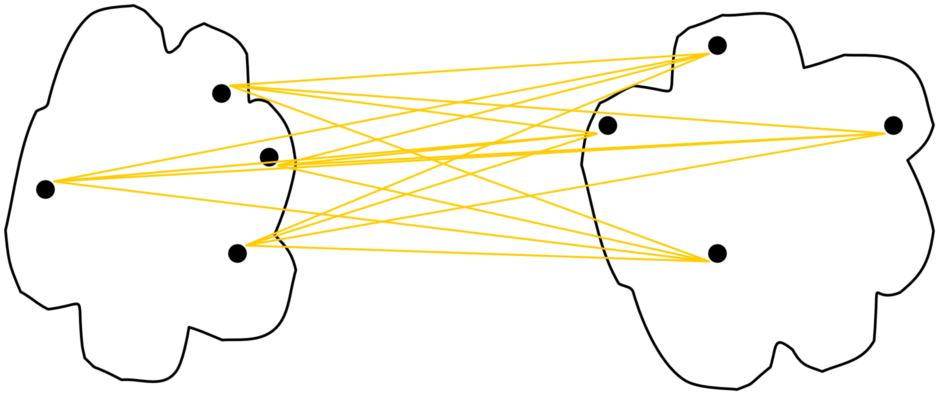




Proximity Matrix

- MIN
- MAX
- Group Average
- Distance Between Centroids

How to Define Inter-Cluster Similarity

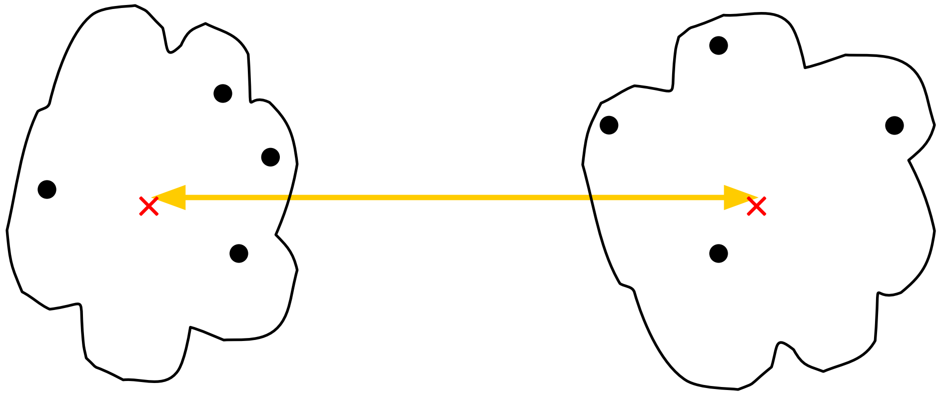


- MIN
- MAX
- Group Average
- Distance Between Centroids

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

How to Define Inter-Cluster Similarity



- MIN
- MAX
- Group Average
- Distance Between Centroids

	p1	p2	p3	p4	p5	. . .
p1						
p2						
p3						
p4						
p5						
.						
.						
.						

Proximity Matrix

Cluster Similarity: MIN or Single Link

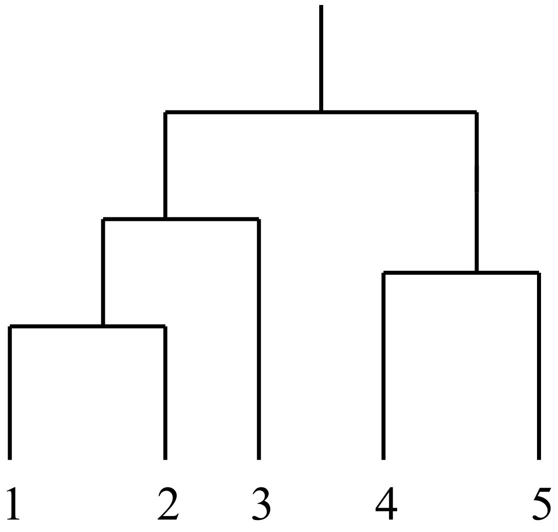
- Similarity of two clusters is based on the two closest points in the different clusters

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

Cluster Similarity: MIN or Single Link

- Similarity of two clusters is based on the two closest points in the different clusters

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Cluster Similarity: MAX or Complete Linkage

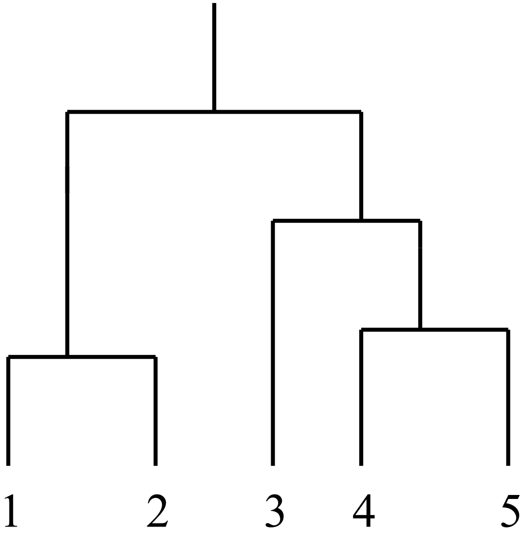
- Similarity of two clusters is based on the two farthest points in the different clusters

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00

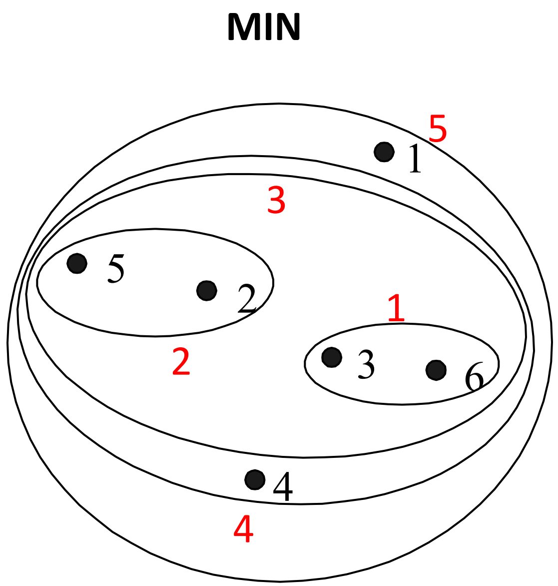
Cluster Similarity: MAX or Complete Linkage

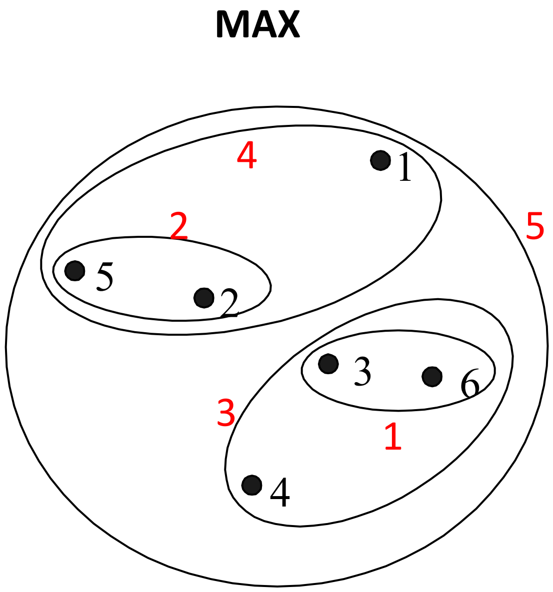
- Similarity of two clusters is based on the two farthest points in the different clusters

	I1	I2	I3	I4	I5
I1	1.00	0.90	0.10	0.65	0.20
I2	0.90	1.00	0.70	0.60	0.50
I3	0.10	0.70	1.00	0.40	0.30
I4	0.65	0.60	0.40	1.00	0.80
I5	0.20	0.50	0.30	0.80	1.00



Hierarchical Clustering: Comparison





Examples

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

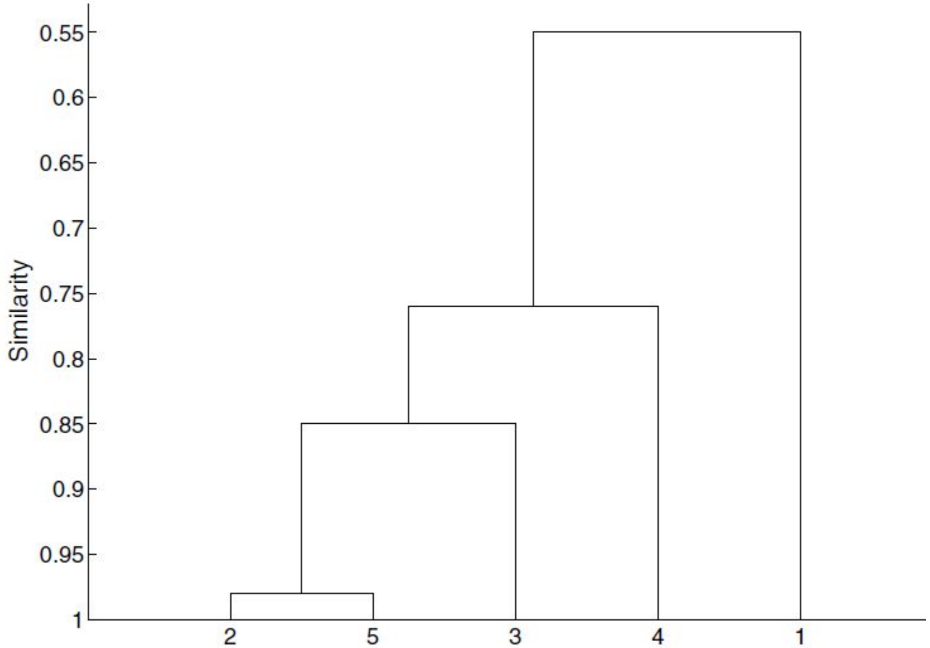
Given the data above, perform single link and complete link hierarchical clustering. Draw dendrogram of your results.

Examples

	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

Given the data above, perform single link and complete link hierarchical clustering. Draw dendrogram of your results.

Single Link:

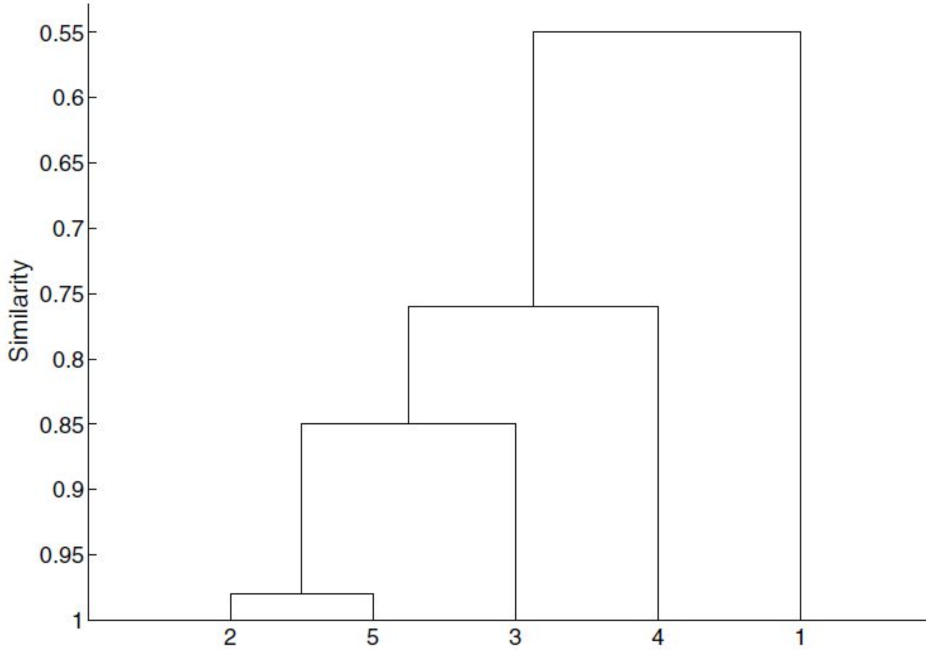


Examples

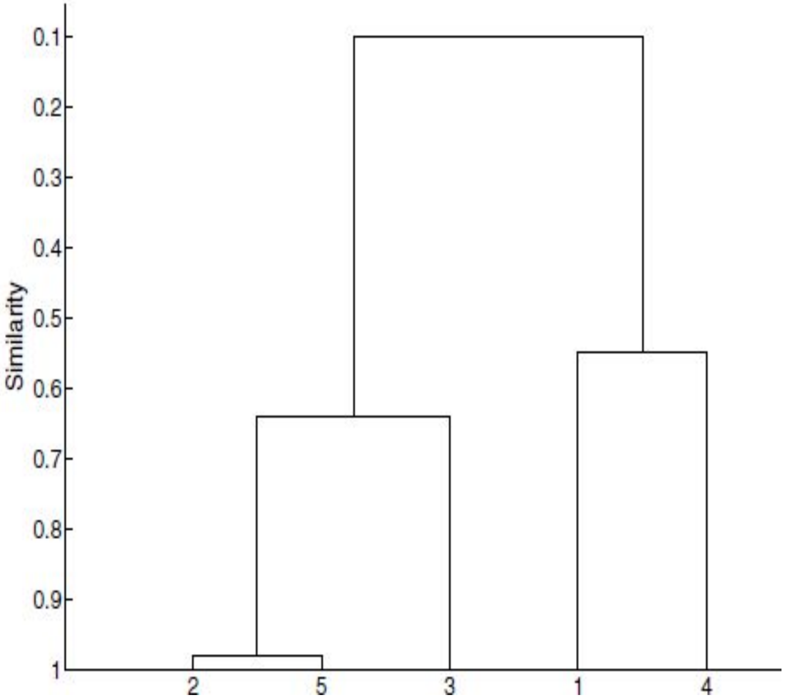
	p1	p2	p3	p4	p5
p1	1.00	0.10	0.41	0.55	0.35
p2	0.10	1.00	0.64	0.47	0.98
p3	0.41	0.64	1.00	0.44	0.85
p4	0.55	0.47	0.44	1.00	0.76
p5	0.35	0.98	0.85	0.76	1.00

Given the data above, perform single link and complete link hierarchical clustering. Draw dendrogram of your results.

Single Link:



Complete Link



Example - HAC on Iris dataset

```
from sklearn import datasets
import matplotlib.pyplot as plt
from sklearn.cluster import AgglomerativeClustering
from sklearn import metrics

iris = datasets.load_iris()
X = iris.data
y = iris.target
plt.scatter(X[:,0], X[:,1], c=y, cmap='spring', s=100)
plt.title('Actual',fontsize=20, fontweight='bold')
plt.xlabel('Sepal Length',fontsize=20)
plt.ylabel('Petal Length',fontsize=20)
plt.figure()

cls = AgglomerativeClustering(n_clusters = 3, linkage='average')
cls.fit(X)

hac_labels = cls.labels_
print metrics.silhouette_score(X, hac_labels)
plt.scatter(X[:,0], X[:,1],c=hac_labels, cmap='spring', s=100)
plt.xlabel('Sepal Length',fontsize=20)
plt.ylabel('Petal Length',fontsize=20)
plt.title('Predicted clusters',fontsize=20, fontweight='bold')
plt.show()
```

Display the Dendrogram

```
from sklearn import datasets
import matplotlib.pyplot as plt
from sklearn.cluster import AgglomerativeClustering
import numpy as np
from scipy.cluster.hierarchy import dendrogram

iris = datasets.load_iris()
X = iris.data
y = iris.target

cls = AgglomerativeClustering(n_clusters = 3, linkage='average')
cls.fit(X)

children = cls.children_
dist = np.arange(children.shape[0])
observations = np.arange(2, children.shape[0]+2)
linkage_mat = np.column_stack([children, dist,
observations]).astype(float)
dendrogram(linkage_mat, labels=cls.labels_)
plt.show()
```

External Evaluation 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 and purity of each of the clusters above.

External Evaluation Measures

Cluster	Entertainment	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

External Evaluation 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.

External Evaluation Measures

Cluster	Entertainment	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!