MACHINE LEARNING

EXERCISES

Elements of data clustering

All the course material is available on the web site

Course web site: https://github.com/unica-ml/ml

Accompanying notebook available at:

https://github.com/unica-ml/ml/blob/master/notebooks/ml07_clustering_exercises.ipynb

Exercise 1

Cluster the following data

X1=(1 1)’
X2=(1 0)’
X3=(0 1)’
X4=(5 0)’
X5=(4 1)’
X6=(3 2)’

assuming c=2 (c is the final number of clusters to obtain)

1) Apply the K-means algorithm, using the L2 distance and C1= (0,0)’; C2= (2,1)’ as initial centroids.

2) Apply the single-linkage clustering algorithm using the L2 distance as the sample-wise distance.

3) Apply the centroid-linkage clustering algorithm using the L2 distance as the sample-wise distance.

1) K-means

begin initialize n, c, $\mathbf{m}_1$, $\mathbf{m}_2$, ... , $\mathbf{m}_c$
do classify n patterns according to nearest $\mathbf{m}_i$
recompute $\mathbf{m}_i$
until no change in $\mathbf{m}_i$
return $\mathbf{m}_1$, $\mathbf{m}_2$, ... , $\mathbf{m}_c$
end

Step 1
C1= [0,0]'; C2= [2,1]';

The table reports squared Euclidean distances of samples vs centroids
(taking the square root is not necessary…)

	C1	C2
x1	2	**1**
x2	**1**	2
x3	**1**	4
x4	25	**10**
x5	17	**4**
x6	13	**2**

Cluster:
C1={x2,x3}; C2={x1,x4,x5,x6}
New centroids:
C1= [0.5, 0.5]'; C2= [3.25, 1]';

Step 2
C1= [0.5, 0.5]'; C2= [3.25, 1]';

	C1	C2
x1	**0.5**	5.06
x2	**0.5**	6.06
x3	**0.5**	10.56
x4	20.5	**4.06**
x5	12.5	**0.56**
x6	8.5	**1.06**

Cluster:
C1={x1,x2,x3}; C2={x4,x5,x6}
New centroids:
C1= [0.667 0.667]'; C2= [4 1]';

Step 3
C1= [0.667 0.667]'; C2= [4 1]';

	C1	C2
x1	**0.22**	9
x2	**0.56**	10
x3	**0.56**	16
x4	19.22	**2**
x5	11.22	**0**
x6	7.22	**2**

Cluster:
C1={x1,x2,x3}; C2={x4,x5,x6}
New centroids:
C1= [0.667 0.667]'; C2= [4 1]';

The algorithm has reached convergence; the final clustering is {x1, x2, x3}; {x4, x5, x6}

3) Single-linkage clustering algorithm

1. Initialize the algorithm by assuming that each sample is a cluster

2. Identify the two most similar clusters and merge them into a new cluster. Then compute distances w.r.t the new cluster, based on the linkage criterion.

3. Repeat step 2 until *c=2* clusters have been found.

STEP 1

(distances in the table are computed using the Euclidean distance – this time using the square root)

	C1{X1}	C2{X2}	C3{X3}	C4{X4}	C5{X5}	C6{X6}
C1{X1}	0.000	1.000	1.000	4.123	3.000	2.236
C2{X2}		0.000	1.414	4.000	3.162	2.828
C3{X3}			0.000	5.099	4.000	3.162
C4{X4}				0.000	1.414	2.828
C5{X5}					0.000	1.414
C6{X6}						0.000

The minimum distance is 1.000, so X2 and X3 are aggregated with X1.

STEP 2

New cluster C1{X1,X2,X3}. The distance between this cluster and C4 is obtained by taking the minimum distance among d(X1,X4), d(X2,X4), d(X3,X4) (**single-linkage criterion**).
The same process is repeated for C5 and C6.
Such distances are reported in bold below.

	C1{X1,X2,X3}	C4{X4}	C5{X5}	C6{X6}
C1{X1,X2,X3}	0.000	**4.000**	**3.000**	**2.236**
C4{X4}		0.000	1.414	2.828
C5{X5}			0.000	1.414
C6{X6}				0.000

The minimum distance is 1.414, so X5 and X6 are aggregated with X4.

STEP 3

New cluster C4{X4,X5,X6}

	C1{X1,X2,X3}	C4{X4,X5,X6}
C1{X1,X2,X3}	0.000	**2.236**
C4{X4,X5,X6}		0.000

The distance between these two remaining clusters is 2.236, which is the distance between X1 and X6 (i.e., the closest points belonging to different clusters).

Final clustering: {x1,x2,x3}, {x4,x5,x6}

3) Centroid-linkage clustering algorithm

1. Initialize the algorithm by assuming that each sample is a cluster

2. Identify the two most similar clusters and merge them into a new cluster. Then compute distances w.r.t the new cluster, based on the linkage criterion.

3. Repeat step 2 until *c=2* clusters have been found.

STEP 1

	C1{X1}	C2{X2}	C3{X3}	C4{X4}	C5{X5}	C6{X6}
C1{X1}	0.000	1.000	1.000	4.123	3.000	2.236
C2{X2}		0.000	1.414	4.000	3.162	2.828
C3{X3}			0.000	5.099	4.000	3.162
C4{X4}				0.000	1.414	2.828
C5{X5}					0.000	1.414
C6{X6}						0.000

STEP 2

New cluster C1{X1,X2,X3} ; m1= (2/3 2/3)'

Distances D(C1,C4), D(C1,C5), and D(C5,C6) have to be updated by computing the distance between the corresponding centroids of each cluster (**centroid-linkage criterion**). They are highlighted in bold below.

	C1{X1,X2,X3}	C4{X4}	C5{X5}	C6{X6}
C1{X1,X2,X3}	0.000	**4.384**	**3.350**	**2.687**
C4{X4}		0.000	1.414	2.828
C5{X5}			0.000	1.414
C6{X6}				0.000

STEP 3

New cluster C4{X4,X5,X6} ; m4= (4 1)'

	C1{X1,X2,X3}	C4{X4,X5,X6}
C1{X1,X2,X3}	0.000	**3.349**
C4{X4,X5,X6}		0.000

Final clustering: {x1,x2,x3}, {x4,x5,x6}

Exercise 2

Cluster 1	Cluster 2
X1=(1 1)’	X4=(5 0)’
X2=(1 0)’	X5=(4 1)’
X3=(0 1)’	X6=(3 2)’

Given the pattern in the table, say whether the division into clusters reflects the 'natural' classes according to the criterion functions

$$J_e = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2 \qquad J_d = \det(\mathbf{S}_W) = \left| \sum_{i=1}^{c} \mathbf{S}_i \right|$$

compared to the case where the pattern X6= (3 2)' is assigned to the cluster 1

Criterion functions in the first case

$$J_e = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2$$

$$m_1 = \begin{pmatrix} 2/3 \\ 2/3 \end{pmatrix}; \quad m_2 = \begin{pmatrix} 4 \\ 1 \end{pmatrix}$$

$$J_e = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2 = \sum_{\mathbf{x} \in D_1} \left\| \mathbf{x} - \mathbf{m}_1 \right\|^2 + \sum_{\mathbf{x} \in D_2} \left\| \mathbf{x} - \mathbf{m}_2 \right\|^2$$

=4/3+4=16/3≅5.333

$$J_d = \det(\mathbf{S}_W) = \left| \sum_{i=1}^{c} \mathbf{S}_i \right|$$

$$S_i = \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t \Rightarrow$$

$$S_1 = \sum_{\mathbf{x} \in D_1} (\mathbf{x} - \mathbf{m}_1)(\mathbf{x} - \mathbf{m}_1)^t; \; S_2 = \sum_{\mathbf{x} \in D_2} (\mathbf{x} - \mathbf{m}_2)(\mathbf{x} - \mathbf{m}_2)^t$$

$$S_W = \frac{1}{3} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} + \begin{pmatrix} 2 & -2 \\ -2 & 2 \end{pmatrix} = \frac{1}{3} \begin{pmatrix} 8 & -7 \\ -7 & 8 \end{pmatrix}$$

$$J_d = \det(\mathbf{S}_W) = 15/9 = 1.6667$$

Criterion functions in the second case

Cluster 1	Cluster 2
X1=(1 1)' X2=(1 0)' X3=(0 1)' X6=(3 2)'	X4=(5 0)' X5=(4 1)'

$$J_e = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2$$

$$m_1 = \begin{pmatrix} 1.25 \\ 1 \end{pmatrix}; \quad m_2 = \begin{pmatrix} 4.5 \\ 0.5 \end{pmatrix}$$

$$J_e = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2 = \sum_{\mathbf{x} \in D_1} \left\| \mathbf{x} - \mathbf{m}_1 \right\|^2 + \sum_{\mathbf{x} \in D_2} \left\| \mathbf{x} - \mathbf{m}_2 \right\|^2$$

=6.75 +1=7.75 (in the first case was 5.333)

$$J_d = \det(\mathbf{S}_W) = \left| \sum_{i=1}^{c} \mathbf{S}_i \right|$$

$$S_i = \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t =$$

$$\sum_{\mathbf{x} \in D_1} (\mathbf{x} - \mathbf{m}_1)(\mathbf{x} - \mathbf{m}_1)^t + \sum_{\mathbf{x} \in D_2} (\mathbf{x} - \mathbf{m}_2)(\mathbf{x} - \mathbf{m}_2)^t$$

$$= \begin{pmatrix} 4.75 & 2 \\ 2 & 2 \end{pmatrix} + \begin{pmatrix} 0.5 & -0.5 \\ -0.5 & 0.5 \end{pmatrix} = \begin{pmatrix} 5.25 & 1.5 \\ 1.5 & 2.5 \end{pmatrix}$$

$J_d = \det(\mathbf{S}_W) = 10.875$ (in the first case was 1.6667)

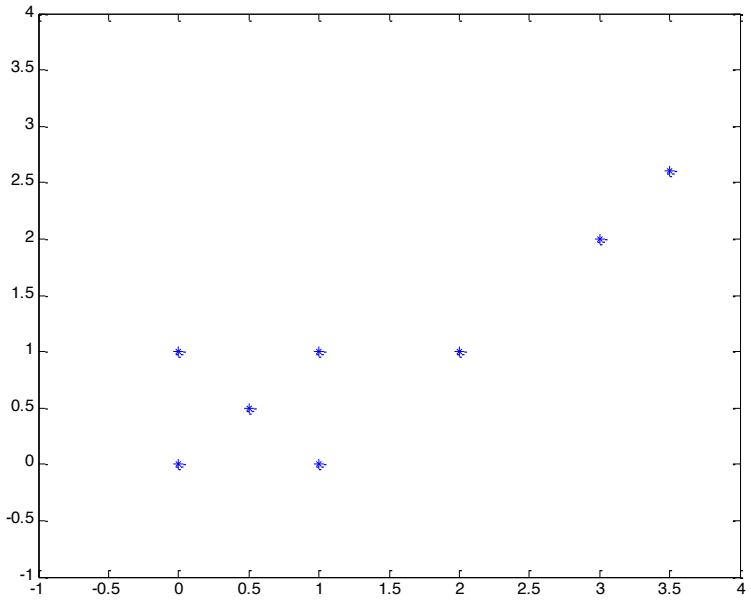
Clusterization '1' **Clusterization '2'**

Je=5.33 **Je=7.75**
Jd=1.66 **Jd=10.87**

The criterion functions show consistent results, indicating that the first division is the best.

Exercise 3

Cluster 1	Cluster 2
X1=(0 0)' X2=(0 1)' X3=(0.5 0.5)' X4=(1 0)' X5=(1 1)' X6=(2 1)'	X7=(3 2)' X8=(3.5 2.6)'



Given the pattern in the table, say whether the division into clusters reflects the 'natural' classes according to the criterion functions

$$J_e = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2 \qquad J_d = \det(\mathbf{S}_W) = \left| \sum_{i=1}^{c} \mathbf{S}_i \right|$$

compared to the case where the pattern X6= (3 2)' is assigned to the cluster 2

Criterion functions in the first case

$$J_e = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2$$

$$m_1 = \begin{pmatrix} 0.75 \\ 0.5833 \end{pmatrix}; \quad m_2 = \begin{pmatrix} 3.25 \\ 2.3 \end{pmatrix}$$

$$J_e = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2 = \sum_{\mathbf{x} \in D_1} \left\| \mathbf{x} - \mathbf{m}_1 \right\|^2 + \sum_{\mathbf{x} \in D_2} \left\| \mathbf{x} - \mathbf{m}_2 \right\|^2 \quad \text{=4.083 + 0.305 =4.388}$$

$$J_d = \det(\mathbf{S}_W) = \left| \sum_{i=1}^{c} \mathbf{S}_i \right|$$

$$S_i = \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t \Rightarrow$$

$$S_1 = \sum_{\mathbf{x} \in D_1} (\mathbf{x} - \mathbf{m}_1)(\mathbf{x} - \mathbf{m}_1)^t; \; S_2 = \sum_{\mathbf{x} \in D_2} (\mathbf{x} - \mathbf{m}_2)(\mathbf{x} - \mathbf{m}_2)^t$$

$$S_W = \begin{pmatrix} 2.875 & 0.625 \\ 0.625 & 1.208 \end{pmatrix} + \begin{pmatrix} 0.125 & 0.15 \\ 0.15 & 0.18 \end{pmatrix} = \begin{pmatrix} 3 & 0.7750 \\ 0.775 & 1.3883 \end{pmatrix}$$

$$J_d = \det(\mathbf{S}_W) = 3.5644$$

Criterion functions in the second case

Cluster 1	Cluster 2
X1=(0 0)' X2=(0 1)' X3=(0.5 0.5)' X4=(1 0)' X5=(1 1)'	X6=(2 1)' X7=(3 2)' X8=(3.5 2.6)'

$$J_e = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2$$

$$m_1 = \begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}; \quad m_2 = \begin{pmatrix} 2.833 \\ 1.866 \end{pmatrix}$$

$$J_e = \sum_{i=1}^{c} \sum_{\mathbf{x} \in D_i} \left\| \mathbf{x} - \mathbf{m}_i \right\|^2 = \sum_{\mathbf{x} \in D_1} \left\| \mathbf{x} - \mathbf{m}_1 \right\|^2 + \sum_{\mathbf{x} \in D_2} \left\| \mathbf{x} - \mathbf{m}_2 \right\|^2$$

= 2.000 + 2.473 =4.473 (in the first case was 4.388)

$$J_d = \det(\mathbf{S}_W) = \left| \sum_{i=1}^{c} \mathbf{S}_i \right|$$

$$S_i = \sum_{\mathbf{x} \in D_i} (\mathbf{x} - \mathbf{m}_i)(\mathbf{x} - \mathbf{m}_i)^t = \sum_{\mathbf{x} \in D_1} (\mathbf{x} - \mathbf{m}_1)(\mathbf{x} - \mathbf{m}_1)^t + \sum_{\mathbf{x} \in D_2} (\mathbf{x} - \mathbf{m}_2)(\mathbf{x} - \mathbf{m}_2)^t =$$

$$= \begin{pmatrix} 2.167 & 1.233 \\ 1.233 & 2.307 \end{pmatrix}$$

$J_d = \det(\mathbf{S}_W) = 3.4767$ (in the first case was 3.564)

Clusterization '1' **Clusterization '2'**

Je=4.388 **Je=4.473**
Jd=3.564 **Jd=3.477**

The criterion functions show **conflicting results**. Je indicates that the first division is the best; Jd indicates that the second division is the best, instead.