K- MEANS CLUSTERING

* The **k-means algorithm** is an algorithm to [cluster](http://en.wikipedia.org/wiki/Data_clustering) *n* objects based on attributes into *k* [partitions](http://en.wikipedia.org/wiki/Partition_of_a_set), where *k* < *n*.
* **Step 1:** Begin with a decision on the value of k =number of clusters.
* **Step 2**: Put any initial partition that classifies the data into k clusters. You may assign the training samples randomly, or systematically as the following:

1. Take the first k training sample as single-element clusters

2. Assign each of the remaining (N-k) training samples to the cluster with the nearest centroid. After each assignment, recomputed the centroid of the gaining cluster.

* **Step 3:** Take each sample in sequence and compute its [distance](http://people.revoledu.com/kardi/tutorial/Similarity/index.html) from the centroid of each of the clusters. If a sample is not currently in the cluster with the closest centroid, switch this sample to that cluster and update the centroid of the cluster gaining the new sample and the cluster losing the sample.
* **Step 4 .** Repeat step 3 until convergence is achieved, that is until a pass through the training sample causes no new assignments.

**EXAMPLE 1:**

Cluster the following eight points (with (x, y)) into three clusters

**A1(2,10) A2(2, 5) A3(8, 4) A4(5, 8) A5(7, 5) A6(6, 4) A7(1, 2) A8(4, 9).**

**SOLUTION**

Here k=3

Initial cluster centers are: **A1 (2, 10), A4 (5, 8) and A7 (1, 2).**

The distance function between two points a=(x1, y1) and b=(x2, y2) is defined as: ρ (a, b) = |x2 – x1| + |y2 – y1|.

Iteration 1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | (2, 10) | (5, 8) | (1, 2) |  |
|  | **Point** | **Dist Mean 1** | **Dist Mean 2** | **Dist Mean 3** | **Cluster** |
| A1 | (2, 10) | 0 | 5 | 9 | 1 |
| A2 | (2, 5) | 5 | 6 | 4 | 3 |
| A3 | (8, 4) | 12 | 7 | 9 | 2 |
| A4 | (5, 8) | 5 | 0 | 10 | 2 |
| A5 | (7, 5) | 10 | 5 | 9 | 2 |
| A6 | (6, 4) | 10 | 5 | 7 | 2 |
| A7 | (1, 2) | 9 | 10 | 0 | 3 |
| A8 | (4, 9) | 3 | 2 | 10 | 2 |

= dist b/w (4, 9) and (2, 10) =| (2-4) + (10-9)|=2+1=3

Now the cluster’s are

Cluster 1 Cluster 2 Cluster 3

(2, 10) (8, 4) (2, 5)

(5, 8) (1, 2)

(7, 5)

(6, 4)

(4, 9)

Clusters: c1: {A1},

C2 :{ A3, A4, A5, A6, A8},

C3: {A2, A7}

Next, we need to re-compute the new cluster centers (means). We do so, by taking the mean of all points in each cluster.

* For Cluster 1, we only have one point A1 (2, 10), which was the old mean, so the cluster center remains the same.
* For Cluster 2, we have ( (8+5+7+6+4)/5, (4+8+5+4+9)/5 ) = (6, 6)
* For Cluster 3, we have ( (2+1)/2, (5+2)/2 ) = (1.5, 3.5)
* Now cluster centers are: **A1 (2, 10), A4 (6, 6) and A7 (1.5, 3.5).**

Iteration 2:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | (2,10) | (6,6) | (1.5,3.5) |  |
|  | **Point** | **Dist Mean 1** | **Dist Mean 2** | **Dist Mean 3** | **Cluster** |
| A1 | (2, 10) | 0 | 8 | 7 | 1 |
| A2 | (2, 5) | 5 | 5 | 2 | 3 |
| A3 | (8, 4) | 12 | 4 | 7 | 2 |
| A4 | (5, 8) | 5 | 3 | 8 | 2 |
| A5 | (7, 5) | 10 | 2 | 7 | 2 |
| A6 | (6, 4) | 10 | 2 | 5 | 2 |
| A7 | (1, 2) | 9 | 9 | 2 | 3 |
| A8 | (4, 9) | 3 | 5 | 8 | 1 |

Now the cluster’s are

Cluster 1 Cluster 2 Cluster 3

(2, 10) (8, 4) (2, 5)

(4, 9) (5, 8) (1, 2)

(7, 5)

(6, 4)

Clusters: C1: {A1, A8},

C2 :{ A3, A4, A5, A6},

C3: {A2, A7}

Next, we need to re-compute the new cluster centers (means). We do so, by taking the mean of all points in each cluster.

* For Cluster 1, we have(( 2+4)/2,(10+9)/2)=(3,9.5)
* For Cluster 2, we have ( (8+5+7+6)/4, (4+8+5+4)/4 ) = (6.5, 5.25)
* For Cluster 3, we have ( (2+1)/2, (5+2)/2 ) = (1.5, 3.5)
* Now cluster centers are: **A1 (3, 9.5), A4 (6.5, 5.25) and A7 (1.5, 3.5).**

Iteration 3:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | (3,9.5) | (6.5,5.25) | (1.5,3.5) |  |
|  | **Point** | **Dist Mean 1** | **Dist Mean 2** | **Dist Mean 3** | **Cluster** |
| A1 | (2, 10) | 1.5 | 9.25 | 7 | 1 |
| A2 | (2, 5) | 5.5 | 4.75 | 2 | 3 |
| A3 | (8, 4) | 9.5 | 2.75 | 7 | 2 |
| A4 | (5, 8) | 3.5 | 4.25 | 8 | 1 |
| A5 | (7, 5) | 8.5 | .75 | 7 | 2 |
| A6 | (6, 4) | 8.5 | 1.75 | 5.5 | 2 |
| A7 | (1, 2) | 9.5 | 8.75 | 2 | 3 |
| A8 | (4, 9) | 1.5 | 6.25 | 5.5 | 1 |

Now the cluster’s are

Cluster 1 Cluster 2 Cluster 3

(2, 10) (8, 4) (2, 5)

(4, 9) (7, 5) (1, 2)

(5, 8) (6, 4)

Clusters: C1: {A1, A4, And A8},

C2 :{ A3, A5, A6},

C3: {A2, A7}

Next, we need to re-compute the new cluster centers (means). We do so, by taking the mean of all points in each cluster.

* For Cluster 1, we have(( 2+4+5)/3,(10+9+8)/3)=(3.66,9)
* For Cluster 2, we have ( (8+7+6)/3, (4+5+4)/3 ) = (7,4.33)
* For Cluster 3, we have ( (2+1)/2, (5+2)/2 ) = (1.5, 3.5)
* Now cluster centers are: **A1 (3.66, 9), A4 (7, 4.33) and A7 (1.5, 3.5).**
* Iteration 4:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | (3.66,9) | (7,4.33) | (1.5,3.5) |  |
|  | **Point** | **Dist Mean 1** | **Dist Mean 2** | **Dist Mean 3** | **Cluster** |
| A1 | (2, 10) | 2.66 | 10.67 | 7 | 1 |
| A2 | (2, 5) | 5.66 | 5.66 | 2 | 3 |
| A3 | (8, 4) | 9.34 | 1.33 | 7 | 2 |
| A4 | (5, 8) | 2.34 | 5.66 | 8 | 1 |
| A5 | (7, 5) | 7.34 | .66 | 7 | 2 |
| A6 | (6, 4) | 7.34 | 1.33 | 5.5 | 2 |
| A7 | (1, 2) | 9.66 | 8.33 | 2 | 3 |
| A8 | (4, 9) | 0.34 | 9.66 | 5.5 | 1 |

The means do not change anymore. Finally we get

Clusters: C1: {A1, A4, and A8}, C2 :{ A3, A5, A6} C3: {A2, A7}

**EXAMPLE II (using mamhatten distance)**

****

|  |  |  |
| --- | --- | --- |
| **Individual** | **Variable 1** | **Variable 2** |
|  | 1.0 | 1.0 |
|  | 1.5 | 2.0 |
|  | 3.0 | 4.0 |
|  | 5.0 | 7.0 |
|  | 3.5 | 5.0 |
|  | 4.5 | 5.0 |
|  | 3.5 | 4.5 |

Assume k=2

Assume initial centroid 1(1.0, 1.0), 4(5.0, 7.0)

|  |  |  |  |
| --- | --- | --- | --- |
| **Individual** | **(1.0,1.0)** | **(5.0,7.0)** | **cluster** |
|  | 0 | 7.21 | 1 |
|  | 1.12 | 6.10 | 1 |
|  | 3.61 | 3.61 | 1 |
|  | 7.21 | 0 | 2 |
|  | 4.72 | 2.5 | 2 |
|  | 5.31 | 2.06 | 2 |
|  | 4.30 | 2.92 | 2 |

Distance between (1.0, 1.0) and individual 2(1.5, 2.0) =√|1.0-1.5| 2+|1.0-2.0|2=1.118

* Thus, we obtain two clusters containing :{ 1, 2, and 3} and {4, 5, 6, 7}.

Next, we need to re-compute the new cluster centers (means). We do so, by taking the mean of all points in each cluster.

* For Cluster 1, we have(( 1.0+1.5+3.0)/3,(1.0+2.0+4.0)/3)=(1.83,2.33)
* For Cluster 2, we have ( (5.0+3.5+4.5+3.5)/4, (7.0+5.0+5.0+4.5)/4 ) = (4.12,5.38)
* Now using these centroids we compute the Euclidean distance of each object, as shown in table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Individual** | **(1.83,2.33)** | **(4.12,5.38)** | **cluster** |
|  | 1.57 | 5.38 | 1 |
|  | 0.47 | 4.28 | 1 |
|  | 2.04 | 1.78 | 2 |
|  | 5.64 | 1.84 | 2 |
|  | 3.15 | 0.73 | 2 |
|  | 3.78 | 0.54 | 2 |
|  | 2.74 | 1.08 | 2 |

Thus, we obtain two clusters containing :{ 1, 2} and {3, 4, 5, 6, 7}.

Next, we need to re-compute the new cluster centers (means). We do so, by taking the mean of all points in each cluster.

* For Cluster 1, we have(( 1.0+1.5)/2,(1.0+2.0)/2)=(1.25,1.5)
* For Cluster 2, we have ( (3.0+5.0+3.5+4.5+3.5)/5, (4.0+7.0+5.0+5.0+4.5)/5 ) = (3.9,5.1 )
* Now using these centroids we compute the Euclidean distance of each object, as shown in table.

|  |  |  |  |
| --- | --- | --- | --- |
| **Individual** | **(1.25,1.5)** | **(3.9,5.1)** | **cluster** |
|  | 0.56 | 5.02 | 1 |
|  | 0.56 | 3.92 | 1 |
|  | 3.05 | 1.42 | 2 |
|  | 6.66 | 2.20 | 2 |
|  | 4.16 | 0.41 | 2 |
|  | 4.78 | 0.61 | 2 |
|  | 3.75 | 0.72 | 2 |

The means do not change anymore. Finally we get

Clusters: C1: {1, 2}, C2 :{ 3, 4, 5, 6, 7}.

**Same problem k=3, initial centriod1 (1.0, 1.0), 2(1.5, 2.0),3(3.0,4.0)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Individual** | **(1.0,1.0)** | **(1.5,2.0)** | **(3.0,4.0)** | **cluster** |
|  | 0 | 1.11 | 3.61 | 1 |
|  | 1.12 | 0 | 2.5 | 2 |
|  | 3.61 | 2.5 | 0 | 3 |
|  | 7.21 | 6.10 | 3.61 | 3 |
|  | 4.72 | 3.61 | 1.12 | 3 |
|  | 5.31 | 4.24 | 1.80 | 3 |
|  | 4.30 | 3.20 | 0.71 | 3 |

New centroid are (1.0, 1.0), (1.5,2.0),(3.9,5.1)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Individual** | **(1.0,1.0)** | **(1.5,2.0)** | **(3.0,4.0)** | **cluster** |
|  | 0 | 1.11 | 5.02 | 1 |
|  | 1.12 | 0 | 3.92 | 2 |
|  | 3.61 | 2.5 | 1.42 | 3 |
|  | 7.21 | 6.10 | 2.20 | 3 |
|  | 4.72 | 3.61 | 0.41 | 3 |
|  | 5.31 | 4.24 | 0.61 | 3 |
|  | 4.30 | 3.20 | 0.72 | 3 |

The means do not change anymore.

Finally we get Clusters: C1: {1}, C2 :{ 2}, c3 :{ 3, 4, 5, 6, 7}.

Real world problem

We have 4 medicines as our training data points object and each medicine has 2 attributes. Each attribute represents coordinate of the object. We have to determine which medicines belong to cluster 1 and which medicines belong to the other cluster.

|  |  |  |
| --- | --- | --- |
| object | Attribute 1 | Attribute 2 |
| Medicine A | 1 | 1 |
| Medicine b | 2 | 1 |
| Medicine c | 4 | 3 |
| Medicine d | 5 | 4 |

K=2 initial centroid A (1, 1), B (2, 1)

|  |  |  |  |
| --- | --- | --- | --- |
| object | (1,1) | 2,1) | cluster |
| Medicine A(1,1) | 0 | 1 | 1 |
| Medicine b(2,1) | 1 | 0 | 2 |
| Medicine c(4,3) | 3.61 | 2.83 | 2 |
| Medicine d(5,4) | 5 | 4.24 | 2 |

New centroid (1,1)

Second (2+4+5/2),(1+3+4)/3=(3.66,2.66)

|  |  |  |  |
| --- | --- | --- | --- |
| object | (1,1) | (3.66,2.66) | cluster |
| Medicine A(1,1) | 0 | 3.14 | 1 |
| Medicine b(2,1) | 1 | 2.36 | 1 |
| Medicine c(4,3) | 3.61 | 0.47 | 2 |
| Medicine d(5,4) | 5 | 1.89 | 2 |

New centroid (1+2/2), (1+1/2) = (1.5, 1)

(4+5/2), (3+4/2) = (4.5, 3.5)

|  |  |  |  |
| --- | --- | --- | --- |
| object | (1.5,1) | (4.5,3.5) | cluster |
| Medicine A(1,1) | 0.5 | 4.30 | 1 |
| Medicine b(2,1) | 0.5 | 3.54 | 1 |
| Medicine c(4,3) | 3.20 | 0.71 | 2 |
| Medicine d(5,4) | 4.61 | 0.71 | 2 |

The means do not change anymore.

Finally we get Clusters: C1: {1,2}, C2 :{ 3,4}