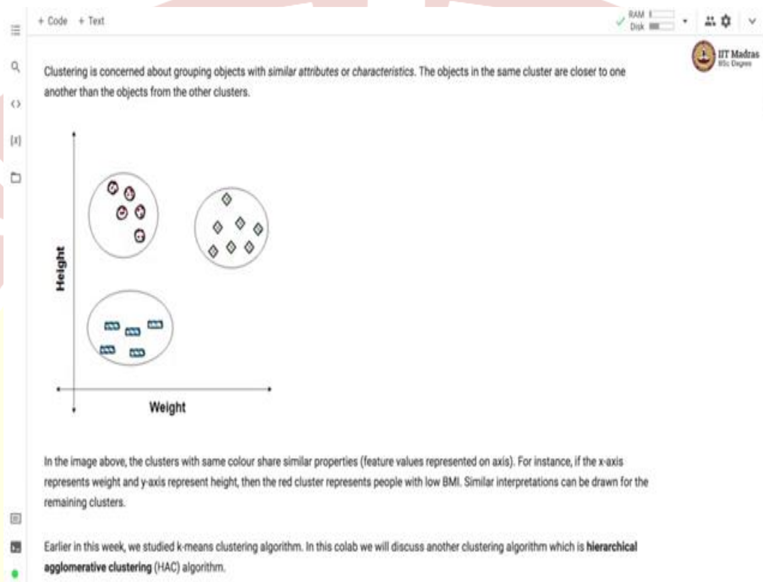
INDIAN INSTITUTE OF TECHNOLOGY MADRAS • सिद्धिर्भवति कर्मजा •

IIT Madras

ONLINE DEGREE

Machine Learning Practice
Indian Institute of Technology, Madras
Programming and Data Science
HAC Demo

(Refer Slide Time: 00:10)



Namaste! Welcome to the next video of Machine Learning Practice Course. In this video, we will study 1 more clustering algorithm, which is hierarchical agglomerative clustering. Clustering, as you know is concerned about grouping objects with similar attributes or characteristics. The objects in the same cluster are closer to one another than the objects from the other clusters.

In the image above, the clusters with the same color share similar properties. The properties are marked on the axis. There is a weight and height, those are the 2 properties. So, you can see that the red cluster probably represent people with low BMI. Similar interpretations can also be drawn from the other 2 clusters.

(Refer Slide Time: 00:59)

Earlier in this week, we studied K-means clustering algorithm. Now we will study hierarchical agglomerative clustering algorithm or in general hierarchical clustering algorithms. There are 2 flavors of hierarchical clustering algorithms. One is a bottom's up approach and another is a top-down approach. In bottoms up approach, we consider each data point as a cluster, and then we combine them to form a larger cluster.

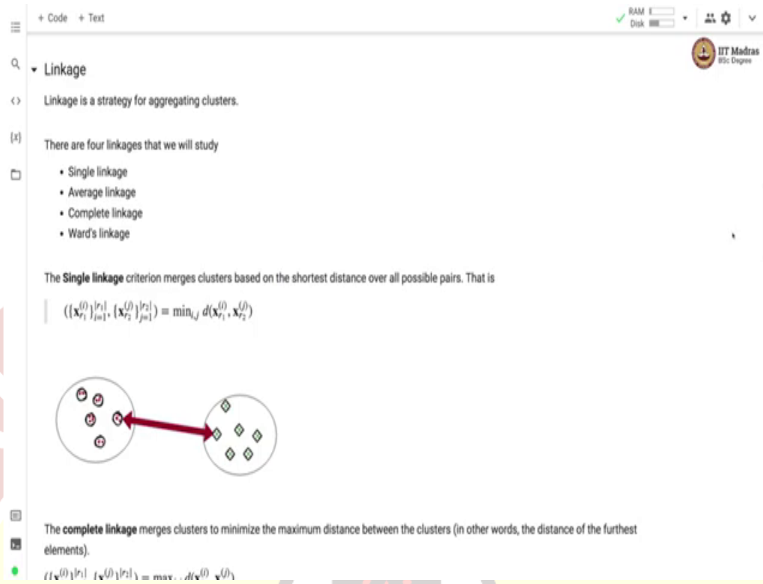
We keep repeating this process until we are left with a single cluster containing all the points. In the top-down approach, the entire data is considered as a single cluster and then it is divided to form smaller clusters in subsequent steps. The merging and splitting decisions are influenced by certain conditions that we will be discussing shortly.

(Refer Slide Time: 01:47)

The merging and splitting decisions are taken based on metrics. Metrics are used for calculating similarity between clusters. Note that, the metric is a generalization of concept of distance. The metrics follow certain properties like that they should be non-negative, they should be symmetric, and they should follow triangle inequality. So, the popular metric functions are Euclidean, Manhattan and cosine metrics.

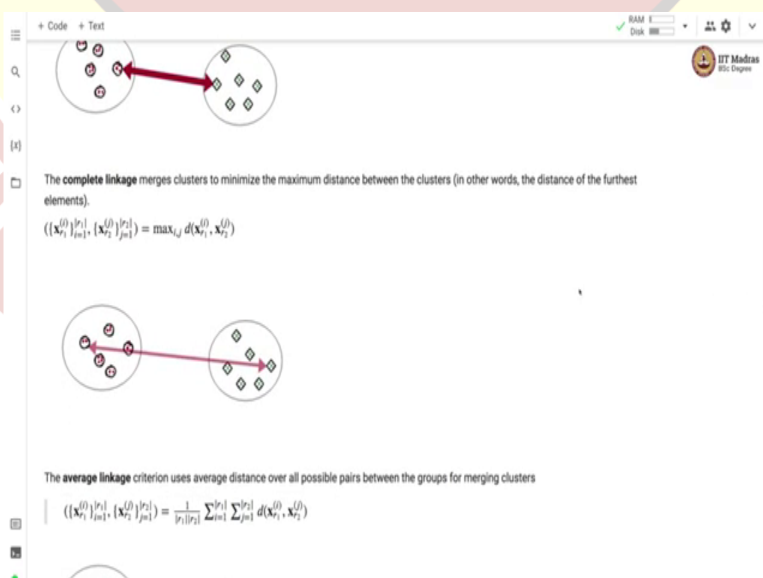
Euclidean and Manhattan matrix are very similar to Euclidean distance and Manhattan distance that we studied in nearest neighbor classifier. Cosine distance, on the other hand, calculates the **.** product between the vector and divide it by the product of the norm of the vectors and the resulting ratio is subtracted from 1 and that gives us the cosine distance.

(Refer Slide Time: 02:40)



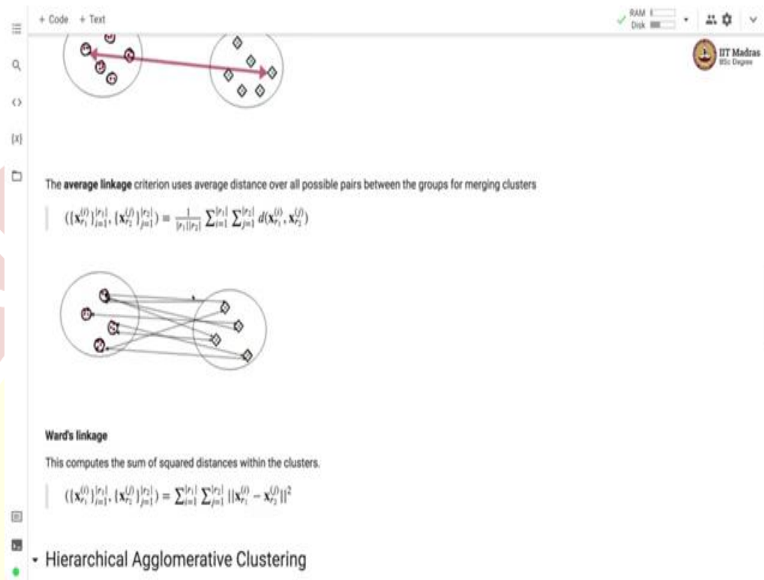
Linkage is a strategy for aggregating clusters; there are 4 linkages that we will study. One is Single linkage, second is Average linkage and it is Complete linkage and Ward's linkage. Single linkage criterion merges clusters based on the shortest distance over all possible pairs. We calculate the distance; the shortest distance between 2 clusters based on the closest point, and based on that we perform the merging of the clusters.

(Refer Slide Time: 03:09)



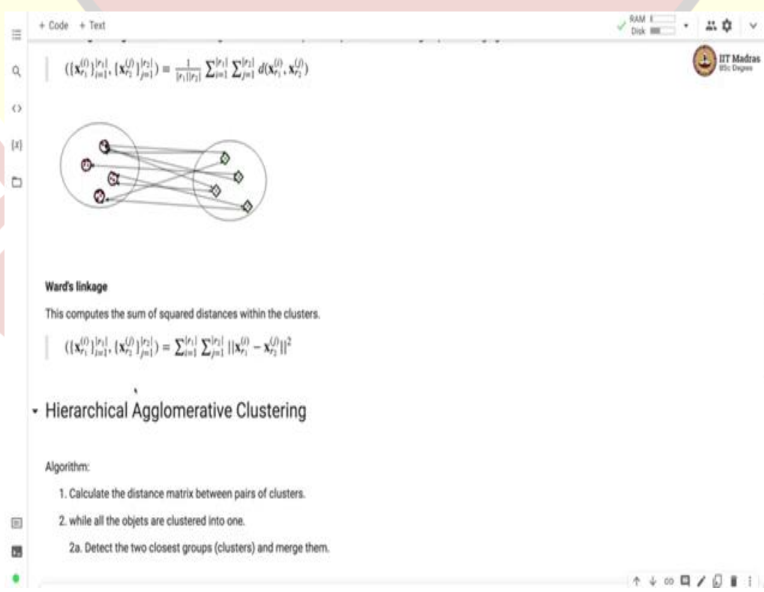
In complete linkage, we calculate distance between the furthest points in the cluster and combined 2 clusters that minimize this maximum distance between 2 furthest points.

(Refer Slide Time: 03:26)



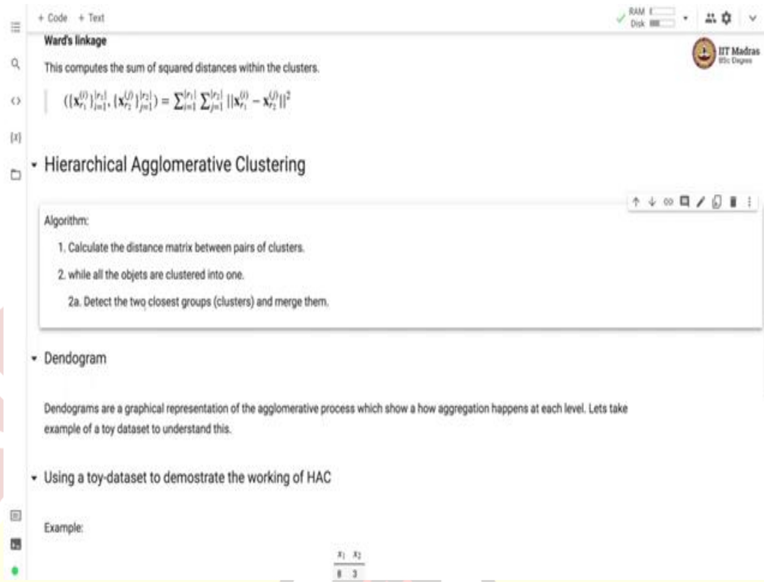
The average linkage criterion uses average distance over all possible pairs between the group of margin clusters.

(Refer Slide Time: 03:33)



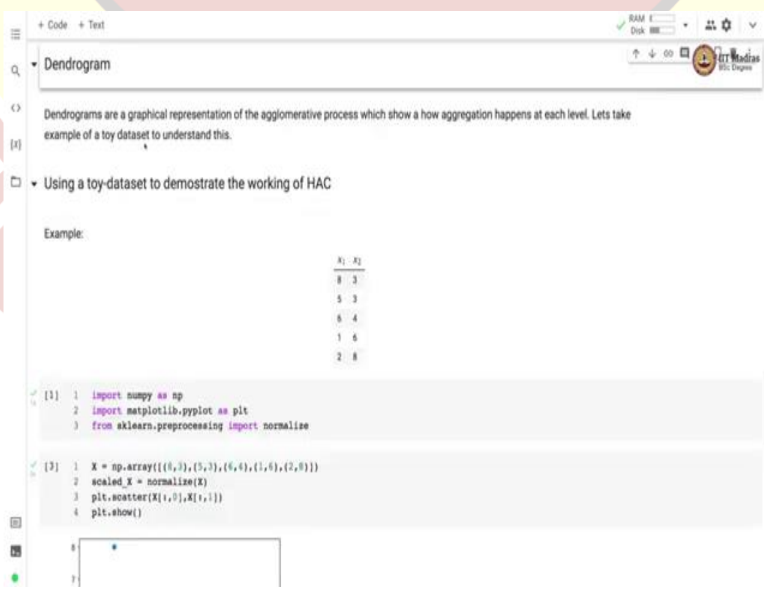
And finally, Ward's linkage calculates the sum of square distance within the clusters.

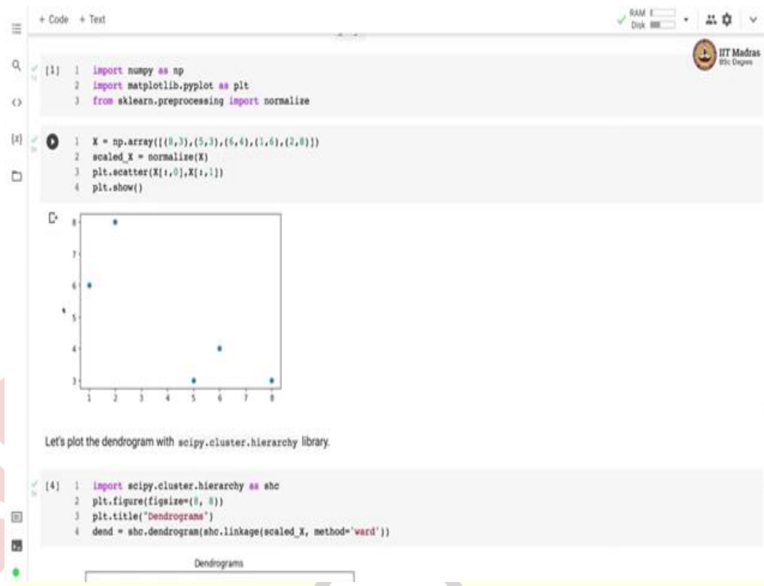
(Refer Slide Time: 03:38)

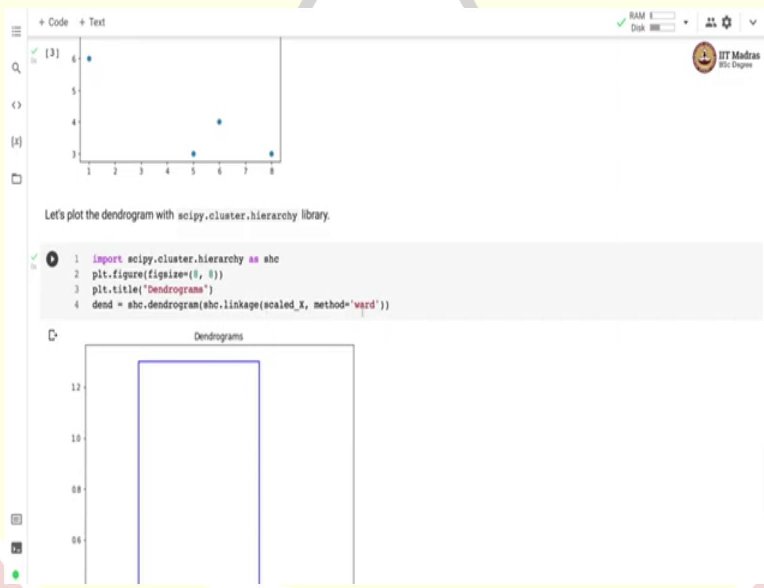


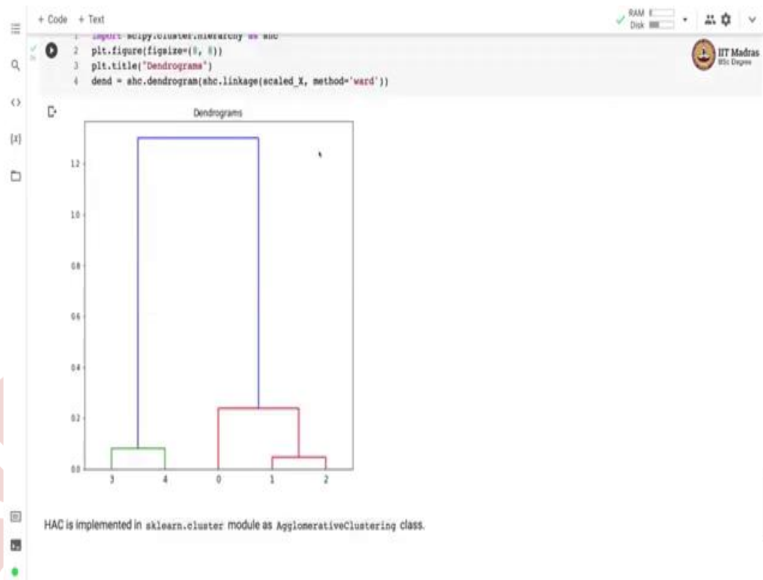
Let us look at hierarchical agglomerative clustering algorithm. In the first step, we calculate distance matrix between pairs of clusters. And this distance matrix is calculated based on the chosen metric. Then the second step, we run a loop while all objects are clustered into one. We first detect 2 closest groups and merge them and we keep repeating this process until we get a single cluster consisting of all the points.

(Refer Slide Time: 04:07)




+ Code + Text
RAM
Disk
IIT Madras
BSc Degree
[1]
import numpy as np
import matplotlib.pyplot as plt
from sklearn.preprocessing import normalize
X = np.array([(8,3),(5,3),(6,4),(1,6),(2,8)])
scaled_X = normalize(X)
plt.scatter(X[:,0],X[:,1])
plt.show()
Let's plot the dendrogram with scipy.cluster.hierarchy library.
[4]
import scipy.cluster.hierarchy as shc
plt.figure(figsize=(8, 8))
plt.title("Dendrograms")
dend = shc.dendrogram(shc.linkage(scaled_X, method='ward'))
Dendrograms


+ Code + Text
RAM
Disk
IIT Madras
BSc Degree
[3]
Let's plot the dendrogram with scipy.cluster.hierarchy library.
import scipy.cluster.hierarchy as shc
plt.figure(figsize=(8, 8))
plt.title("Dendrograms")
dend = shc.dendrogram(shc.linkage(scaled_X, method='ward'))
Dendrograms
1.2
1.0
0.8
0.6



Dendrogram are a graphical representation of agglomerative process which shows how aggregation happens at each level. Let us take an example to understand article clustering and the resulting dendrogram. Here we take a toy dataset consisting of 5 examples. We scale these examples; we plot these points in 2d plot with feature x1 on X-axis and feature x2 on Y-axis. Then you plot dendrogram with scipy.cluster.hierarchy library.

We first link the points using Ward method and then plot the dendrogram. This is the resulting dendrogram. You can see that point 1 and 2 are the closest one and they got merged in the initial stage. Then the point 3 and 4 were the second closest point and they got merged later. Point 0 and the cluster of 1 and 2 were closest so they got merged in the next step. And finally, we merge these 2 resulting clusters.

So, this is how the aggregation of cluster happened and which is represented in dendrograms. So, HAC is implemented in a sklearn.cluster module as AgglomerativeClustering class. Now, what you can do is you can cut the dendrogram as at arbitrary places and form clusters. So, if you cut the dendrogram over here, we will get these 2 clusters one containing point 3 and 4 and second containing point 0, 1 and 2.

If we cut the dendrogram over here, you can see that there are 3 clusters that will get formed, one which is with point 3, 4 the midpoint 0 and third with point 1 and 2. In this video, we studied hierarchical agglomerative clustering, and various linkages and metrics used in hierarchical agglomerative clustering. We also looked at dendrograms that shows the aggregation of clusters in the bottoms up approach.