Unsupervised Learning

Nipun Batra

July 12, 2020

IIT Gandhinagar

Lecture heavily adapted from Kevin Murphy's book

Unsupervised Learning

- Unsupervised learning: we are just given output data, without any inputs

Unsupervised Learning

- Unsupervised learning: we are just given output data, without any inputs
- The goal is to discover “interesting structure” in the data; this is sometimes called knowledge discovery.

Unsupervised Learning

- Unsupervised learning: we are just given output data, without any inputs
- The goal is to discover “interesting structure” in the data; this is sometimes called knowledge discovery.
- Unlike supervised learning, we are not told what the desired output is for each input.

Unsupervised Learning

- Unsupervised learning: we are just given output data, without any inputs
- The goal is to discover "interesting structure" in the data; this is sometimes called knowledge discovery.
- Unlike supervised learning, we are not told what the desired output is for each input.
- Instead, we will formalize our task as one of density estimation, that is, we want to build models of the form $p\left(\mathrm{x}_i \mid \boldsymbol{\theta}\right)$.

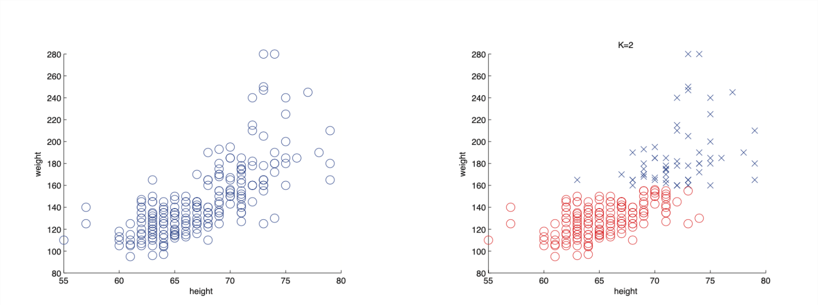
Differences between supervised and unsupervised learning

- We have written $p\left(\mathrm{x}_i \mid \boldsymbol{\theta}\right)$ instead of $p\left(y_i \mid \mathrm{x}_i, \boldsymbol{\theta}\right)$; that is, supervised learning is conditional density estimation, whereas unsupervised learning is unconditional density estimation.

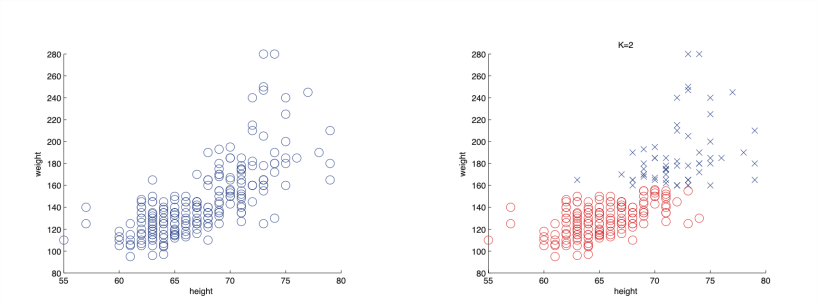
Differences between supervised and unsupervised learning

- We have written $p(\mathbf{x}_i \mid \boldsymbol{\theta})$ instead of $p(y_i \mid \mathbf{x}_i, \boldsymbol{\theta})$; that is, supervised learning is conditional density estimation, whereas unsupervised learning is unconditional density estimation.
- $\mathbf{x}_i$ is a vector of features, so we need to create multivariate probability models. By contrast, in supervised learning, y_i is usually just a single variable that we are trying to predict. This means that for most supervised learning problems, we can use univariate probability models (with input-dependent parameters), which significantly simplifies the problem.

Some categories of unsupervised algorithms: Clustering

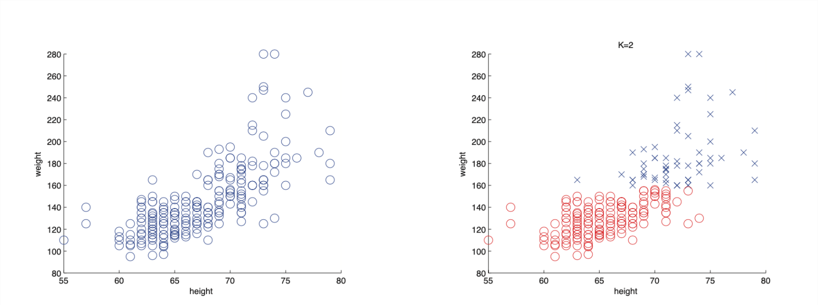


Some categories of unsupervised algorithms: Clustering



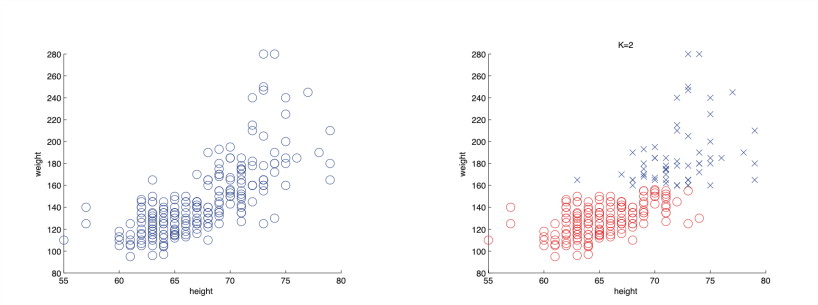
- Divide data into groups or clusters

Some categories of unsupervised algorithms: Clustering



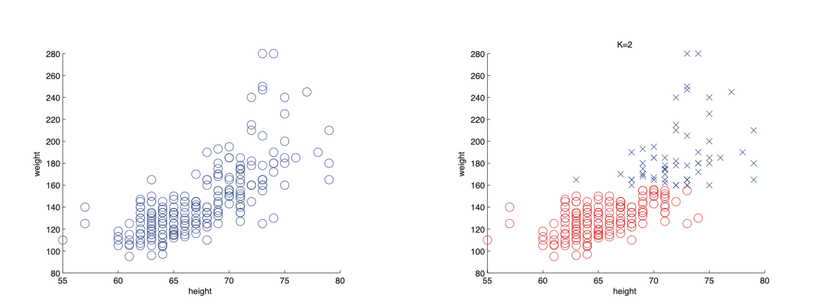
- Divide data into groups or clusters
- Assuming K clusters, we have two goals:

Some categories of unsupervised algorithms: Clustering



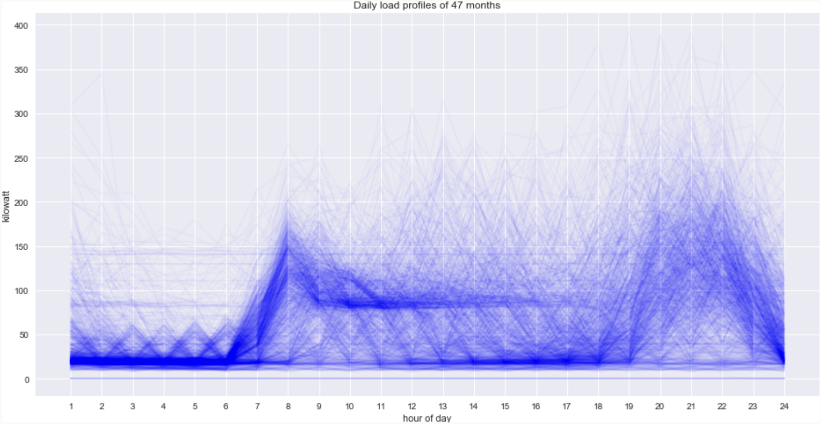
- Divide data into groups or clusters
- Assuming K clusters, we have two goals:
    1. estimate the distribution over the number of clusters, $p(K|D)$

Some categories of unsupervised algorithms: Clustering

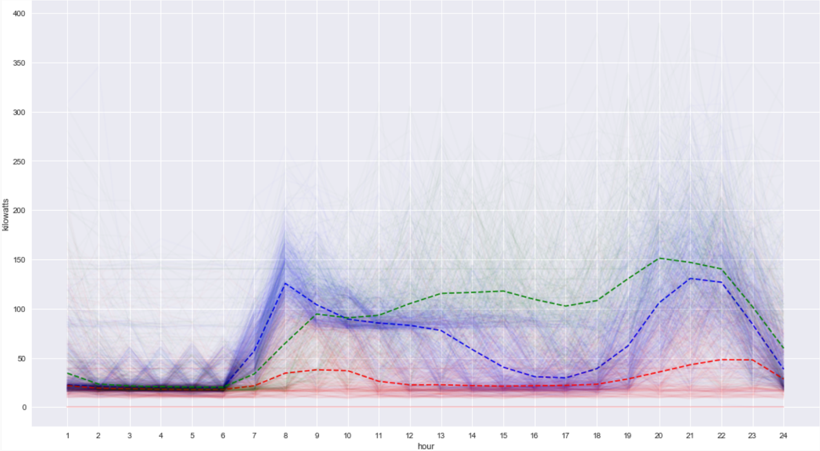


- Divide data into groups or clusters
- Assuming K clusters, we have two goals:
  1. estimate the distribution over the number of clusters, p(K|D)
  2. estimate which cluster each point belongs to. Let $z_i \in \{1, \ldots, K\}$ represent the cluster to which data point i is assigned.

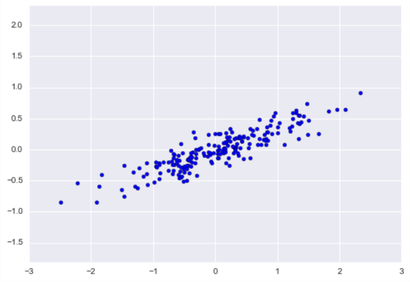
Some categories of unsupervised algorithms: Clustering



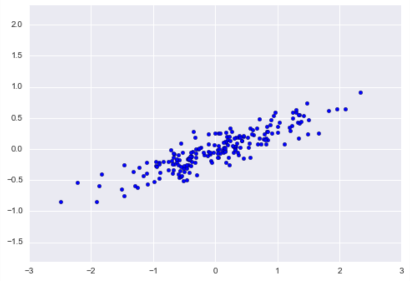
Some categories of unsupervised algorithms: Clustering



Some categories of unsupervised algorithms: Dimensionality reduction

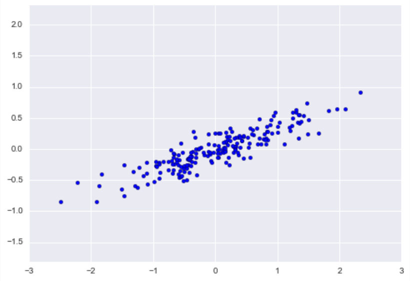


Some categories of unsupervised algorithms: Dimensionality reduction



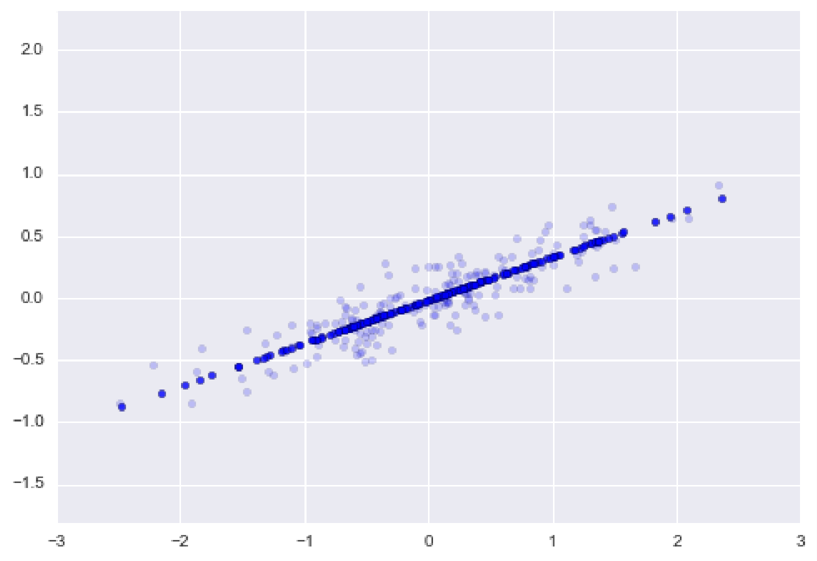
- Definition: reduce the dimensionality by projecting the data to a lower dimensional subspace which captures the “essence” of the data.

Some categories of unsupervised algorithms: Dimensionality reduction

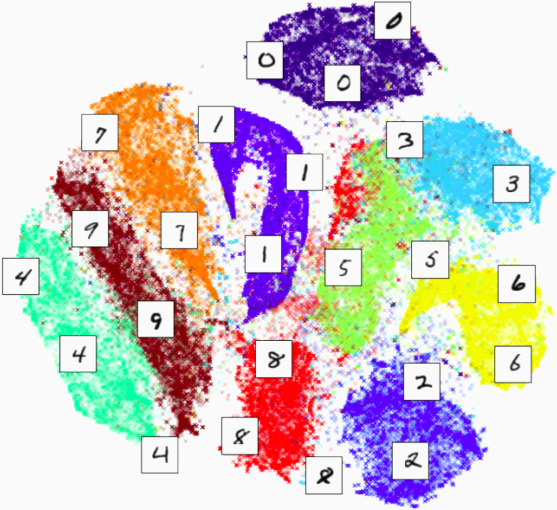


- Definition: reduce the dimensionality by projecting the data to a lower dimensional subspace which captures the "essence" of the data.
- Motivation: although the data may appear high dimensional, there may only be a small number of degrees of variability, corresponding to latent factors.

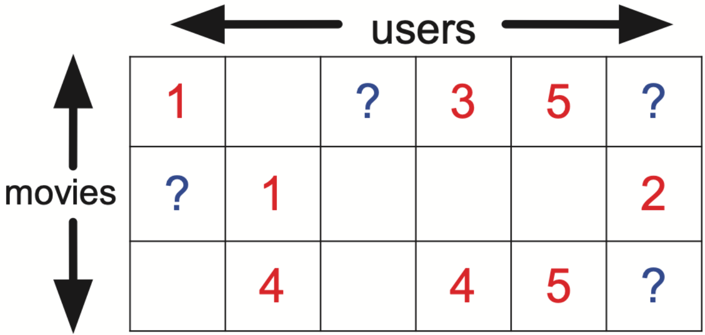
Some categories of unsupervised algorithms: Dimensionality reduction



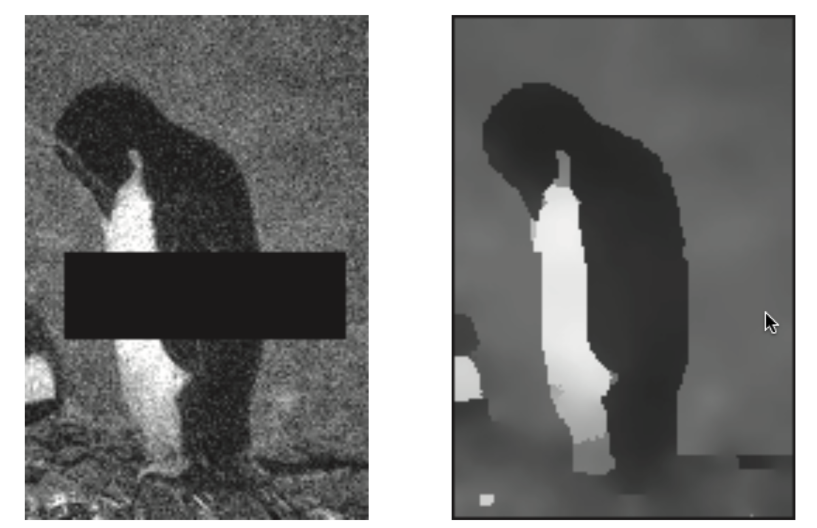
Some categories of unsupervised algorithms: Dimensionality reduction



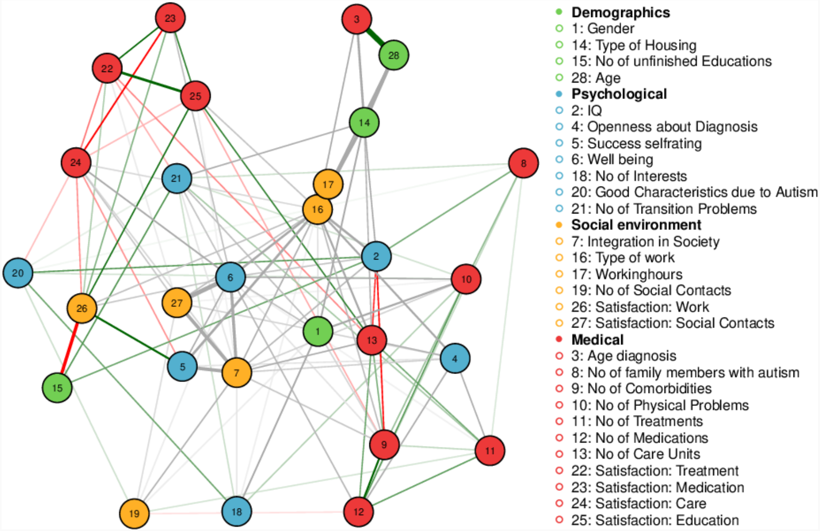
Some categories of unsupervised algorithms: Matrix completion



Some categories of unsupervised algorithms: Matrix completion



Some categories of unsupervised algorithms: Discovering Graph Structure



The need for Unsupervised Learning

- Aids the search of patterns in data.
- Find features for categorization.
- Easier to collect unlabeled data.

The need for Unsupervised Learning

- Aids the search of patterns in data.
- Find features for categorization.
- Easier to collect unlabeled data.

Places where you will see unsupervised learning

- It can be used to segment the market based on customer preferences.
- A data science team reduces the number of dimensions in a large data set to simplify modeling and reduce file size.

Clustering

Clustering

- **AIM:** To find groups/subgroups in a data set.

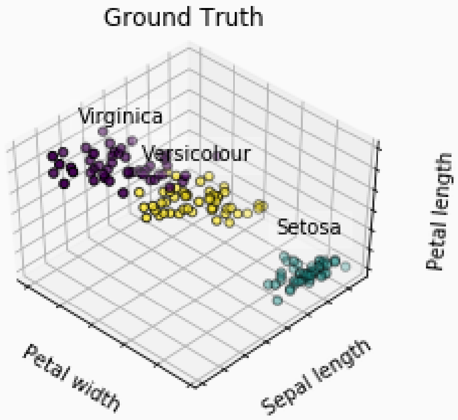
Clustering

- AIM: To find groups/subgroups in a data set.
- REQUIREMENTS: A predefined notion of similarity/dissimilarity.

Clustering

- AIM: To find groups/subgroups in a data set.
- REQUIREMENTS: A predefined notion of similarity/dissimilarity.
- Examples: Market Segmentation: Customers with similar preferences in the same groups. This would aid in targeted marketing.

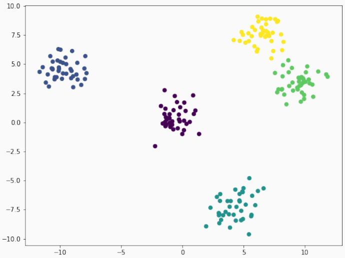
Clustering



Iris Data Set with ground truth

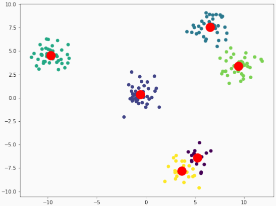
K-Means Clustering

- N points in a R^d space.
- C_i: set of points in the i^{th} cluster.
- $C_1 \cup C_2 \cup \ldots C_k = \{1, \ldots, n\}$
- $C_i \cap C_j = \{\phi\}$ for $i \neq j$

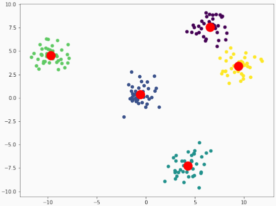


Dataset with 5 clusters

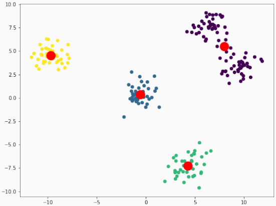
K-Means Clustering



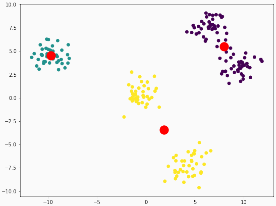
K=6



K=5



K=4



K=3

K-Means Intuition

- Good Clustering: Within the cluster the variation (*WCV*) is small.

K-Means Intuition

- Good Clustering: Within the cluster the variation (WCV) is small.
- Objective:

$$\min_{C_1,\ldots,C_k}\left(\sum_{i=1}^{k} WCV\left(C_i\right)\right)$$

K-Means Intuition

- Good Clustering: Within the cluster the variation (WCV) is small.
- Objective:

$$\min_{C_1,\ldots,C_k}\left(\sum_{i=1}^{k} WCV\left(C_i\right)\right)$$

Minimize the WCV as much as possible

K-Means Intuition

Objective:

$$\min_{C_1,\ldots,C_k} \left(\sum_{i=1}^{k} WCV\left(C_i\right) \right)$$

Objective:

$$\min_{C_1,\ldots,C_k} \left(\sum_{i=1}^{k} WCV\left(C_i\right) \right)$$

$$WCV\left(C_i\right) = \frac{1}{|C_i|} \texttt{ (Distance between all points)}$$

$$WCV\left(C_i\right) = \frac{1}{|C_i|} \sum_{a \in C_i} \sum_{b \in C_i} ||x_a - x_b||_2^2$$

where $|C_i|$ is the number of points in C_i

K-Means Algorithm

1. Randomly assign a cluster number i to every point (where $i \in \{1, \ldots n\}$)

K-Means Algorithm

1. Randomly assign a cluster number i to every point (where $i \in \{1, \ldots n\}$)

    2.1 For each cluster C_i compute the centroid (mean of all points in C_i over d dimensions)

K-Means Algorithm

1. Randomly assign a cluster number i to every point (where $i \in \{1, \ldots n\}$)

   2.1 For each cluster C_i compute the centroid (mean of all points in C_i over d dimensions)
   2.2 Assign each observation to the cluster which is the closest.

K-Means Algorithm

1. Randomly assign a cluster number i to every point (where $i \in \{1, \ldots n\}$)
2. Iterate until convergence:
   2.1 For each cluster C_i compute the centroid (mean of all points in C_i over d dimensions)
   2.2 Assign each observation to the cluster which is the closest.

Working of K-Means Algorithm

Why does K-Means work?

$$\text{Let, } x_i \in R^d = \text{Centroid for} i^{th} \text{cluster}$$
$$= \frac{1}{|C_i|} \sum_{a \in C_i} x_a$$

Why does K-Means work?

$$\text{Let, } x_i \in R^d = \text{Centroid for} i^{th} \text{cluster}$$
$$= \frac{1}{|C_i|} \sum_{a \in C_i} x_a$$

Then,

$$WCV\left(C_i\right) = \frac{1}{|C_i|} \sum_{a \in C_i} \sum_{b \in C_i} ||x_a - x_b||_2^2$$
$$= 2 \sum_{a \in C_i} ||x_a - x_i||_2^2$$

Why does K-Means work?

$$\begin{aligned}\text{Let, } x_i \in R^d &= \text{Centroid for} i^{th} \text{cluster} \\ &= \frac{1}{|C_i|}\sum_{a \in C_i} x_a\end{aligned}$$

Then,

$$\begin{aligned}WCV\left(C_i\right) &= \frac{1}{|C_i|}\sum_{a \in C_i}\sum_{b \in C_i} ||x_a - x_b||_2^2 \\ &= 2\sum_{a \in C_i} ||x_a - x_i||_2^2\end{aligned}$$

K-Means gives the **local minima**.

Hierarchal Clustering

Hierarchal Clustering

Gives a clustering of all the clusters

Hierarchal Clustering

Gives a clustering of all the clusters

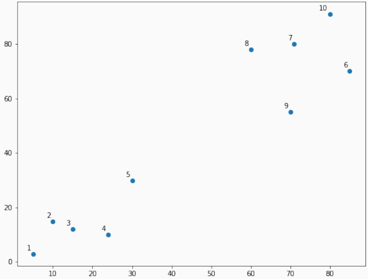
There is no need to specify K at the start

Hierarchal Clustering

Gives a clustering of all the clusters
There is no need to specify K at the start

Algorithm for Hierarchal Clustering

1. Start with all points in a single cluster

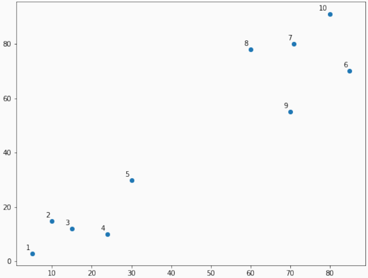


Example Dataset

Algorithm for Hierarchal Clustering

1. Start with all points in a single cluster

    2.1 Identify the 2 closest points

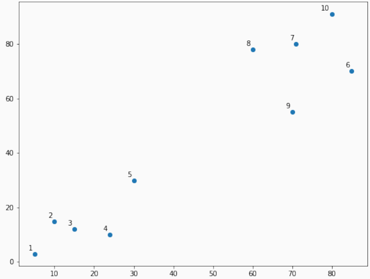


Example Dataset

Algorithm for Hierarchal Clustering

1. Start with all points in a single cluster

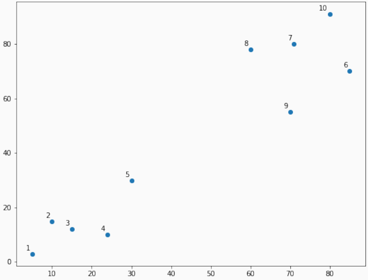
    2.1 Identify the 2 closest points
    2.2 Merge them



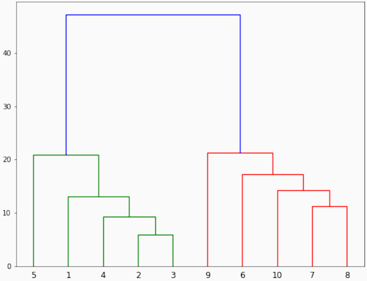
Example Dataset

Algorithm for Hierarchal Clustering

1. Start with all points in a single cluster
2. Repeat until all points are in a single cluster
    2.1 Identify the 2 closest points
    2.2 Merge them



Example Dataset



Final Clustering

Joining Clusters/Linkages

Complete
Max inter-cluster similarity

Single
Min inter-cluster similarity

Centroid
Dissimilarity between cluster centroids

More Code

Google Colab Link