ML@LSE - Bootcamp 4: Unsupervised learning

Ref. for today: Chapter 10, Hastie et al.

Unsupervised learning

A- Principal Component Analysis (PCA)

1. Intuition of PCA

a. How to represent a galaxy in 2D?
b. PCA idea: maximize the variance

2. PCA algorithm

a. Finding the components
b. Proportion of variance explained

B- K-means Clustering

1. The intuition of K-MC

c. Why clustering?
d. K-MC finds partitions our data

2. K-MC algoritm

a. Within cluster variation
b. K-MC in practice

C- Hierarchical Clustering

1. The intuition of HC

c. A more flexible approach
d. The idea behind: agglomerative clustering

2. HC algorithm

a. Summary of the algorithm
b. Dendrograms

So far we've studied supervised learning.

Each time we had a set of inputs (features) and we wanted to predict an output.

Today we're going to study a couple of unsupervised learning techniques. In unsupervised learning, we only have features and no output variable. We want to make sense of the set of features: finding useful way to represent the data or finding patterns in the data.

PCA intuition

A- Principal Component Analysis (PCA)

1. **Intuition of PCA**

a. **How to represent a galaxy in 2D?**
b. PCA idea: maximize the variance

2. **PCA algorithm**

a. Finding the components
b. Proportion of variance explained

B- K-means Clustering

1. **The intuition of K-MC**

c. Why clustering?
d. K-MC finds partitions our data

2. **K-MC algoritm**

a. Within cluster variation
b. K-MC in practice

C- Hierarchical Clustering

1. **The intuition of HC**

c. A more flexible approach
d. The idea behind: agglomerative clustering

2. **HC algorithm**

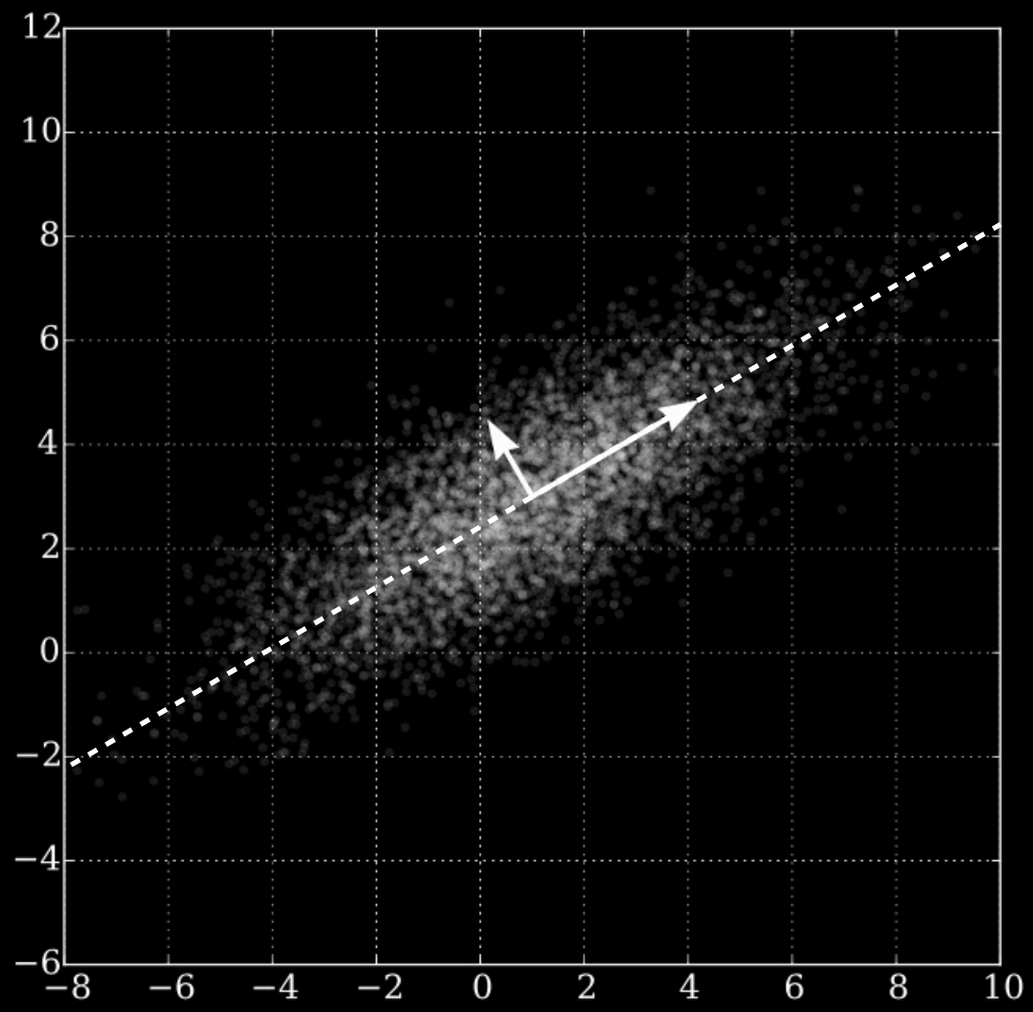
a. Summary of the algorithm
b. Dendrograms

How to represent a galaxy in two dimensions?

PCA finds the plane that maximizes the variance

The goal of PCA is to represent datasets on a lower dimension while preserving a maximum of information.

This is done by projecting the data onto the plane that maximizes the variance of the data.



Finding the components

The goal of PCA is to represent datasets on a lower dimension while preserving a maximum of information.

Let's look at an animation to understand how to find the components.

1) Center the data around its mean

2) Create a unit vector $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^T, \sum_{j=1}^{p} \lambda_j^2 = 1$

the $\lambda_j s$ are called the loadings of the first principal component.

3) Project all the datapoints onto the plane formed by the loading vector by taking the inner product. (As $\boldsymbol{\lambda}$ has a norm of one, the inner profuct yields the norm of the projected data).

$$\boldsymbol{\lambda} \cdot \boldsymbol{x} = \sum_{j=1}^{p} \lambda_j x_{ij}$$

4) Find the loading vector the maximizes the variance of the projected data:

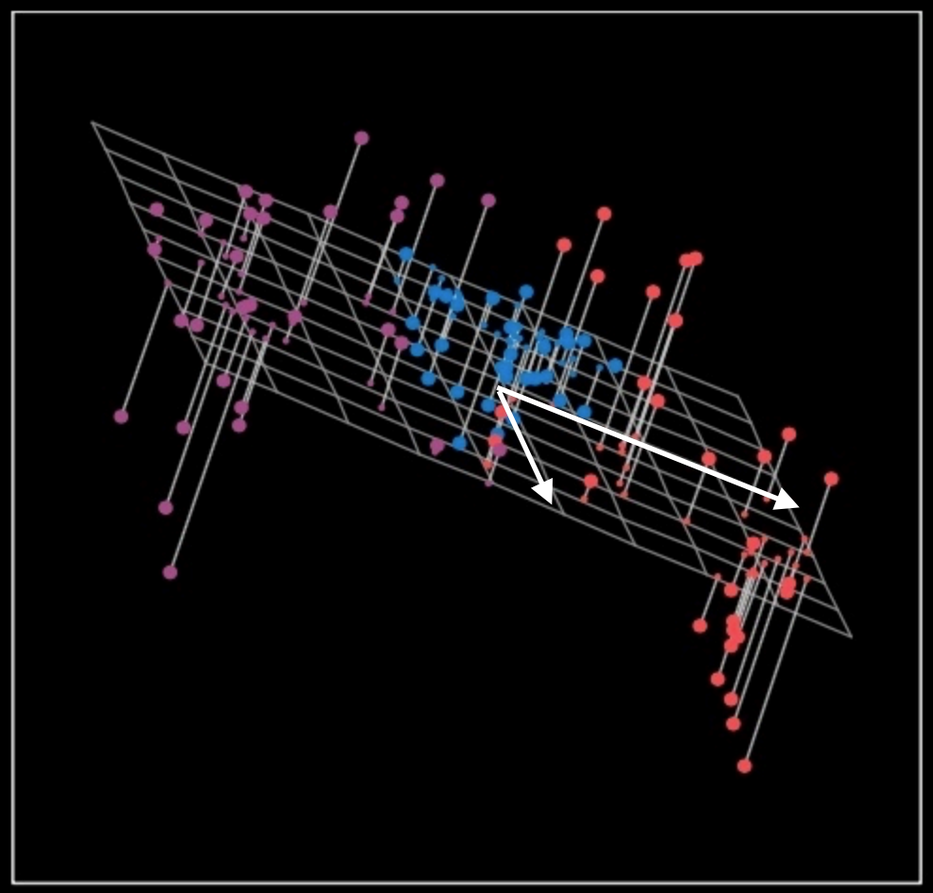
$$max_{\lambda_1,\ldots,\lambda_p} \sum_{i=1}^{n} (\sum_{j=1}^{p} \lambda_j x_{ij})^2 \quad s.t. \sum_{j=1}^{p} \lambda_j^2 = 1$$

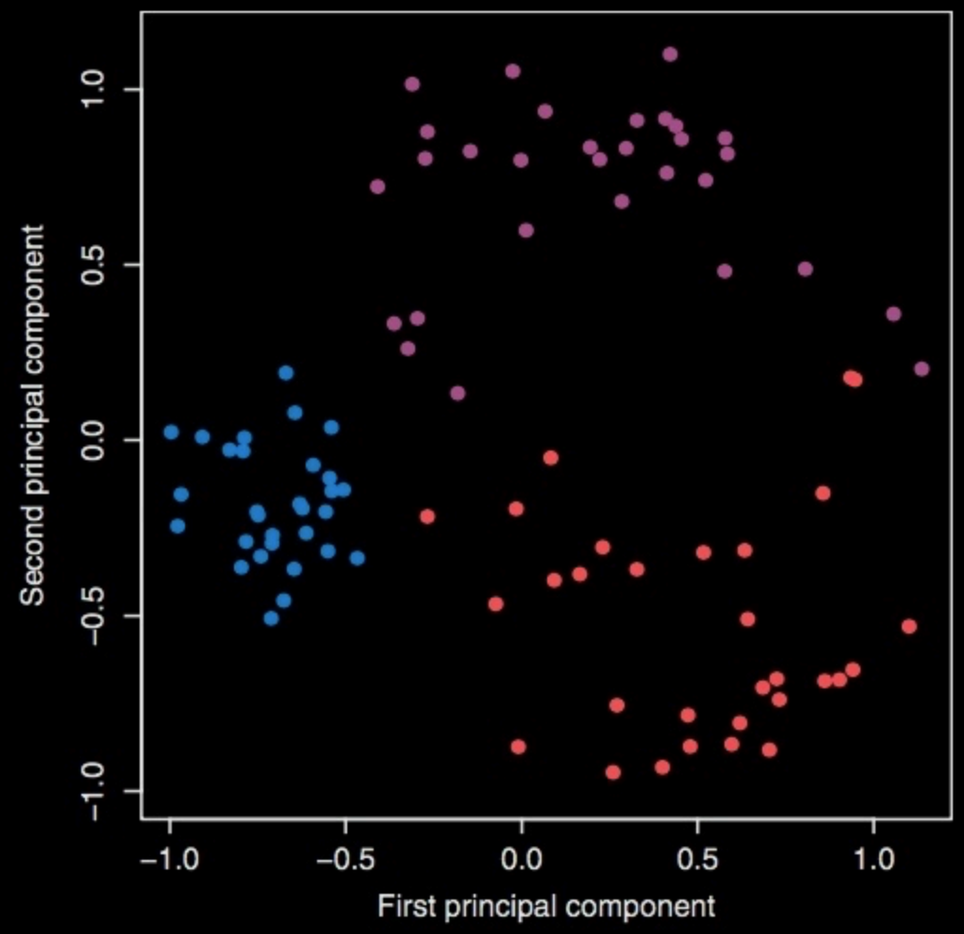
This optimal loading vector is called the First Principal Component, Z_1

← For reference

Finding the components

The second component Z_2 has to be orthogonal to the first component Z_1, the third component Z_3 has to orthogonal to both Z_1 and Z_2 etc…





Proportion of variance explained

We now have a set of principal components $Z_1, \ldots, Z_p$.

Question: How much information do we loose by projecting our data onto the first few Principal Components?
Similarly, how much of the total variance is not contained into the first few principal components?

Proportion of variance explained (PVE):

$$\frac{\text{Variance explained by the mth principal component}}{\text{Total variance}}$$

Proportion of variance explained

Total variance:

Variance = average of the squared deviations

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x_i}-\bar{\mathbf{x}}\|^2$$

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x_i}\|^2$$

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{(\sum_{j=1}^{p}x_{ij}^2)}^2$$

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}x_{ij}^2$$

Proportion of variance explained

Proportion of variance explained (PVE):

$$\frac{\text{Variance explained by the mth principal component}}{\text{Total variance}}$$

$$\text{PVE} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\lambda_j x_{ij}\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}x_{ij}^2}$$

Proportion of variance explained: scree plots

A- Principal Component Analysis (PCA)

1. Intuition of PCA

a. How to represent a galaxy in 2D?
b. PCA idea: maximize the variance

2. PCA algorithm

a. Finding the components
b. Proportion of variance explained

B- K-means Clustering

1. The intuition of K-MC

c. Why clustering?
d. K-MC finds partitions our data

2. K-MC algoritm

a. Within cluster variation
b. K-MC in practice

C- Hierarchical Clustering

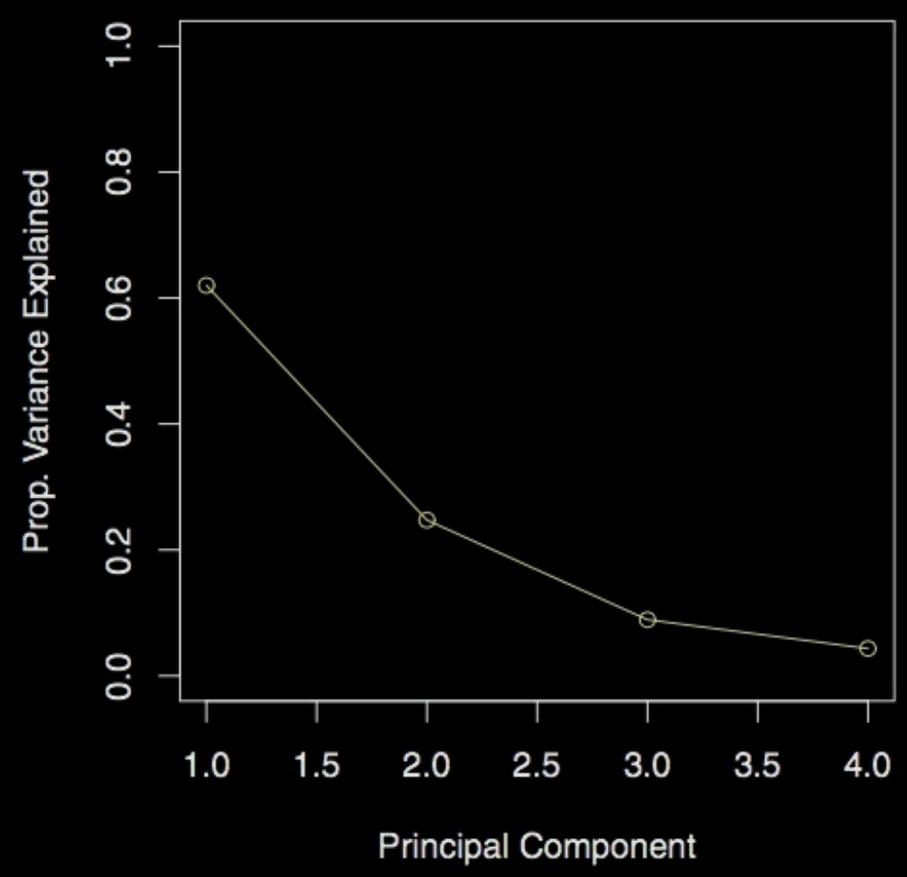
1. The intuition of HC

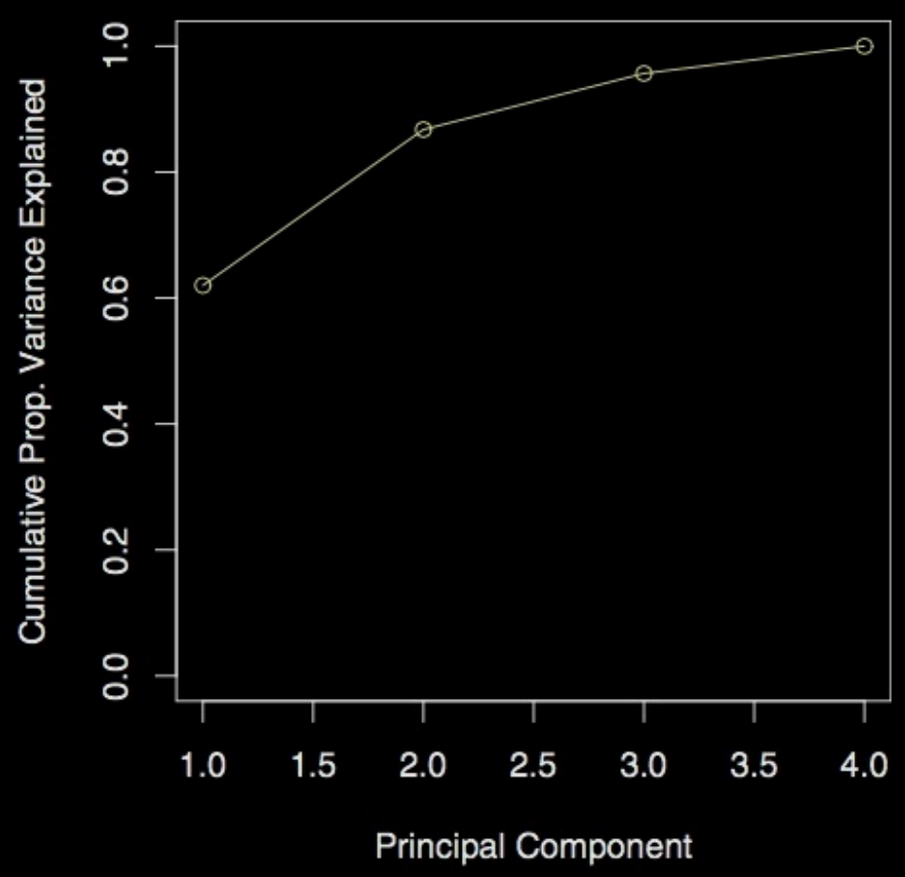
c. A more flexible approach
d. The idea behind: agglomerative clustering

2. HC algorithm

a. Summary of the algorithm
b. Dendrograms

$$\text{PVE} = \frac{\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\lambda_j x_{ij}\right)^2}{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}x_{ij}^2}$$





Clustering

A- Principal Component Analysis (PCA)

1. Intuition of PCA

a. How to represent a galaxy in 2D?
b. PCA idea: maximize the variance

2. PCA algorithm

a. Finding the components
b. Proportion of variance explained

B- K-means Clustering

1. The intuition of K-MC

c. Why clustering?
d. K-MC finds partitions our data

2. K-MC algoritm

a. Within cluster variation
b. K-MC in practice

C- Hierarchical Clustering

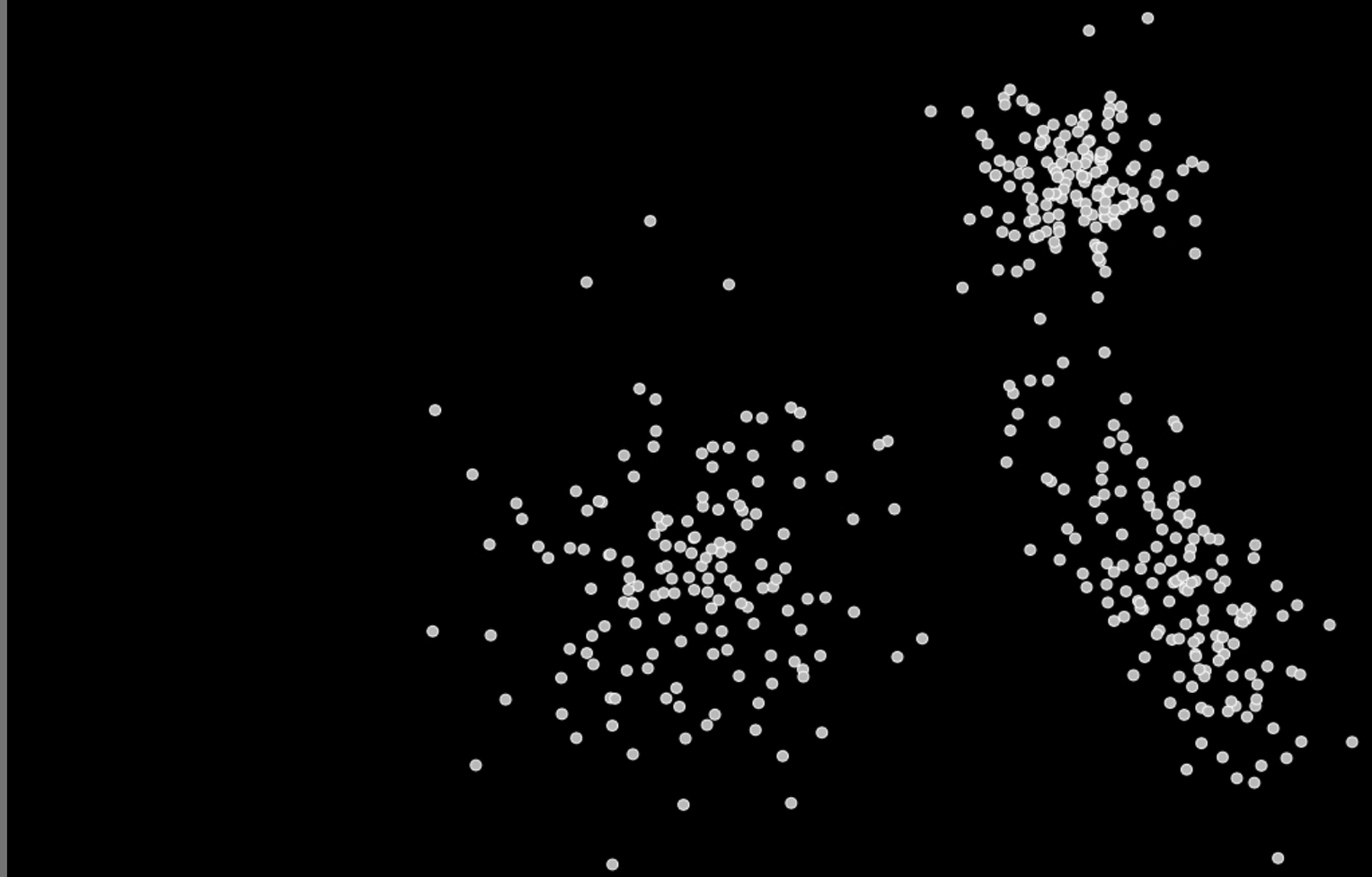
1. The intuition of HC

c. A more flexible approach
d. The idea behind: agglomerative clustering

2. HC algorithm

a. Summary of the algorithm
b. Dendrograms

PCA allowed us to reduce the dimension of our dataset without losing too much information.
We might as well find patterns in our dataset. For example, we may want to identify subgroups in the data, or clusters: this is called clustering.



Clusters partition our data

First clustering method: K-mean clustering

How does it work?

- First we decide the number of clusters K we want to create
- We want to put each of our datapoints into a cluster, hence the K clusters must not overlap and each datapoint must belong to one cluster.
- We call our clusters C_1, …, C_K

Clusters partition our data

First clustering method: K-mean clustering

How does it work?

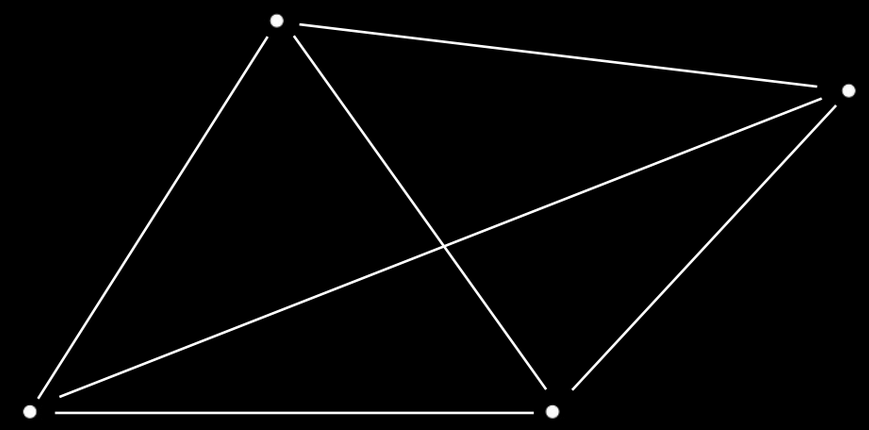
- First we decide the number of clusters K we want to create
- We want to put each of our datapoints into a cluster, hence the K clusters must not overlap and each datapoint must belong to one cluster.
- We call our clusters C_1, …, C_K

K-mean clustering main idea: we want the observations in one cluster to be similar, i.e. we want the within-cluster variation to be as small as possible!

Within-Cluster variation

Now, what measure would you suggest to evaluate within-cluster variation?

We use the average square distances between each pair of observations in the cluster:



$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \|\mathbf{x}_i - \mathbf{x}_i'\|^2 = \frac{1}{|C_k|} \sum_{i,i' \in C_k} (\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2)$$

Ideally, we want to solve the program:

$$min_{C_1,...C_K} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} (\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2)$$

K-MC Algorithm

Unfortunately, the program

$$min_{C_1,...C_K} \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} (\sum_{j=1}^{p} (x_{ij} - x_{i'j})^2)$$

is difficult to solve since there are about Kn ways to combine the observations into K clusters.

BUT the following algorithm provides a locally optimal solution:
(locally optimal means that even if the solution is not the best of all solutions, it's a pretty good solution)

K-Mean algorithm:

1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.

2. Iterate until the cluster assignments stop changing:

   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.

   (b) Assign each observation to the cluster whose centroid is closest

K-MC Algorithm

A- Principal Component Analysis (PCA)

1. Intuition of PCA

a. How to represent a galaxy in 2D?
b. PCA idea: maximize the variance

2. PCA algorithm

a. Finding the components
b. Proportion of variance explained

B- K-means Clustering

1. The intuition of K-MC

c. Why clustering?
d. K-MC finds partitions our data

2. K-MC algoritm

a. Within cluster variation
b. K-MC in practice

C- Hierarchical Clustering

1. The intuition of HC

c. A more flexible approach
d. The idea behind: agglomerative clustering

2. HC algorithm

a. Summary of the algorithm
b. Dendrograms

Hierarchical clustering

A- Principal Component Analysis (PCA)

1. Intuition of PCA

a. How to represent a galaxy in 2D?
b. PCA idea: maximize the variance

2. PCA algorithm

a. Finding the components
b. Proportion of variance explained

B- K-means Clustering

1. The intuition of K-MC

c. Why clustering?
d. K-MC finds partitions our data

2. K-MC algoritm

a. Within cluster variation
b. K-MC in practice

C- Hierarchical Clustering

1. The intuition of HC

c. A more flexible approach
d. The idea behind: agglomerative clustering

2. HC algorithm

a. Summary of the algorithm
b. Dendrograms

Problem with K-mean clustering: it requires us to pre-specify some structure on the data (by choosing the number of clusters K).

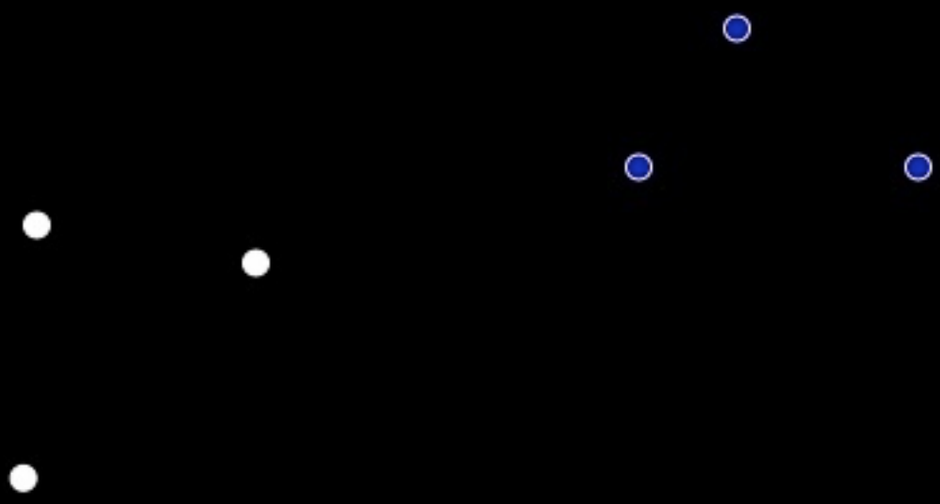
Hierarchical clustering is another clustering method that doesn't require this specification, and it is therefore more general.

Agglomerative clustering

The idea of hierarchical clustering is to progressively build up clusters of observations.

To do this we need an important concept: the dissimilarity between two clusters.

Minimal Intercluster Dissimilarity: Compute all pairwise distances between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities.



The Algorithm

A- Principal Component Analysis (PCA)

1. Intuition of PCA

a. How to represent a galaxy in 2D?
b. PCA idea: maximize the variance

2. PCA algorithm

a. Finding the components
b. Proportion of variance explained

B- K-means Clustering

1. The intuition of K-MC

c. Why clustering?
d. K-MC finds partitions our data

2. K-MC algoritm

a. Within cluster variation
b. K-MC in practice

C- Hierarchical Clustering

1. Preliminaries

c. A more flexible approach
d. Agglomerative clustering and dissimilarity between clusters

2. HC algorithm

a. Summary of the algorithm
b. Dendrograms

Let’s run the algorithm manually

Dendrogram

A- Principal Component Analysis (PCA)

1. Intuition of PCA

a. How to represent a galaxy in 2D?
b. PCA idea: maximize the variance

2. PCA algorithm

a. Finding the components
b. Proportion of variance explained

B- K-means Clustering

1. The intuition of K-MC

c. Why clustering?
d. K-MC finds partitions our data

2. K-MC algoritm

a. Within cluster variation
b. K-MC in practice

C- Hierarchical Clustering

1. Preliminaries

c. A more flexible approach
d. Agglomerative clustering and dissimilarity between clusters

2. HC algorithm

a. Summary of the algorithm
b. Dendrograms

H-Clustering allows for a useful representation: the Dendrogram