Knowledge and solutions for a changing world

Be boundless

Advancing data-intensive discovery in all fields

UNSUPERVISED MACHINE LEARNING

UW DIRECT

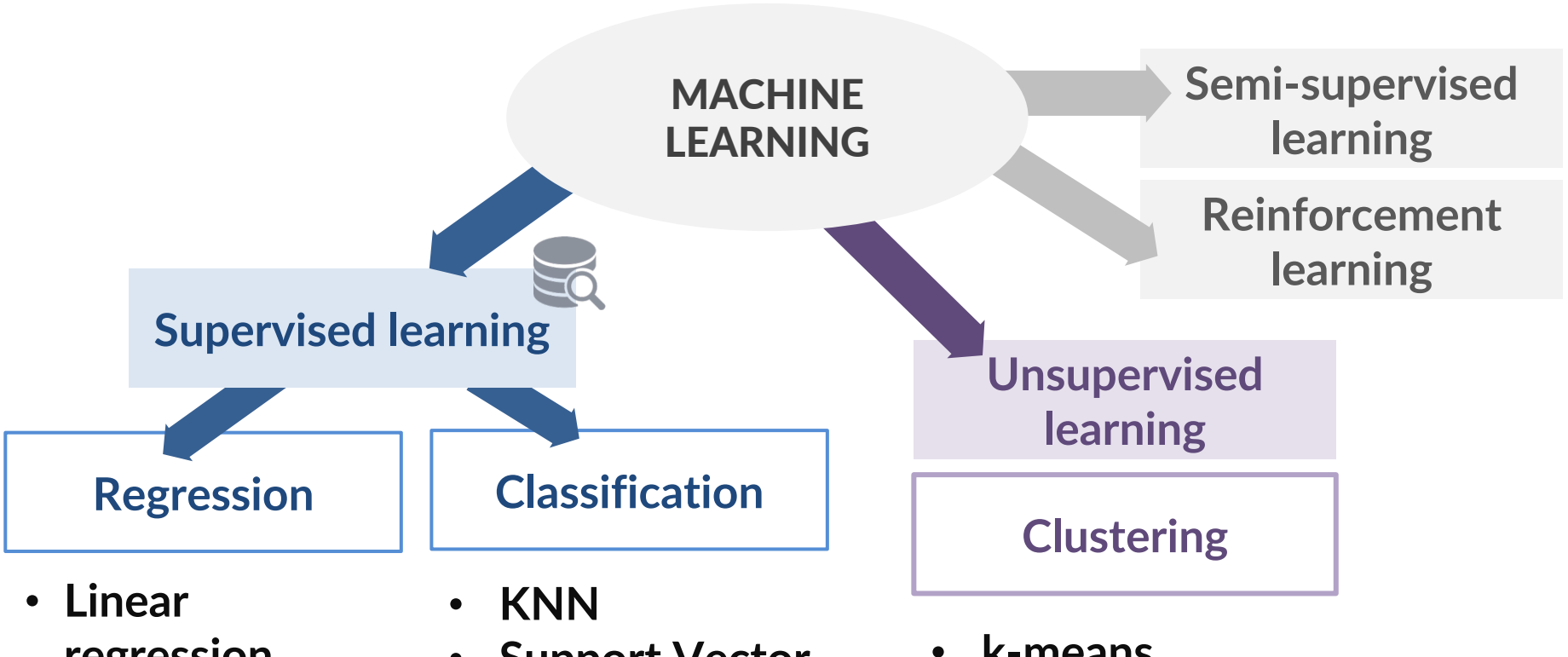
(Data Intensive Research Enabling Cutting-edge Tech)

https://uwdirect.github.io

Stéphanie Valleau

Chemical Engineering

Supervised vs Unsupervised



- Linear regression
- Neural network regression
- Decision tree regression

- KNN
- Support Vector Machines
- Decision trees

- k-means
- DBSCAN
- Hierarchical clustering

Supervised vs Unsupervised

***X* input feature/s – independent variables**
e.g. pictures, numbers
(weight / shape etc.)

SUPERVISED
ML models

PREDICT TARGET

target *Y* output
response / dependent variable
e.g. cost of fruit

***Unlabeled* *X* input feature/s**
e.g. pictures, numbers
(weight / shape etc.)

UNSUPERVISED
ML models

FIND PATTERNS

No predefined target Y !

e.g. find **clusters** in input data

Unsupervised learning

Main goal

Use a large set of features (X) [there are no longer any responses, Y!] and determine how the data may be grouped together

Central challenge in unsupervised learning

How to validate the data?

Two types of unsupervised learning

- **Principal components analysis (PCA) -** group the data in reduced dimensionality so that sub-groups describe most of the variance

- **Clustering -** find similar sub-groups of data within our total data set

Onto Jupyter!

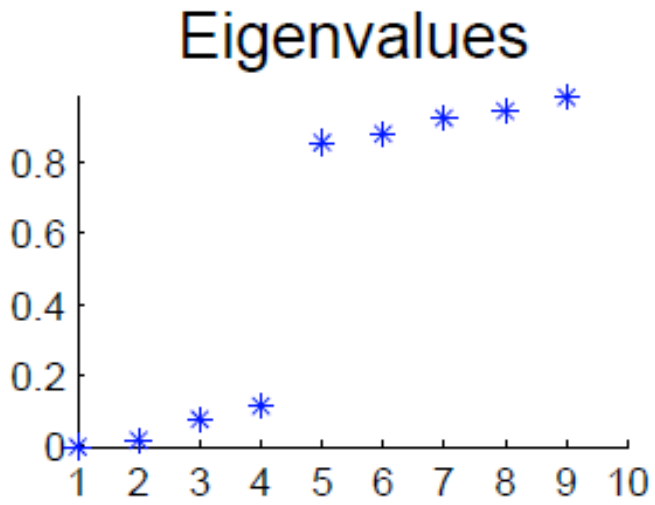
Read the following slides to get more information on PCA and K-Means

For the lecture we will look at these methods from the Jupyter notebook: `L13_PCA_Kmeans.ipynb`

PCA summary

There are many principal components, they are usually determined through eigen-decomposition of the covariance matrix of X

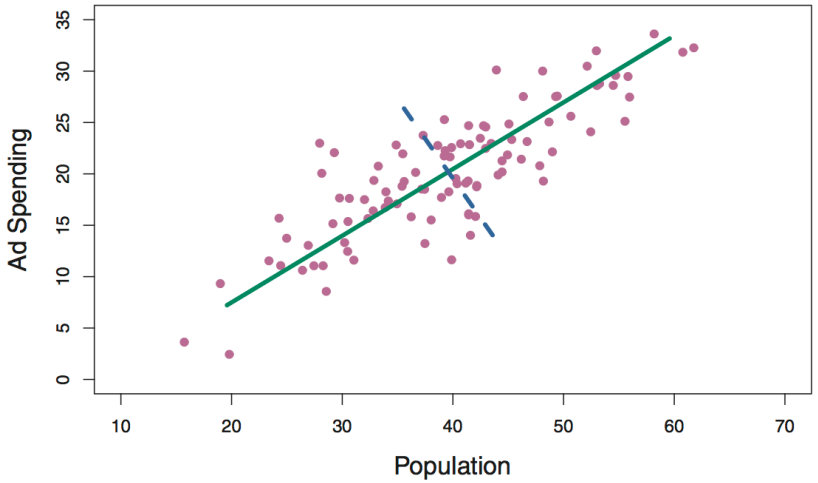
The eigenvalues, when ordered, often display a spectral gap, which can be useful in determining which set are the more useful to focus on



Random internet picture

PCA summary

When used for dimensionality reduction you always report the proportion of variance explained (PVE)



Why do you think PVE matters?

PCA summary

When used for dimensionality reduction you always report the proportion of variance explained (PVE)

Scree plots can be useful for this

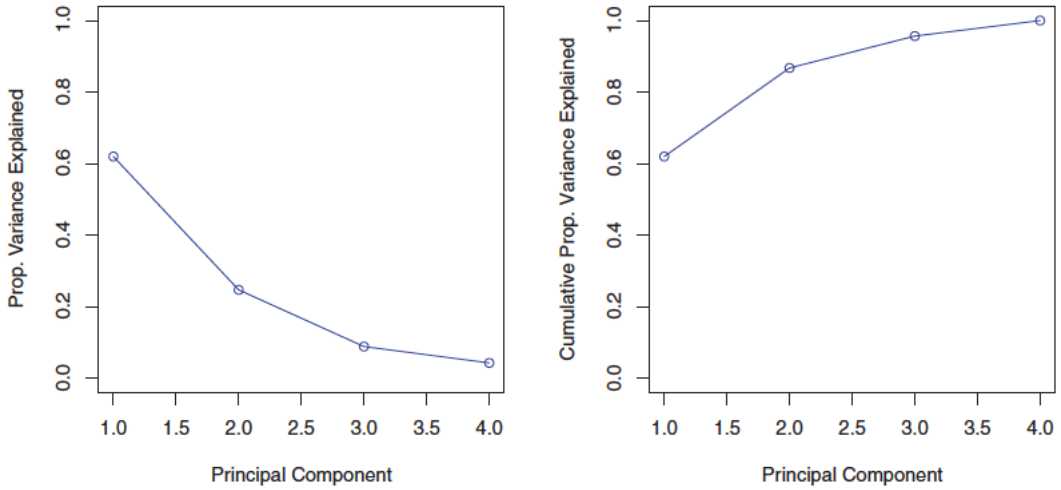
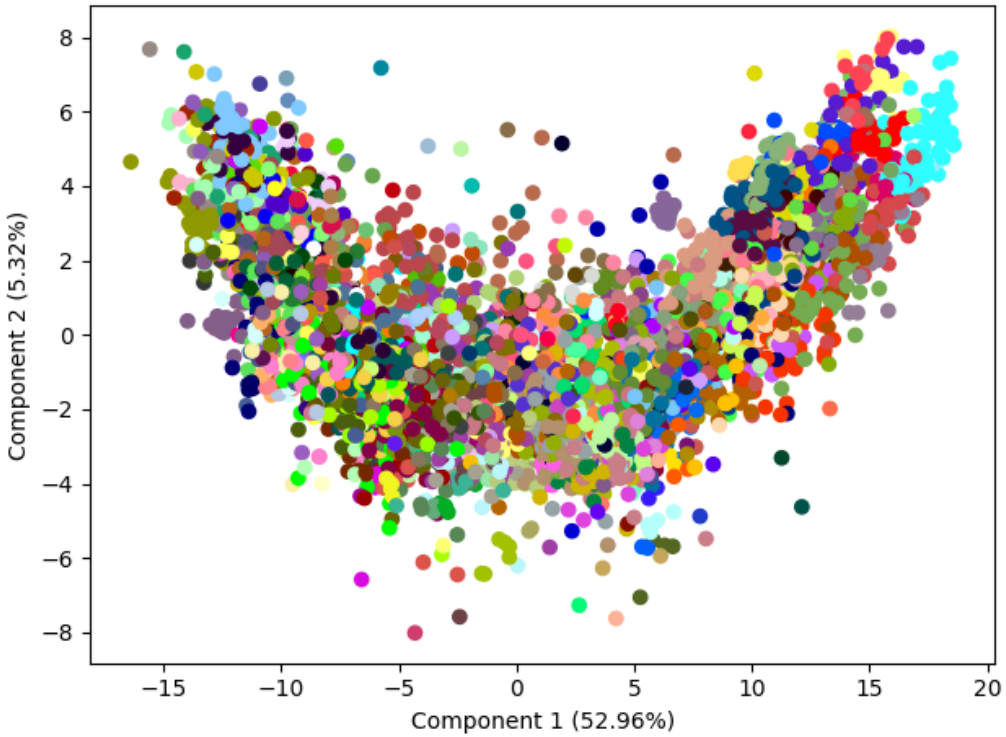


FIGURE 10.4. Left: *a scree plot depicting the proportion of variance explained by each of the four principal components in the* `USArrests` *data.* Right: *the cumulative proportion of variance explained by the four principal components in the* `USArrests` *data.*

PCA summary

When used for dimensionality reduction you always report the proportion of variance explained (PVE)

Or you can report it on the figure



PCA analysis results may depend on the scale of X_i!

As in Ridge and LASSO regression, the ultimate answer of your PCA analysis is not scale invariant!

Prior to conducting PCA you should:

1. Set all the means of each X_i equal to zero: transformation $\acute{X}_i = X_i - \bar{X}$
2. Set the variance of each X_i equal to one: transformation $\acute{X}_i = X_i/\sigma_i$

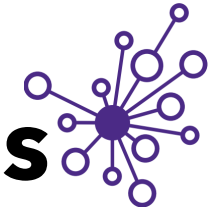
We are looking for variables that explain the variance and don't want the order of magnitude (or choice of units!) to numerically swamp out an important effect

How to use principal components

PCA is usually an “exploratory” method

Sometimes you may do PCA to reduce the dimensionality of your data then apply another clustering or regression method

Python implementation & tips



PCA can become expensive to calculate, especially as the data set size grows

Advice

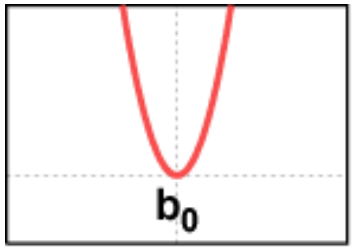
- Go slow and use a subset of your data (if you have many points)
- Use PCA as a guide and as an exploratory tool
- Constantly interrogate the results and ask if they make sense! You don't have "test set error" to fall back on, so you need to use your brain!

There are other types of dimensionality reduction

"Multidimensional scaling"

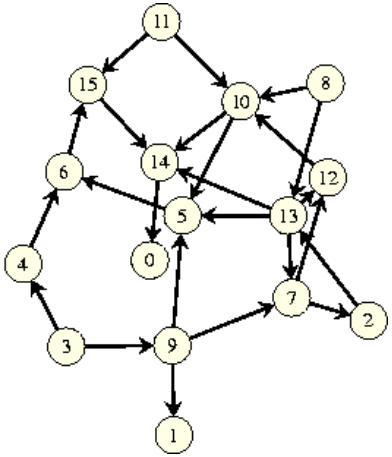
Spring embedding

- Compute distance in *p* dimensional space
- Use that as idea distance between two points in embedded space
- Optimize the system to minimize the RSS
  - E.g. simulated annealing



Bond

$$\sum_{i}^{bonds} K_{b,i}(b_i - b_{0,i})^2$$



Always want to report some metric of "strain"

Clustering

Our other main tool in unsupervised learning is clustering

Clustering seeks to group items by minimizing a distance metric between groups of observations or groups of features

K means

- Algorithm
- How to use it
- What the results mean
- Warnings: size K , many trials

Hierarchical clustering

Lots of other approaches // clustering vs PCA

K-means clustering, a simple algorithm

One of the most common clustering methods

Requires, as an input, specification of the final number of clusters you want (K)

Rules:

- Each observation must be placed in at least one of the clusters
- No clusters may overlap, each observation can only be placed in a single cluster
- The goal is to minimize the variance of observations within each of the clusters

Same data different values of K

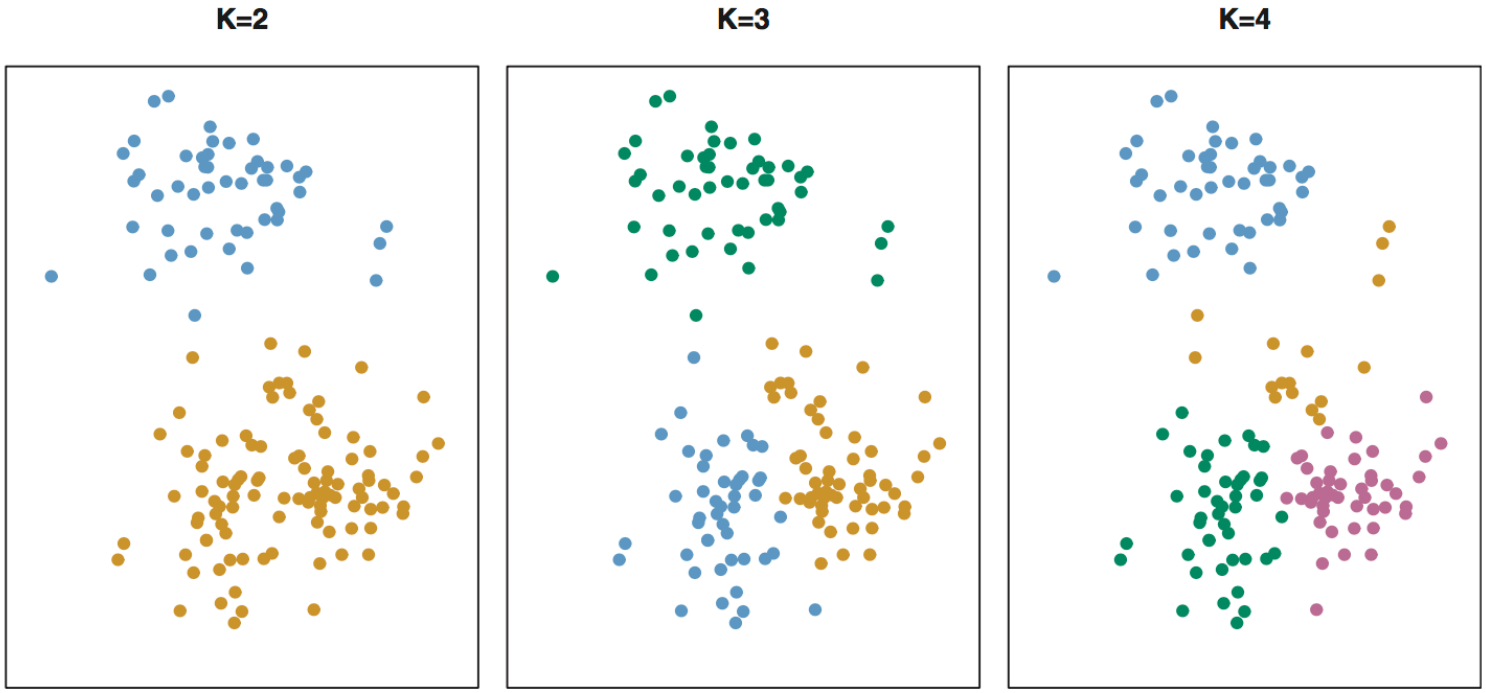


FIGURE 10.5. *A simulated data set with 150 observations in two-dimensional space. Panels show the results of applying K-means clustering with different values of K, the number of clusters. The color of each observation indicates the cluster to which it was assigned using the K-means clustering algorithm. Note that there is no ordering of the clusters, so the cluster coloring is arbitrary. These cluster labels were not used in clustering; instead, they are the outputs of the clustering procedure.*

Algorithm for K-means

Algorithm 10.1 *K-Means Clustering*

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 (where *closest* is defined using Euclidean distance).

$$\frac{1}{|C_k|}\sum_{i,i' \in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i \in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2, \qquad (10.12)$$

Algorithm for K-means (e.g., K=3)

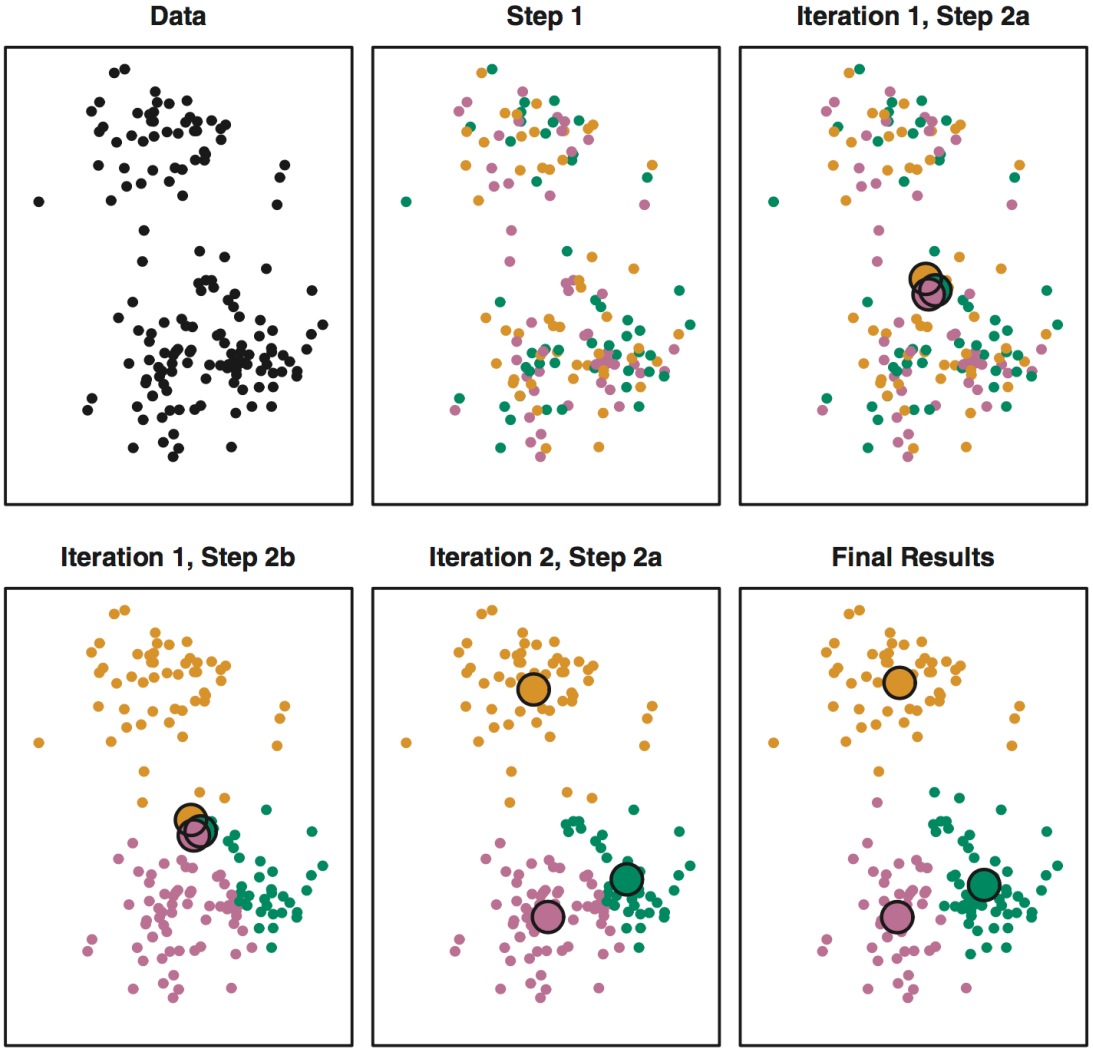


FIGURE 10.6. *The progress of the K-means algorithm on the example of Fig-*

Importance of sampling

K-means clustering is a stochastic process

- Initial random assignment of data to classes
- The optimization scheme leads to a local optimization , it is not guaranteed to find the global minimum

Process

- Re-seed different initial clusters and repeat optimization of cluster centers / assignments
- Monitor the sum of distances (each point's distance from each cluster center) as your error metric
- Choose clustering arrangement with lowest error

Importance of sampling

FIGURE 10.7. *K-means clustering performed six times on the data from Fig-*

What to do with your clusters

As in PCA, clustering is an **exploratory analysis tool**

If you have clustered your observations (most common): you can **interrogate the different clusters** to see if they have common features

For observations with many features, you can also cluster the features and look @ common observations…

In K-means, **you have to choose the number of clusters**: you must assess the degree to which this choice makes an impact on your scientific conclusions!

Performing K-means clustering

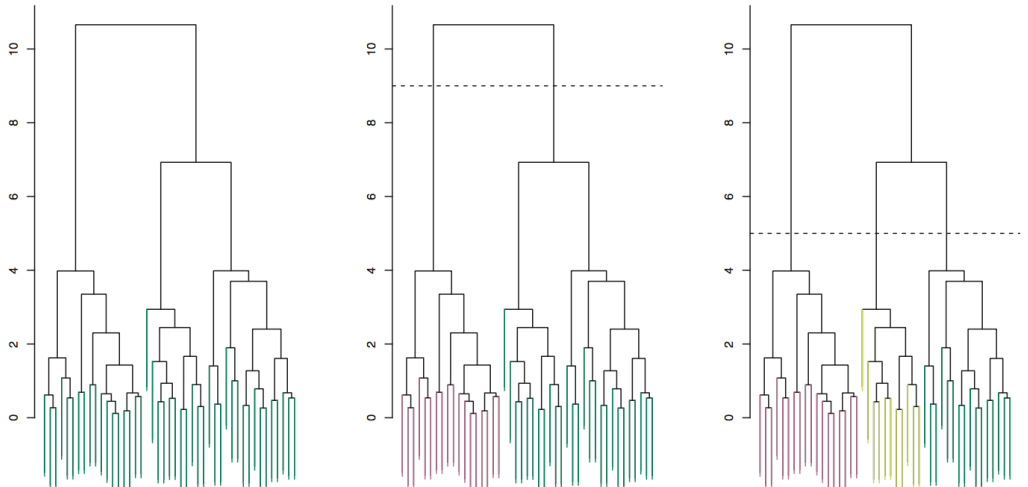
- The module `sklearn.clustering` can perform K-means and has many variants
- Take care as with PCA if the size of your data set grows, the computational cost to complete the clustering can become prohibitive…

Hierarchical clustering vs K-means

The main limitation of K-means,

Section 10.3.2 discusses hierarchical clustering, an approach that clusters all the data using a tree type structure

- Choice of how many clusters can be made after the clustering is completed…



Clustering vs PCA

Simple definition in ISL (p385):

> "*PCA looks to find a low-dimensional representation of the observations that explain a good fraction of the variance*"

> "Clustering looks to find homogeneous subgroups among the observations"

Unsupervised learning, especially use of results, can be a bit of an art…

- In some cases, it might be appropriate to look at both approaches