Lecture 20: Unsupervised learning I – Clustering

Isabel Valera

Machine Learning Group
Department of Mathematics and Computer Science
Saarland University, Saarbrücken, Germany

28.06.2021

Outline

Main references

- Bishop - Chapter 9 & 12
- EML - Chapter 14

Outline

Unsupervised learning

Unsupervised Learning:
Given a set of input points $(X_i)_{i=1}^n$:

- **Clustering:** Construction of a grouping of the points into sets of *similar* points, the so called *clusters*. (**Today!**)
- **Density Estimation:** Estimation of the distribution of the input points over the input space $\mathcal{X}$. Related to *outlier detection*. (**Partially today!**)
- **Dimensionality Reduction:** Construction of a mapping $\phi : \mathcal{X} \to \mathbb{R}^m$, where the dimensionality m of the target space is usually much smaller than that of the input space $\mathcal{X}$. Generally, the mapping should preserve properties of the input space $\mathcal{X}$, e.g., distances.

Clustering

Clustering approaches aim at grouping of the observed data into sets of *similar* points, the so called *clusters*. In general,

- there is not a broadly accepted objective for clustering – without specifying a suitable objective, clustering is ill-defined,
- clustering objective depends usually on application,
- in clustering the modeling aspect is even more important than in supervised learning – thus, do not use a clustering method if you do not understand what the objective implies!

Outline

K-means

K-means clustering is a prototype based clustering approach:

- **Goal:** find prototypes (**centroids**) $\boldsymbol{\mu}_i$, $i = 1, \ldots, k$ which represent the data in an optimal way (what does that mean?),
- **Objective:** denote by C_i the i-th cluster (set of points) which is represented by the prototype $\boldsymbol{\mu}_i$,

$$\underset{(C_1,\boldsymbol{\mu}_1),\ldots,(C_k,\boldsymbol{\mu}_k)}{\arg\min} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in C_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2,$$

  where $\|\cdot\|$ is the Euclidean norm,
- **True Goal:**
  1. finds sphere-like clusters in the data,
  2. heavily influenced by outliers,
  3. non-sphere like clusters are hard to fit.

K-means algorithm

K-means clustering:

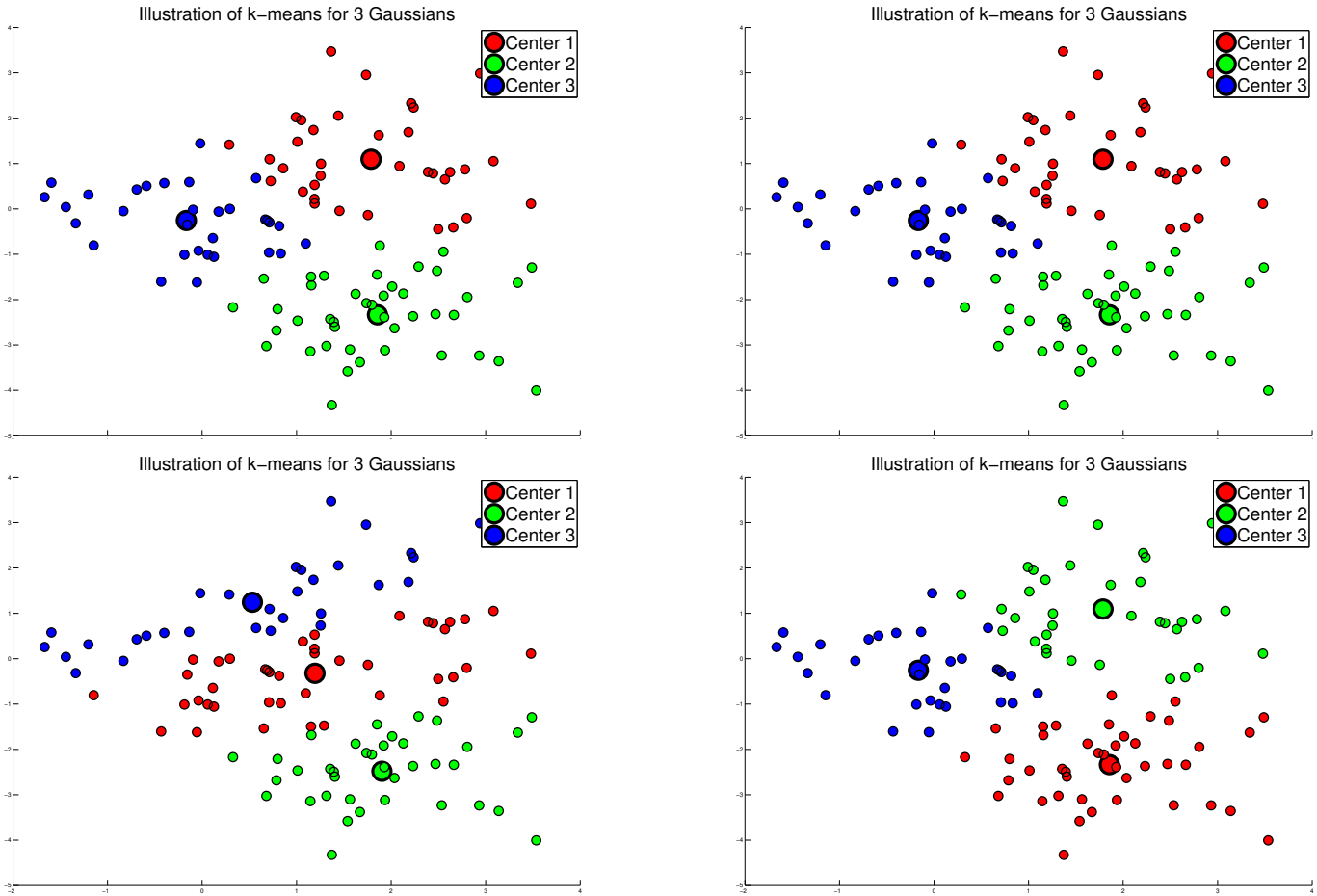
- k-means is combinatorial optimization problem,
- simple iterative algorithm – converges fast but finds only local minimum.

Lloyd's algorithm for k-means clustering:

1. initialize centers $\boldsymbol{\mu}_i$,
2. **while** changing $\boldsymbol{\mu}_i$, $i = 1, \ldots, k$ (i.e, iterate until clusters stop changing),
   1. group all samples according to closest $\boldsymbol{\mu}_i$, $i = 1, \ldots, k$
   2. recompute $\boldsymbol{\mu}_i$ as the mean of the observations in cluster C_i for $i = 1, \ldots, k$
3. **return** $\mu_1, \ldots, \mu_k$,

Steps are optimal for fixed centroids

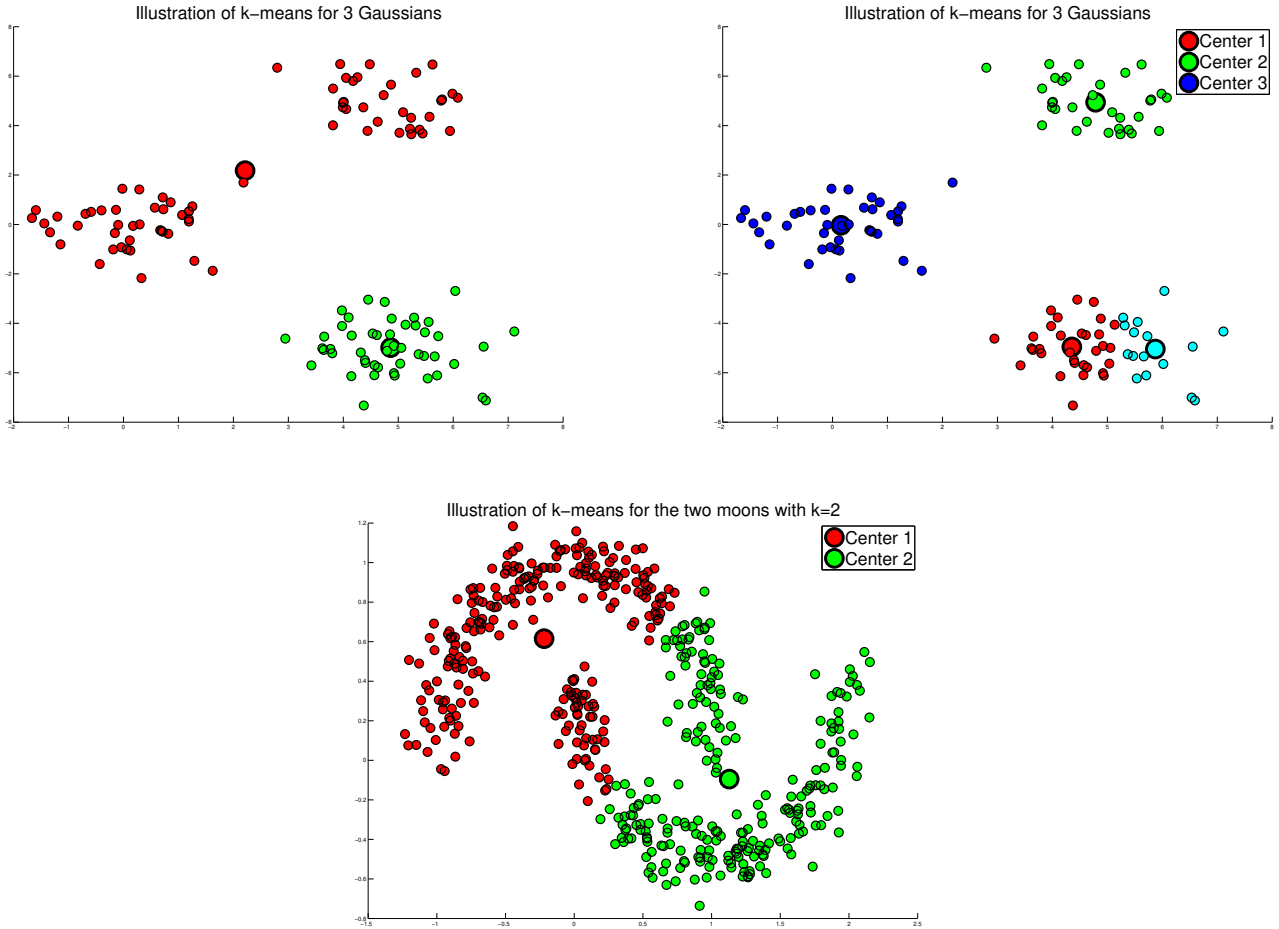
Example I



(Image by prof. Hein).

Four independent runs for k-means algorithm for the same dataset

Example II



Left: k is chosen too small. Middle: k is chosen too large. Right: The two moons dataset - clusters are not of spherical shape (thus, decreasing k not useful to find true k). (Image by Prof. Hein)

Outline

Hierarchical clustering

Hierarchical clustering generates a hierarchical grouping of the n data points. Two approaches:

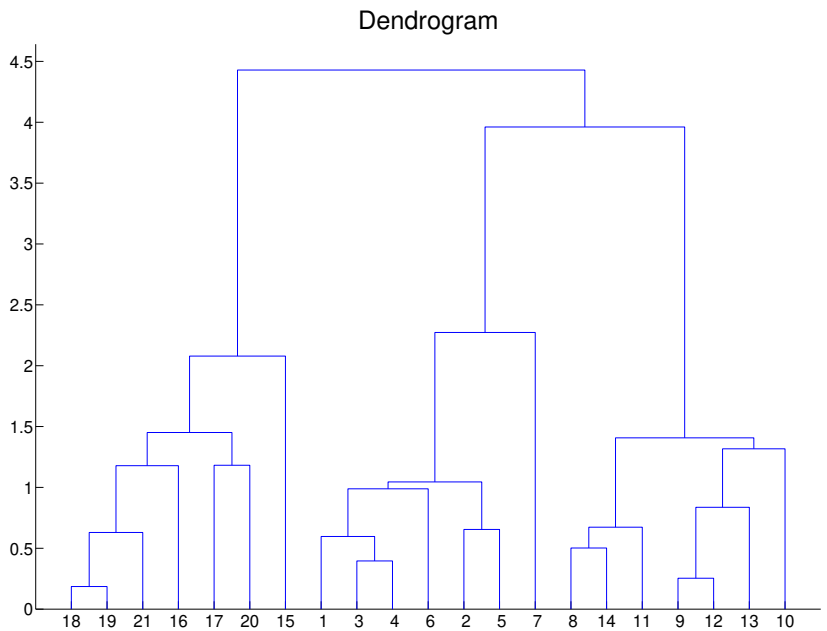
- **agglomerative:** start with all n points as individual clusters and consecutively join cluster which are *most similar*;
- **divisive:** start with one cluster containing all n points and consecutively divide the clusters so that they are *most dissimilar*.

Generates a tree structure on the data – the **dendrogram**.

Dendrogram

Definition

A **dendrogram** is a binary tree with a distinguished root, that has the data points as its leaves. The height where two clusters are merged is equal to their dissimilarity.



Agglomerative hierarchical clustering

Requirement: a distance measure between point sets.

Definition

A **dissimilarity measure** D **between finite subsets** of $\mathcal{X}$ is defined as $D : 2^{\mathcal{X}} \times 2^{\mathcal{X}} \to \mathbb{R}$ with

- $D(A, B) \geq 0$ for all $A, B \subseteq \mathcal{X}$,
- $D(A, B) = 0$ if and only if $A = B$,
- $D(A, B) = D(B, A)$.

Note: triangle inequality not required – not necessarily a metric.

Agglomerative hierarchical clustering II

Algorithm:

- *Input*: set of n points in $\mathcal{X}$, dissimilarity D between subsets of $\mathcal{X}$.
- *Initialize* with n clusters at level n, $C_1^{(n)}, \ldots, C_n^{(n)}$ with $C_i^{(n)} = \{\mathbf{x}_i\}$.
- while $l > 1$, do
  1. compute for all l clusters in $C_1^{(l)}, \ldots, C_l^{(l)}$ their dissimilarity $d_{ij} = D(C_i^{(l)}, C_j^{(l)})$
  2. merge the least dissimilar clusters, with indices $(r, s) = \underset{1 \leq i,j \leq l,\, i \neq j}{\arg\min}\ d_{ij}$.
  3. for $i \neq r$ and $i \neq s$, $C_i^{(l-1)} = C_i^{(l)}$ and $C_r^{(l-1)} = C_r^{(l)} \cup C_s^{(l)}$.
  4. height in the dendrogram of the merger between $C_r^{(l)}$ and $C^{(l)}$ is
  $$\alpha^{(l)} = d_{rs} = \min_{i,j} d_{ij}.$$
  5. relabel the clusters of level $l-1$ from 1 to $l-1$,
- *Output*: the sets of clusters $C^{(l)}$ for each level $l = 1, \ldots, n$.

Similarity between clusters

Agglomerative clustering: iteratively join *most similar* clusters.

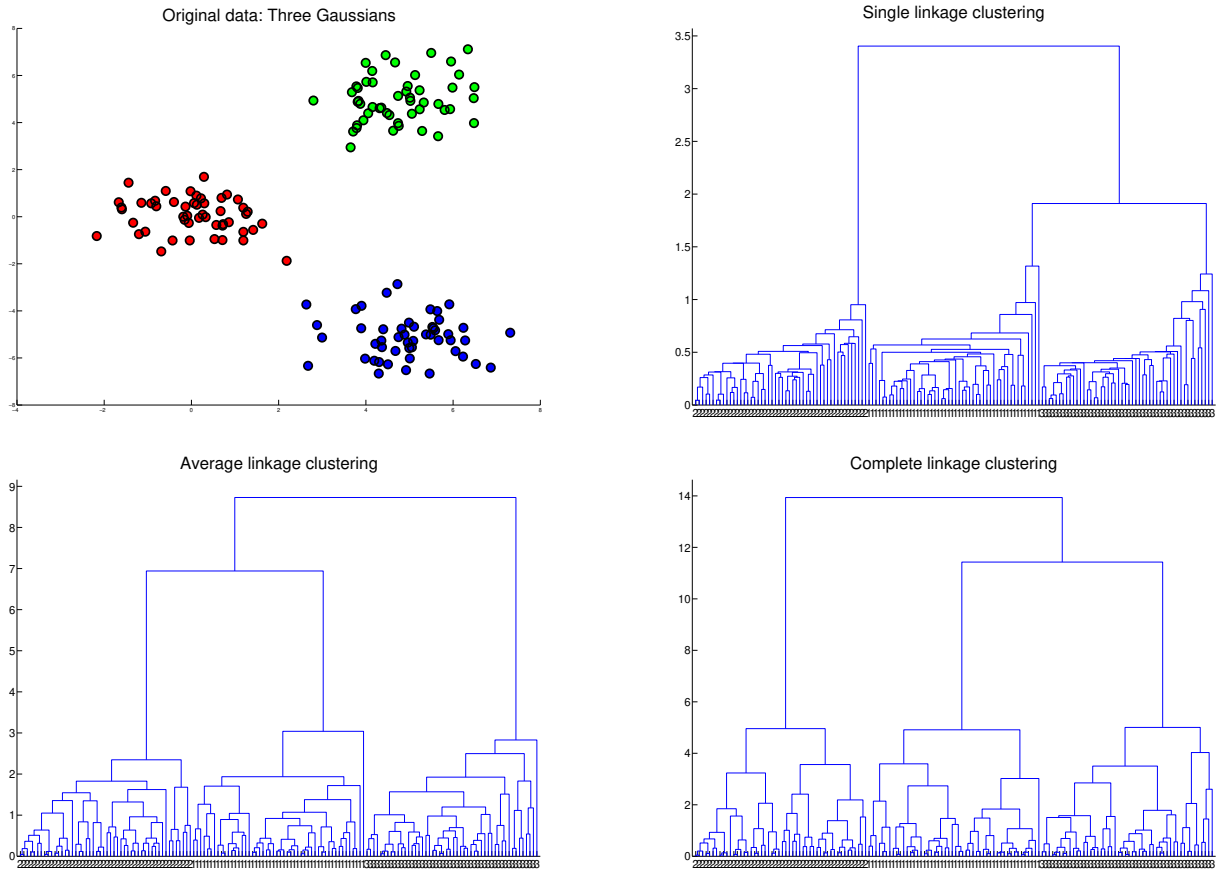
How to measure dissimilarity of clusters C_1 and C_2 ?

- **Single-linkage:** $d_{\min}(C_1, C_2) = \min_{i \in C_1, j \in C_2} d(\mathbf{x}_i, \mathbf{x}_j)$,
- **Complete-linkage:** $d_{\max}(C_1, C_2) = \max_{i \in C_1, j \in C_2} d(\mathbf{x}_i, \mathbf{x}_j)$,
- **Average-linkage:** $d_{\text{avg}}(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{i \in C_1, j \in C_2} d(\mathbf{x}_i, \mathbf{x}_j)$,

Two clusters are similar:

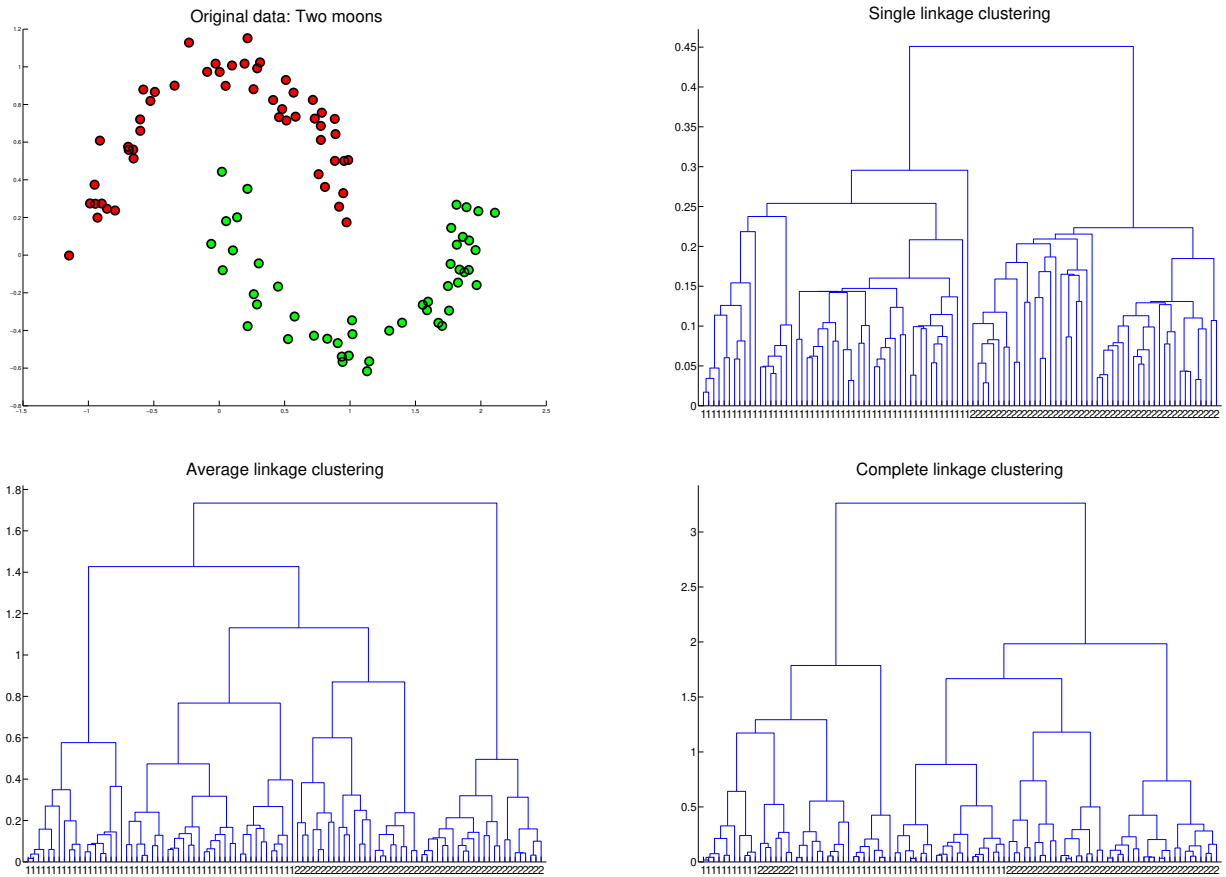
- single linkage: if for all points in each cluster there exists a path so that all points in the path are similar;
- complete-linkage: if all points for both clusters are similar; or
- average-linkage: if on average the points of both clusters are similar.

Example I



Top, left: dataset are three Gaussians. Top, right: single-linkage clustering. Bottom, left: average-linkage clustering. Bottom, right: complete-linkage clustering. (Image by Prof. Hein)

Example II



Top, left: Two moons dataset. Top, right: single-linkage clustering Bottom, left: average-linkage clustering. Bottom, right: complete-linkage clustering. (Image by Prof. Hein)

Problems of dendrograms

Problems of dendrograms

- **instability** small changes in the data can lead to huge changes in the dendrogram,
- **hierarchy:** multi-scale partitioning but different distance measures are hard to interpret.
- **dissimilarity:** the dissimilarity of clusters at which one joins clusters encodes their dissimilarity – comparing data using this distance is highly non-intuitive.

Overview

Pros:

- nice hierarchical representation of the data,
- single-linkage has a nice theoretical foundation,
- computationally relatively cheap.

Cons:

- single-linkage and complete very sensitive to data fluctuations,
- complete linkage has problems with non-spherical clusters,
- interpretation of the data requires profound understanding of the cluster similarity measures.

Outline

Density-based clustering

Probabilistic setting:

- sample $\{\mathbf{x}_i\}_{i=1}^n$ is drawn i.i.d. from probability measure in $\mathbb{R}^d$,
- the probability measure has a density in $\mathbb{R}^d$,

Clustering model: The simplest approach is to assume that the density p is a *Gaussian mixture*, i.e.,

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),$$

where π_k is teh prior probability of cluster k.

Main difference to approaches up to now is that we have clusters and "background noise".

Gaussian mixture model

- The Gaussian mixture model (GMM) assumes that the observed data is an i.i.d. sample from a Gaussian mixture distribution.
- Given observed data $\{\mathbf{x}_i\}_{i=1}^n$, we can get the ML (as well as the MAP) estimate the GMM parameters (i.e., $\{\pi_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}_{k=1}^K$) using the expectation-maximization (EM) algorithm.
- To perfom **model selection** (i.e., select number of components/clusters), one may rely to cross-validation using, e.g., *Bayessian Information Criterion (BIC)* as criteria.
- **More details in additional notes and Jupyter notebook.**

Observations

- Note that GMMs are an approach for **density estimation**, that in turn allows us to perform clustering. We, however, need to make the explicit assumption of the likelihood (which is equivalent to select the dismilarity measure for k-means or hierarchical clustering)
- Mixture models can be generalized beyond the Gaussian distribution to accommodate, e.g., categorical or binary data.
- EM algorithm is a general and powerful iterative method to find (local) maximum likelihood or maximum a posteriori (MAP) estimates of parameters in probabilistic models, where the model depends on unobserved latent variables.