Unsupervised Learning: Clustering

ML Instruction Team, Fall 2022

CE Department
Sharif University of Technology

Clustering: An Overview

- Clustering algorithms can be classified into different categories, based on the following criteria:
  - Whether each point is assigned to exactly one cluster or several clusters with certain probabilites that add up to 1:
    - **Hard**
    - **Soft**
  - Whether all clusters are on the same level or several clusters are built in a hierarchical way:
    - **Partitional**
    - **Hierarchical**

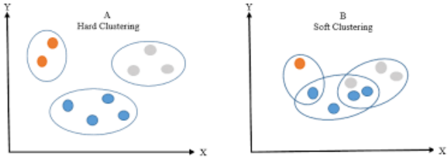

Figure: Hard vs Soft source.

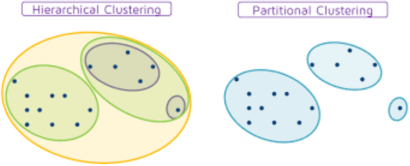

Figure: Partitional vs Heirarchical source.

Clustering: An Overview

- Hierarchical clustering is usually done in two different ways:
  - **Agglomerative**: This is a ”bottom-up” approach, Each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
  - **Divisive**: This is a ”top-down” approach, All observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

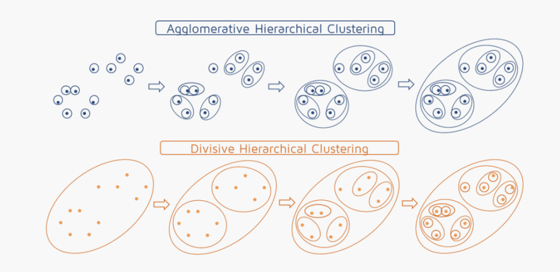

Figure: Agglomerative vs Divisive source.

Hard Partitional Clustering: K-Means

- A particularly simple method for clustering is K-means, The idea is **to represent each cluster k by a center point $\mathbf{c}_k$ and assign each data point $\mathbf{x}_n$ to one of the clusters k** which can be written in terms of index sets $\mathcal{C}_k$
- The center points and the assignment are then chosen such that the mean squared distance between data points and center points **is minimized**:

$$J := \sum_{n=1}^{N} \sum_{k=1}^{K} r_{nk} \|\mathbf{x}_n - \mathbf{c}_k\|^2$$

- Here we introduced a corresponding binary indicator variable $r_{nk} \in \{0, 1\}$ where $k = 1, 2, \ldots, K$ describing which of the K clusters the data point $\mathbf{x_n}$ is assigned to, so that if data point $\mathbf{x_n}$ is assigned to cluster k then $r_{nk} = 1$, and $r_{nj} = 0$ for $j \neq k$
- Now, Our goal is to find values for the $\{r_{nk}\}$ and the $\{\mathbf{c}_k\}$ so as to minimize J. we can do this through an **Iterative Procedure**

Hard Partitional Clustering: K-Means

- To minimize J through iterating, we have to do the following algorithm:
  1. **Initialize** $\mathbf{c}_k$ with **Random Value** for all $k = 1, 2, \dots, K$, It could be chosen from data values either.
  2. Minimize J with respect to r_{nk}, keeping the $\mathbf{c}_k$ fixed. because J is a linear function of r_{nk} this optmization can be performed easily to give a closed form solution:

$$r_{nk} = \begin{cases} 1 & \text{if } k = \mathtt{argmin}_j \|\mathbf{x}_n - \mathbf{c}_j\|^2 \\ 0 & \text{otherwise} \end{cases}$$

  3. Minimize J with respect to $\mathbf{c}_k$, keeping the r_{nk} fixed. if the assignment is fixed, it is easy to show that the optimal choice of the center positions is given by:

$$\mathbf{c}_k = \frac{\sum_n r_{nk}\mathbf{x}_n}{\sum_n r_{nk}}$$

  4. Check the convergence criteria, otherwise go to step 2.

Hard Partitional Clustering: K-Means

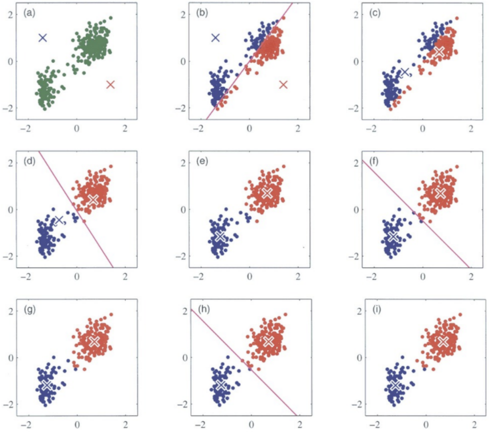

Figure: Illustration of K-Means with $K = 2$ [1].

Hard Partitional Clustering: K-Means

- Note that the result of the algorithm is **not necessarily a global optimum** of the objective function J
- It is therefore advisable to **run the algorithm several times** with different initial center locations and **pick the best result**.
- A drawback of this and many other clustering algorithms is that **the number of clusters is not determined**.
- One has to decide on a proper K in advance, or one simply runs the algorithm with several different K-values and picks the best according to some criterion.

Soft Partitional Clustering: Gaussian Mixture Model (GMM)

- The K-means algorithm is a very simple method with sharp boundaries between the clusters, and no particular characterization of the shape of individual clusters.
- In a more refined algorithm, one might want to model each cluster with a Gaussian, capturing the shape of the clusters.
- **This leads naturally to a probabilistic interpretation** of the data as a superposition of Gaussian probability distributions.

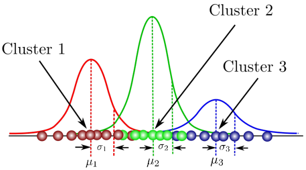

Figure: $1D$ Gaussian Mixture Model source.

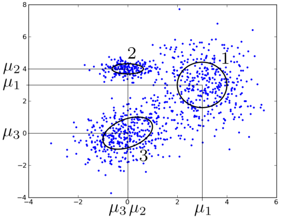

Figure: $2D$ Gaussian Mixture Model source.

Soft Partitional Clustering: Gaussian Mixture Model (GMM)

- Recall the probability, we assume that the probability density function (pdf) of cluster k can be written as:

$$\mathcal{N}(\mathbf{x} \mid \mathbf{c}_k, \Sigma_k) = \frac{1}{(2\pi)^{D/2}} \frac{1}{(\mathrm{det}(\Sigma_k))^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mathbf{c}_k)^T \Sigma_k^{-1} (\mathbf{x}-\mathbf{c}_k)\right)$$

- Here $\mathbf{c}_k, \Sigma_k$ are the mean and covariance matrix of the given k cluster respectively. There is also a prior probability $P(k) = \pi_k$ that a data point belongs to a particular cluster k. **The overall pdf for the data is** then given by the total probability:

$$p(\mathbf{x}) = \sum_{k=1}^{K} P(k) p(\mathbf{x} \mid k) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \mathbf{c}_k, \Sigma_k)$$

$$\text{where} \quad 0 \leq \pi_k \leq 1\,, \sum_{k=1}^{K} \pi_k = 1$$

Soft Partitional Clustering: Gaussian Mixture Model (GMM)

- The problem now is that we do not know the parameters of the model, i.e. the values of the centers $\{\mathbf{c}_k\}$ and the covariance matrices $\{\Sigma_k\}$ of the Gaussians and the probabilities $\{\pi_k\}$ for the clusters.
- The simple idea is to choose the parameters such, that the **probability density of the data is maximized**. In other words we want to choose the model such that the data becomes most probable. This is referred to as **Maximum Likelihood Estimation**
- We know that our data points were drawn indepenedently, assume that we put these $\{\mathbf{x}_n\}$ into the rows of the $\mathbf{X}_{n\times d}$, so as a result the likelihood function would be:

$$L\Big(\{(\mathbf{c}_k, \Sigma_k, \pi_k)\}\Big) = ln\Big(p(\mathbf{X} \mid \{(\mathbf{c}_k, \Sigma_k, \pi_k)\})\Big) = \sum_{n=1}^{N} ln\Big(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \mathbf{c}_k, \Sigma_k)\Big)$$

Soft Partitional Clustering: Gaussian Mixture Model (GMM)

- Unfortunately, the **Maximum Likelihood Estimation** has not a closed form solution. because the parameters on the left-hand side will occur implicitely also on the right-hand side.
- Beside of the lackness of a closed form solution, Maximum Likelihood Estimation would probably have singularity and identifiability problems.
- However, one can start with some initial parameter values and then **iterate** through these equations to improve the estimate.
- One can actually show that the likelihood increases with each iteration, if a change occurs. This iterative scheme is referred to as the **expectation-maximization algorithm**, or simply EM algorithm

Soft Partitional Clustering: Gaussian Mixture Model (GMM)

- To maximize Likelihood function through EM , we have to do the following algorithm:
  1. Initialize $\{\mathbf{c}_k\}$, $\{\Sigma_k\}$ and $\{\pi_k\}$ with Random Value and evaluate the initial value of log-likelihood.
  2. Evaluate the responsibilities using the current parameter values:

$$\gamma(z_{nk}) = \frac{\pi_k \mathcal{N}(\mathbf{x}_n|\mathbf{c}_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(\mathbf{x}_n|\mathbf{c}_j, \Sigma_j)}$$

  3. Re-Estimate the parameters using the current responsibilities:

$$\mathbf{c}_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk}) \mathbf{x}_n$$

$$\Sigma_k^{new} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(\mathbf{x}_n - \mathbf{c}_k^{new})(\mathbf{x}_n - \mathbf{c}_k^{new})^T$$

$$\pi_k^{new} = \frac{N_k}{N} \quad \text{where} \quad N_k = \sum_{n=1}^{N} \gamma(z_{nk})$$

  4. Evaluate the log-likelihood:

$$ln\left(p(\mathbf{X}|\{(\mathbf{c}_k, \Sigma_k, \pi_k)\})\right) = \sum_{n=1}^{N} ln\left(\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n|\mathbf{c}_k, \Sigma_k)\right)$$

  5. Check the convergence criteria, otherwise go to step 2.

Soft Partitional Clustering: Gaussian Mixture Model (GMM)

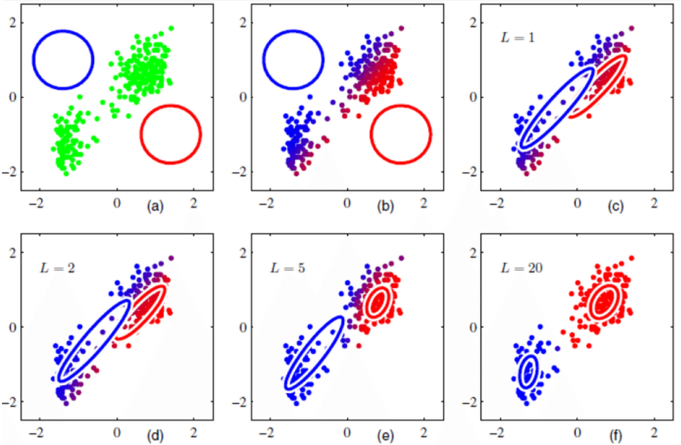

Figure: Illustration of EM using $K = 2$ [1].

References

- **[1]**. Bishop, Christopher. Pattern Recognition and Machine Learning (Information Science and Statistics). 2018.

Thank You!

Any Question?