Unsupervised Learning Methods 2022
Problem Set II –
Clustering

Due: 11.04.2022

Guidelines

- Answer all questions (PDF + Jupyter notebook).
- You must type your solution manual (handwriting is not allowed).
- Submission in pairs (use the forum if needed).
- You **may** submit the entire solution in a single ipynb file (or PDF + ipynb files).
- You **may** (and should) use the forums if you have any questions.
- Good luck!

1 K-Means

Objective
The K-Means objective is given by:

$$\arg \min_{\{\mathcal{D}_k\},\{\boldsymbol{\mu}_k\}} \sum_{k=1}^{K} \sum_{\boldsymbol{x}_i \in \mathcal{D}_k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2$$

1.1

Show that the following two objectives are equivalent to the K-Means objective:

1. As a sole function of the clusters:

$$\arg \min_{\{\mathcal{D}_k\}} \sum_{k=1}^{K} \frac{1}{|\mathcal{D}_k|} \sum_{\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathcal{D}_k} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2$$

2. As a sole function of the centroids:

$$\arg \min_{\{\boldsymbol{\mu}_k\}} \sum_{i=1}^{N} \min_{k} \|\boldsymbol{x}_i - \boldsymbol{\mu}_k\|_2^2$$

Solution:
1. Type your solution here...
2. Type your solution here...

1.2

Prove or disprove:
The K-Means algorithm **always** converges to a global minimum.

Solution:
Type your solution here...

1.3 K-Means implementation and Super-pixels

 Solve this section in the attached notebook. 

2 GMM

Gaussian random vector

- Let $\underline{X} \sim \mathcal{N}_d(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$ be a Gaussian random vector.
- Let $Y = \boldsymbol{a}^T \underline{X} + b$ be a random variable

2.1

Find $f_Y(y)$, the pdf of Y (as a function of $\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x, \boldsymbol{a}, b$).

Solution:
Type your solution here...

– –

Covariance

A matrix $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ is called Symmetric Positive Semi-Definite (SPSD) if $\boldsymbol{A}^T = \boldsymbol{A}$ and for any $\boldsymbol{v} \in \mathbb{R}^d$:

$$\boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} \geq 0$$

In other words:

$$\boldsymbol{A} \succeq 0 \iff \begin{cases} \boldsymbol{A}^T = \boldsymbol{A} \\ \boldsymbol{v}^T \boldsymbol{A} \boldsymbol{v} \geq 0 \quad \forall \boldsymbol{v} \end{cases}$$

Let $\underline{X}$ be a random vector with covariance $\boldsymbol{\Sigma}_x$.

2.2

Prove that $\boldsymbol{\Sigma}_x$ is an SPSD matrix.

Solution:
Type your solution here...

– –

2.3 GMM implementation

Solve this section in the attached notebook.

– –

3 Hierarchical Clustering

Complete-linkage

The complete-linkage distance between the two clusters $\mathcal{C}_1 = \{\boldsymbol{x}_i\}_{i=1}^{N_1}$ and $\mathcal{C}_2 = \{\boldsymbol{x}_j\}_{j=1}^{N_2}$:

$$d^2_{\text{complete-link}}(\mathcal{C}_1, \mathcal{C}_2) = \begin{cases} 0 & \mathcal{C}_1 = \mathcal{C}_2 \\ \max_{\boldsymbol{x}_i \in \mathcal{C}_1, \boldsymbol{x}_j \in \mathcal{C}_2} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2^2 & \text{else} \end{cases}$$

3.1

Prove that the complete-linkage is indeed a metric.

Solution:
Type your solution here...

Lance-Williams

The Lance-Williams update rule (see the full algorithm in the lecture notes):

$$D_{\widetilde{ij},k} \leftarrow \alpha_i D_{i,k} + \alpha_j D_{j,k} + \beta D_{i,j} + \gamma |D_{i,k} - D_{j,k}|$$

Consider the three clusters $\mathcal{C}_1, \mathcal{C}_2$ and $\mathcal{C}_3$ with

$$D_{i,j} = d_{\text{single-link}}(\mathcal{C}_i, \mathcal{C}_j)$$

3.2 (Bouns 4%)

Prove that

$$D_{\widetilde{12},3} = d_{\text{single-link}}(\mathcal{C}_1 \cup \mathcal{C}_2, \mathcal{C}_3)$$

In words, show that the Lance-Williams algorithm is correct for the single-linkage dissimilarity.

Solution:
Type your solution here...

---

4 DBSCAN

4.1 DBSCAN implementation

 Solve this section in the attached notebook.