Elements of Machine Learning, WS 2024/2025
Prof. Dr. Isabel Valera and Dr. Kavya Gupta
Assignment Sheet #4: *Unsupervised Learning*

UNIVERSITÄT DES SAARLANDES

Deadline: Wednesday, January 8th, 2025 23:59 hrs

This problem set is worth a total of 50 points, consisting of 3 theory questions and 1 programming question. Please carefully follow the instructions below to ensure a valid submission:

- You are encouraged to work in groups of two students. Register your team (of 1 or 2 members) on the CMS at least ONE week before the submission deadline. You have to register your team for each assignment.
- All solutions, including coding answers, must be uploaded individually to the CMS under the corresponding assignment and problem number. On CMS you will find FOUR problems under each assignment. Make sure you upload correctly each of your solution against $Assignment\#X - Problem\ Y$ (where X- Assignment number and Y is the problem number) on CMS. In total you have to upload THREE PDFs (theoretical problems) and ONE ZIP file (programming problem).
- For each **theoretical question**, we encourage using LaTeX or Word to write your solutions for clarity and readability. Scanned handwritten solutions will be accepted as long as they are clean and easily legible. Final submission format must always be in a single PDF file per theoretical problem. Ensure your name, team member's name (if applicable), and matriculation numbers are clearly listed at the top of each PDF.
- For **programming question**, you need to upload a ZIP file to CMS under $Assignment\#X - Problem$ 4. Each ZIP file must contain a PDF or HTML exported from Jupyter Notebook and the .ipynb file with solutions. Make sure all cells in your Jupyter notebook contain your final answers. For creating PDF/HTML, use the export of the Jupyter notebook. Before exporting, ensure that all cells have been computed. To do this:
  - Go to the "Cell" menu at the top of the Jupyter interface.
  - Select "Run All" to execute every cell in your notebook.
  - Once all cells are executed, export the notebook: Click on "File" in the top menu.
  - Choose "Export As" and select either PDF or HTML.

  The submission should include your name, team member's name, and matriculation numbers at the top of both PDF/HTML and .ipynb file document.
- Finally, ensure academic integrity is maintained. Cite any external resources you use for your assignment.
- If you have any questions follow the instructions here.

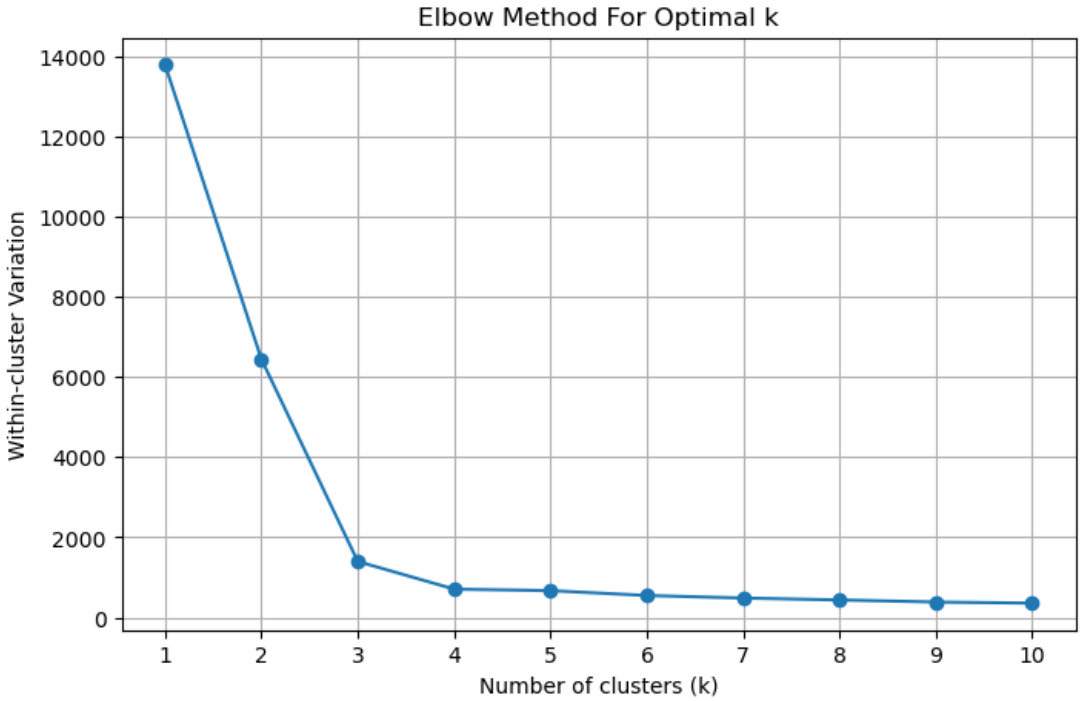
Problem 1 (K-means). (15 Points)

1. Given a dataset D of 4 points: (7 Points)

$$\mathbf{x}_1 = (1,\ 1) \quad \mathbf{x}_2 = (1,\ 4)$$
$$\mathbf{x}_3 = (4,\ 1) \quad \mathbf{x}_4 = (4,\ 4)$$

Perform 2 iterations of $k-$means clustering, using the Euclidean distance metric. Partition the dataset into two clusters (i.e., $k = 2$). Use the data points $\mathbf{x}_1$ and $\mathbf{x}_3$ as the initial cluster centroids $(\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2)$. In each iteration, calculate the required distances, show the cluster assignments for each data point, and indicate the centroid of each cluster. Has the algorithm converged? If so, why?

2. Below, you are given a plot representing the within-cluster variation (also known as inertia or within-cluster sum-of-squares, WCSS) for different numbers of clusters (k) in $k-$means.



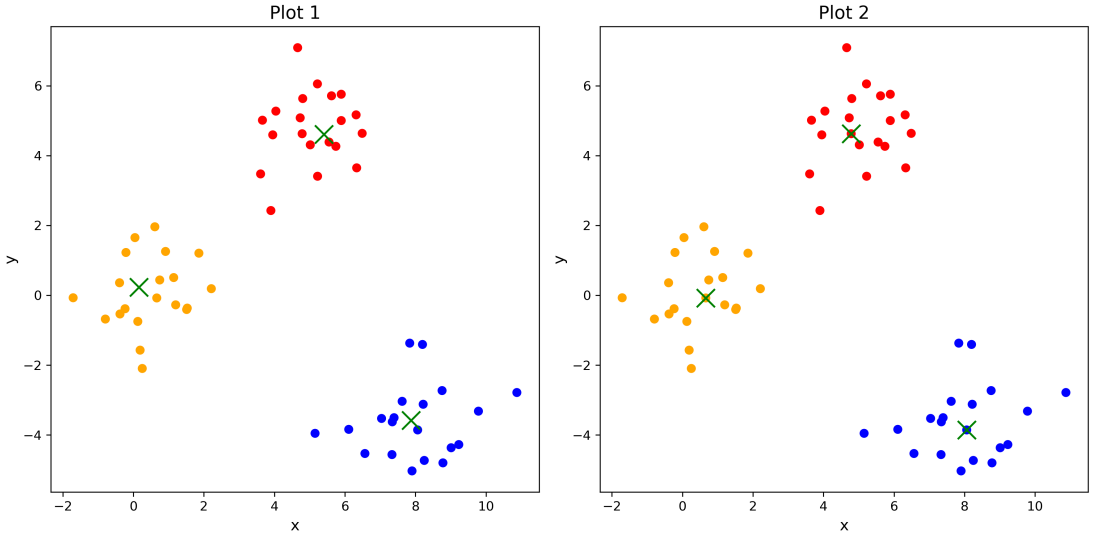
Assuming that: (1) Euclidean distance is used, and (2) the dataset has been preprocessed to remove duplicates (i.e., only unique points are present):

(a) According to the elbow heuristic, what is the optimal number of clusters for this dataset? Explain why did you choose this value. (Note: A range of values is also acceptable.) (2 points)

(b) Intuitively explain how the within-cluster variation changes as the number of clusters increases. (2 points)

(c) Intuitively explain under what conditions the within-cluster variation equals to zero. (2 points)

3. The figure below shows the resulting clusters for a random dataset using both $k-$means and $k-$medoids. Identify which of the two plots corresponds to $k-$medoids and explain your reasoning. (2 points)



Problem 2 (Hierarchical clustering and dissimilarity). (15 Points)

1. To perform $k-$means clustering, $k-$medoids, or agglomerative clustering, is it necessary to know the coordinates of the elements being clustered, or is it sufficient to have only their dissimilarity matrix? Explain your answer in detail. (3 Points)

2. The hierarchical clustering algorithm has already identified one cluster, and the dissimilarity matrix below represents the pairwise dissimilarities among the elements:

	$[A, B]$	C	D	E
$[A, B]$	0	1.2	1.8	2.3
C	1.2	0	1	1.7
D	1.8	1	0	1.4
E	2.3	1.7	1.4	0

   Using this dissimilarity matrix, perform one step of hierarchical clustering based on the following linkage methods

   (a) Complete linkage

   (b) Single linkage

   For each method, determine which clusters should merge in this step and show the updated dissimilarity matrix. Explain your answer. (8 Points)

3. You are given five data points: {A, B, C, D, E} which we will cluster. Using agglomerative hierarchical clustering, the clusters and similarity values at each step are as follows:

   (a) Merge A and B: Dissimilarity = 1.0

   (b) Merge C and D: Dissimilarity = 1.0

   (c) Merge {A, B} and {C, D}: Dissimilarity = 1.2

   (d) Merge {A, B, C, D} and E: Dissimilarity = 1.4

   Construct the dendrogram illustrating these steps. Ensure the height (y-axis values) of merges is clear. (4 Points)

Problem 3 (Dimensionality Reduction). (10 Points)
Given 3 data points in 2-dimensional space, $(1, 1)$, $(2, 2)$, and $(3, 3)$:

1. What information does the first principal component capture in terms of the data variance and the data explaining? (1 Points)
2. Calculate the first principal component. (3 Points)
3. Can PCA be used to reduce the dimensionality of a highly nonlinear dataset? Explain. (2 Points)
4. When might be sensible to chain two different dimensionality reduction algorithms? You can support your answer with an example. (2 Points)
5. How can you assess the effectiveness of a dimensionality reduction algorithm, used as a preprocessing step, on your dataset by considering the accuracy or error of a downstream model? (2 Points)

Problem 4 (Unsupervised Learning: Dimensionality Reduction and Clustering). (10 Points)

In this assignment, you will learn different clustering and dimensionality reduction techniques.

Please refer to the file `assignment_4_handout.ipynb` and **only** complete the sections marked in red and missing codes denoted with `#TODO`. Once you have filled in the required parts, revisit the submission instructions to check how to submit it.