HW 9: ECE 601 Machine Learning for Engineers

Important Notes:

1. When a HW question asks for writing a code, you would need to include the entire code as well as the output of the program as well as any other analysis requested in the question.
2. Don't panic about the length of the HW assignment. HW assignments are treated as opportunities for improving learning and understanding, so I might include some extra text to help you better understand the concepts or learn about a point that was not covered during the class. The actual work needed from you is indeed manageable.
3. Combine your solutions in **one** zip file called `homework9_UMassUSERNAME.zip`.

This homework consists of two coding questions designed to reinforce your understanding of clustering techniques using Python. You will work with real-world datasets from `scikit-learn`, applying K-means clustering. Each question is provided as a Jupyter Notebook (`.ipynb`) file as a starter code. Your task is to complete the missing sections, run the code, and submit the completed notebooks.

1. The Iris dataset is a classic dataset containing measurements of 150 iris flowers across three species: Setosa, Versicolor, and Virginica. Each flower is described by four features: sepal length, sepal width, petal length, and petal width (all in centimeters). In this task, you will apply K-means clustering to segment the flowers based on two features: sepal length and petal length. The goal is to group similar flowers into clusters, determine the optimal number of clusters, and visualize the results.

   Your objectives are as follows:

   (a) Load the Iris dataset from `scikit-learn` using the provided code in `Iris_Clustering_Question_1.ipynb`.
   (b) Select the features `sepal length (cm)` and `petal length (cm)` for clustering.
   (c) Use the Elbow Method to determine the optimal number of clusters (k) by plotting the Within-Cluster Sum of Squares (WCSS) for $k = 1$ to 10. Identify the "elbow" point where adding more clusters yields diminishing returns.
   (d) Apply K-means clustering with your chosen k to assign each flower to a cluster.
   (e) Visualize the clusters in a scatter plot with `sepal length (cm)` on the x-axis and `petal length (cm)` on the y-axis, using different colors for each cluster and marking the centroids.
   (f) Interpret the clusters in 2-3 sentences based on their patterns in the scatter plot.

   - Use the starter code in `Clustering_Homework_Iris_Question.ipynb`.
   - Include comments in your code to explain each step.
   - Submit the completed `.ipynb`.

Bonus (Optional): Add `petal width (cm)` as a third feature, perform 3D clustering, and visualize it using a 3D scatter plot. Discuss how this affects the clustering results.

2. The Digits dataset contains 1,797 images of handwritten digits (0-9), each represented as an 8x8 pixel grid flattened into a 64-dimensional vector of pixel intensities (0-16). Directly clustering 64-dimensional data is computationally intensive and hard to visualize, so you will first reduce the dimensionality to 2D using Principal Component Analysis (PCA) before applying K-means clustering.

   Your objectives are as follows:

   (a) Load the Digits dataset from `scikit-learn` using the provided code in `Handwritten_Clustering_Question_2.ipynb`.
   (b) Use the Elbow Method to determine the optimal number of clusters (k) by plotting the WCSS for $k = 1$ to 15.
   (c) Use the Silhouette Score to find the optimal number of clusters (k) for K-means clustering.
   (d) Apply PCA to reduce the 64-dimensional data to 2 dimensions (PCA Component 1 and PCA Component 2).
   (e) Apply K-means clustering with your chosen k to the PCA-transformed data to assign each digit to a cluster.
   (f) Visualize the clusters in a scatter plot with PCA Component 1 on the x-axis and PCA Component 2 on the y-axis, using different colors for each cluster and marking the centroids.

   - Use the starter code in `Handwritten_Clustering_Question_2.ipynb`.
   - Include comments in your code to explain each step.
   - Submit the completed `.ipynb`.