CSCI 6223 Practical Data Science

Homework #1

Due 11/01/2023

In this homework, you will use ***K-means for unsupervised learning***.

You will work together with your team to solve this problem.

Turn in one pdf document for the whole team!

The Business Problem

You are a data scientist working at a battery manufacturing company and the engineering department has asked for your help. The engineers have noticed that some of the batteries are failing sooner than expected. They perform a standard voltage test for quality assurance and have provided the latest 100 data samples from their last shipment. The data is included in the batteries.csv file. There are only two values in the dataset, V1 is voltage change at the input (mV), and V2 is voltage change at the output (mV).

They typically use a rating scale with 3 quality categories (Grade 1, Grade 2, Grade 3). They normally ship batteries in grades 1 and 2 and reject batteries in grade 3. They suspect that they need more grades but don't know how many to use. Can you help them determine the right number of grades they should use?

Project Deliverables (100 points)

Using k-means unsupervised learning, make a recommendation on the best number of grades to use. Specifically, provide the following in sections in your submission.

1. **Group Members** – list your group members as authors
2. **Business Problem** – include a description of the business problem and why you believe unsupervised learning is the right approach for this problem.
3. **Method** – describe the k-means algorithm and how it can help you solve this problem. Include any appropriate code snippets that you believe are relevant to your solution.
4. **Recommendations** – State your recommendations along with appropriate results to support your recommendations. Remember you are trying to convince the engineering management to adopt your recommendations.

   At a minimum, include the following plots/figures. Be sure to label the x and y axis appropriately and all plots should have a meaningful title.
   a. A ***scatter plot showing the un-labeled data*** with V1 on the x-axis and V2 on the y-axis. The plot should be like the one demonstrated in the lecture for the iris data.
   b. A plot showing your ***final clusters*** with centroids shown for each cluster.
   c. A plot showing the ***decision boundaries*** for your clusters.
   d. Plots supporting your decision for the ***optimum number of clusters*** (elbow and/or silhouette method). Be sure to explain why you chose the number of clusters based on these plots.