

Data Mining Final Practical File

24/48029

Vivaan Singh Adhikari November 22, 2025

Assignment Overview

Document Contents:

1. Introduction	2
1.1. Code	2

Assignment Details:

- **Course:** Data Mining DSE
- **Instructor:** Prof. Archana Gahalaut
- **Hardware:** No specifications
- **Software:** Python, Pandas, Typst(documentation)

Introduction

This assignment entails my solutions to the question assigned as per the course's guidelines. All the final files are available on <https://github.com/user7537/coursework/>

Code

QApply simple K-means algorithm for clustering any dataset. Compare the performance of clusters by varying the algorithm parameters. For a given set of parameters, plot a line graph depicting MSE obtained after each iteration.

```

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt

colnames = ["A1", "A2", "A3", "A4", "A5", "A6", "A7", "A8", "A9", "A10",
            "A11", "A12", "A13", "A14", "A15", "class"]

df = pd.read_csv("japanese_credit_screening/crx.data",
                 names=colnames,
                 na_values="?")

df = df.fillna(df.mode().iloc[0])

# Use only numeric attributes for clustering
num_df = df.select_dtypes(include=["float64", "int64"])

X = num_df.values

k = 3
max_iter = 20

# Random init
rng = np.random.RandomState(42)
centers = X[rng.choice(len(X), k, replace=False)]

mse_history = []

for i in range(max_iter):
    # Assign
    dist = np.linalg.norm(X[:, None] - centers[None, :], axis=2)
    labels = dist.argmin(axis=1)

    # Compute MSE
    mse = ((X - centers[labels])**2).sum() / len(X)
    mse_history.append(mse)

    # Update
    new_centers = np.vstack([X[labels==j].mean(axis=0) for j in
                             range(k)])
    if np.allclose(new_centers, centers):
        break
    centers = new_centers

plt.plot(mse_history)
plt.xlabel("Iteration")
plt.ylabel("MSE")
plt.title("K-Means MSE per iteration")
plt.savefig("q5_kmeans_mse.png")
print("Saved q5_kmeans_mse.png")

```