Tashfeen, Ahmad
CSCI 6213: Data Science Fundamentals

Homework 6

Let $x_1, \dots, x_n$ be a set of instances in $\mathbb{R}^p$ and $c_1, \dots, c_K$ be a set of K cluster centroids. Note that c_i are also in $\mathbb{R}^p$. Let $\mathbb{I}_{ij} \in \{0, 1\}$ be a variable such that $\mathbb{I}_{ij} = 1$ if x_i is assigned to cluster centred at c_j and $\mathbb{I}_{ij} = 0$ otherwise. In layman's terms, $\mathbb{I}_{ij}$ indicates whether x_i is in the j^{th} cluster. The K–Means Clustering algorithm minimises the following objective function,

$$\sum_{i=1}^{n}\sum_{j=1}^{K} \mathbb{I}_{ij}[(x_i - c_j)^T(x_i - c_j)]$$

This objective function is essentially a measure of how far are the members of each cluster from their respective centroids or means. The K–Means Clustering algorithm is given bellow,

1) Initialise centroids or means c_j for $1 \leq j \leq K$ randomly in $\mathbb{R}^p$, i. e., $c_j \leftarrow x_j$.
2) Assign all x_i to the closest centroid c_u. I. e., for some u and all j,

$$\textbf{if} \quad [(x_i - c_u)^T(x_i - c_u)] < [(x_i - c_j)^T(x_i - c_j)] \quad \textbf{then} \quad \mathbb{I}_{iu} \leftarrow 1$$

3) Adjust the centroids or means according the new assignment in the last step,

$$c_j \leftarrow \sum_{i=1}^{n} \mathbb{I}_{ij}\left(\frac{x_i}{n}\right)$$

   Note that this is just assigning c_j to the new mean of the j^{th} cluster.
4) Repeat from (2) till c_j stop changing or the change is less than some ε.

This algorithm is in section 6.12.1 of Mitchell or optionally algorithm 12.2 on page 523 of Hatie et al.

Question 1. Listing 1 plots the clusters shown in figure 1. Give your best guesses for the optimal centroids $c_j \in \mathbb{R}^2$ and $K = 4$.

```python
import numpy as np
from matplotlib import pyplot as plt

# don't change the seed
RNG = np.random.default_rng(3)
normal = RNG.multivariate_normal

t, n = 3, 2500
spread = np.identity(2)
X = normal([-t, t], np.identity(2), n)
X = np.vstack((X, normal([t, t], spread, n)))
X = np.vstack((X, normal([t, -t], spread, n)))
X = np.vstack((X, normal([-t, -t], spread, n)))
plt.title('Unlabelled Data Clusters')
plt.xlabel('x--axis')
plt.ylabel('y--axis')
plt.plot(X[:, 0], X[:, 1], 'k.', alpha=0.3)
plt.tight_layout()
plt.show()
```

Listing 1. Synthetic 2D clusters. Download.

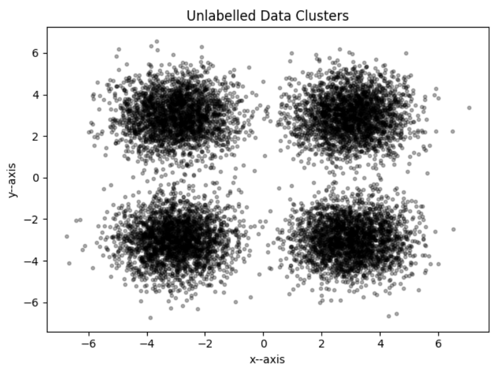

Figure 1. Clusters

Question 2. Implement the K–Means Clustering algorithm for the data in listing 1. Run the algorithm for $K \in \{1, 2, 3, 4\}$ and $\varepsilon = 10^{-5}$. Give your final converged centroids plotted over the clustered data. Plot each cluster with a distinct colour[1]. At the end you should have four converged (with different coloured clusters) plots for each value of K.

Question 3. Download the file `t10k-images-idx3-byte`. This is the binary file containing 10,000 images of the test set from the the MNIST database of handwritten digits. You can refer to their website for more information about the data and its organisation. However, it is read and plotted for your inspection and further usage to implement the K–Means Clustering algorithm in the linked file.

1) Why do you think $K = 10$ maybe a good choice for this data?
2) What are the values of n and p?
3) What is represented by each of the p features these data points?
4) For $\varepsilon = 3.5$, finish implementing the K–Means Clustering algorithm after the line indicated by the comment in the Python file given in question 3. The file has code that (when completed) should plot your 10 approximately converged cluster centroids as 28 by 28 pixel images comprising of 2 rows and 5 columns. Give and discuss this plot. Why do you think the centroids c_j look the way they do?

Submission Instructions

1) Submit a PDF that answers the questions and contains all the plots that the assignment asks for.
2) Submit your `kmeans.py`. This is your K–Means implementation for the code in listing 1.
3) Submit your `mnsit.py`. This is your K–Means implementation for the MNSIT test set in question 3.

Computer Science, Petree College of Arts & Sciences, Oklahoma City University

[1] use different shades of grey if you have difficulty seeing colour