Lab Course Machine Learning

Exercise Sheet 12

Prof. Dr. Dr. Lars Schmidt-Thieme
Kiran Madhusudhanan

HiWi: Basharat Mubashir Ahmed
Sophia Damilola Lawal

Submission deadline : **February 12, 2023**

General Instructions

1. Please perform any data pre-processing required.
2. Unless explicitly noted, you are not allowed to use scikit, sklearn or any other library for solving any part.
3. Please refrain from plagiarism.
4. The last exercise for this term!. And all the best for the exams.

1. K-Means (10 points)

A. `K-Means` algorithm splits a dataset $X \in \{x_1, \ldots, x_N\}$ into K many partitions, where each $X_k \subseteq X \quad \forall k \in \{1, \ldots, K\}$. Clustering algorithms like the `K-Means` is a useful technique when the true labels are unknown. Or in other words, we are basically interested in analyzing patterns within the data and make useful inferences.

In this task, you will implement a `K-Means` algorithm from scratch using the dataset "HTRU 2.csv". The dataset contains 8 continuous variables describing a pulsar candidate[1]. The task is to identify multiple (K) clusters that might best describe the classes within the data. Being a simple algorithm, we strongly advise you to implement the algorithm as per the lecture slides.

1. Initialize the cluster centers by selecting the first center at random and the rest sequentially based on the largest sum of distances to the selected cluster center.
2. Run with different random initialization.
3. Plot, a figure showing the selection of the best number of clusters K for each initialization.
4. Optimize the algorithm and show runtime improvements.

Try to compare your results (cluster centers, loss/distortion) and runtime to the sklearn implementation of `KMeans` clustering algorithm for the same dataset.

B. Principal Components Analysis (`PCA`) is a widely used method for reducing the number of dimensions to a low dimensional representation of the data. (You are allowed to use numpy.linalg.svd for single value decomposition). Use `PCA` to reduce the dimensionality of the data and represent the clusters (from the `K-Means`) in a 2D or 3D graph.

Compare your results with sklearn implementation of **PCA**.

[1] https://archive.ics.uci.edu/ml/datasets/HTRU2

2. Gaussian Mixtures (10 points)

In this exercise, you are required to implement `Gaussian Mixtures` for Soft Clustering using the Expectation Maximization (EM) Algorithm. We will use the same data as the one from the `K-Means` exercise.

1. Initialize clusters by drawing randomly from a uniform distribution.
2. Clearly specify the Expectation step and the Maximization step.
3. Plot, a figure showing the selection of the best number of clusters K
4. Plot the optimal cluster by assigning points to the cluster with the highest responsibility (Hard Clustering) using PCA.