Lab Course Machine Learning

Exercise Sheet 8

Prof. Dr. Dr. Lars Schmidt-Thieme
Jung Min Choi

HiWi: Harish Malik

Submission deadline : December 12, 2024

General Instructions

1. Data should be normalized.
2. Train to Test split should be 80-20 / with Validaiton 70-15-15
3. Convert any non-numeric values to numeric values. For example you can replace a country name with an integer value or more appropriately use one-hot encoding.
4. Please refrain from plagiarism.

1 K-Means (14 + 2 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.

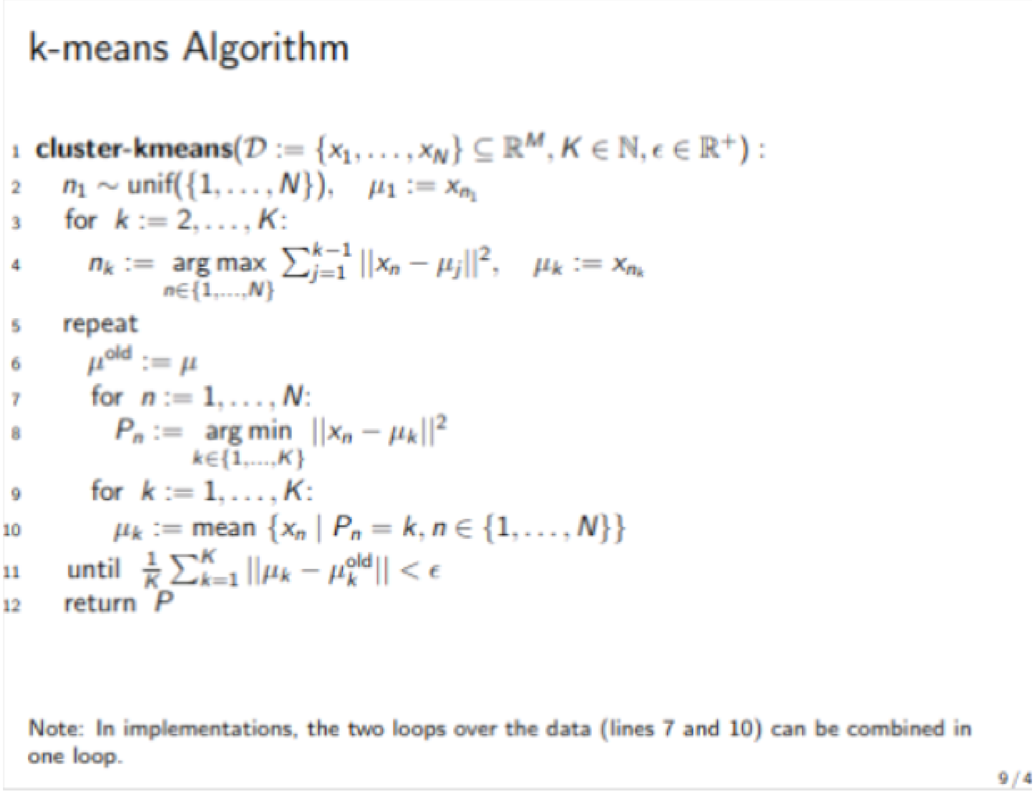
1. **[8 Points]** Initialize the cluster centers by selecting the first center at random and the rest sequentially based on the largest sum of distances to the selected cluster center. Run with different random initialization. Plot, a figure showing the selection of the best number of clusters K for each initialization.
2. **[Bonus** 2 Points]** Optimize the algorithm and show runtime improvements.
3. **[1 Point]** Try to compare your results (cluster centers, loss/distortion) and runtime to the sklearn implementation of `KMeans` clustering algorithm for the same dataset.
4. **[5 Points]** 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.

2 Gaussian Mixtures (6 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. **[5 Points]** Initialize clusters by drawing randomly from a uniform distribution and implement Gaussain Mixtures for Soft Clustering using the EM Algorithm. Plot a figure showing the selection of the best number of clusters K
2. **[1 Point]** Plot the optimal cluster by assigning points to the cluster with the highest responsibility (Hard Clustering) using PCA as the first question.

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



k-means Algorithm

```
1  cluster-kmeans($\mathcal{D} := \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^M, K \in \mathbb{N}, \epsilon \in \mathbb{R}^+$) :
2    $n_1 \sim$ unif($\{1, \ldots, N\}$),   $\mu_1 := x_{n_1}$
3    for $k := 2, \ldots, K$:
4      $n_k := \underset{n \in \{1,\ldots,N\}}{\arg\max} \sum_{j=1}^{k-1} ||x_n - \mu_j||^2$,   $\mu_k := x_{n_k}$
5    repeat
6      $\mu^{\text{old}} := \mu$
7      for $n := 1, \ldots, N$:
8        $P_n := \underset{k \in \{1,\ldots,K\}}{\arg\min} ||x_n - \mu_k||^2$
9      for $k := 1, \ldots, K$:
10       $\mu_k :=$ mean $\{x_n \mid P_n = k, n \in \{1, \ldots, N\}\}$
11   until $\frac{1}{K} \sum_{k=1}^{K} ||\mu_k - \mu_k^{\text{old}}|| < \epsilon$
12   return $P$
```

Note: In implementations, the two loops over the data (lines 7 and 10) can be combined in one loop.

9 / 4