Applied Machine Learning
Exercise 10

Prof. Dr. Amr Alanwar

December 2024

Datasets

1. **Document Dataset:**
    - IRIS dataset D1
    - rcv1v2 (topics; subsets) D2
    - 20Newsgroups dataset D3

Exercise 1: Implement K-Means Clustering Algorithm (5 Points)

Implement he K-Means (`cluster-kmeans`) algorithm using datasets D1 or D2. Your implementation should handle sparse data (**Note:** D2 is a sparse dataset; see Annex below for more details). Finally, choose a criterion for selecting an optimal value of k (number of clusters).

Exercise 2: Cluster News Articles (5 Points)

The dataset D3 (20Newsgroups) can be downloaded as `20news-bydate.tar.gz`. Each news article is stored in its group folder (e.g., articles corresponding to "alt.atheism" are in the `alt.atheism` folder). Pre-process the data and extract features for each document. Store the data in a libsvm file format. Use the provided train and test splits.

Tasks:

1. Cluster the 20Newsgroups dataset using your implementation of the K-Means algorithm. Use the test data to measure the clustering quality.
2. Use a K-Means implementation from a software library of your choice. Compare the results of your implementation with the library's implementation:
    - What is the optimal value of k in each case?
    - Which implementation takes longer (time your program)?

Hint: Use Python's `time` or `timeit` libraries for timing your code. Note that you are not allowed to use `sklearn.datasets.fetch_20newsgroups` for Exercise 1 or 2.

Annex

rcv1v2 (topics; subsets) D2: This dataset, available at: `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel.html#rcv1v2(topics;subsets)`, contains multiple labels. An alternative version is available at: `https://archive.ics.uci.edu/ml/datasets/Reuters+RCV1+RCV2+Multilingual,+Multiview+Text+Categorization+Test+collection`. Multiple files and folders are provided; you can use `Index_EN-EN` for the original English documents in the `EN` folder.

Algorithm 1 CLUSTER-KMEANS

Require: $D = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^M$, $K \in \mathbb{N}, \epsilon \in \mathbb{R}^+$

$i_1 \sim \text{unif}\{1, \ldots, N\}$, $\mu_1 := x_{i_1}$
for $k = 2, \ldots, K$ **do**
    $i_k := \arg\max_{n \in \{1,\ldots,N\}} \sum_{k'=1}^{k-1} \|x_n - \mu_{k'}\|$, $\mu_k := x_{i_k}$
end for
repeat
    $\mu^{\text{old}} := \mu$
    for $n = 1, \ldots, N$ **do**
        $P_n := \arg\min_{k \in \{1,\ldots,K\}} \|x_n - \mu_k\|$
    end for
    for $k = 1, \ldots, K$ **do**
        $\mu_k := \text{mean}\{x_n \mid P_n = k\}$
    end for
until $\frac{1}{K} \sum_{k=1}^{K} \|\mu_k - \mu_k^{\text{old}}\| < \epsilon$
return P