Lab Course Machine Learning
Exercise 8

Prof. Dr. Dr. Lars Schmidt-Thieme,
Mofassir ul Islam Arif
Information Systems and Machine Learning Lab
University of Hildesheim
Submission: 03.02.2020 LearnWeb 3115

January 28, 2020

Instructions

Please following these instructions for solving and submitting the exercise sheet.

1. You should submit two things a) python scripts(zipped) / jupyter notebook and b) a pdf document.
2. In the pdf document you will explain your approach (i.e. how you solved a given problem), and present your results in form of graphs and tables.
3. The submission should be made before the deadline, only through learnweb.
4. Unless explicitly mentioned, you are not allowed to use scikit, sklearn or any other library for solve any part. All implementations must be done yourself.

1 Exercise Sheet 11

Datasets

- 1. Sparse dataset :

  (a) IRIS dataset D1:
  `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass/iris.scale`

  (b) rcv1v2 (topics; subsets D2:
  `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel.html#rcv1v2(topics;subsets)`
  (c) 20Newsgroups dataset D3: `http://qwone.com/~jason/20Newsgroups/`

Exercise 1: Implement K Means clustering algorithm (10 Points)

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

Figure 1: K-Means Alogrithm

The K Means algorithm (cluster-kmeans) is given in here. Implement this algorithm. You should use D1 or D2 datasets. Your algorithm should be able to handle sparse data ([note: D2 is a sparse dataset, more details in Annex below). Finally, you should also choose a criterion for selecting an optimal value of k (number of clusters).

Also, please explicitly differentiate between K-Means and KNN. There is ONE crucial difference that you need to mention here.

Exercise 2: Cluster news articles(10 Points)

D3 is 20Newsgroups dataset (download "20news-bydate.tar.gz"). Each news article is stored as a file in its group folder i.e. all articles corresponding to "alt.atheism" are placed in "alt.atheism folder". Do appropriate pre-processing of the data and extract features for each document (we have covered this in the SVM lab). After preprocessing you need to store data in a libsvm file format. Note that you are provided with train and test splits. Use these train and test splits. Cluster the 20newsgroup dataset using your own implementation of Kmeans algorithm. Use test data to measure the quality of the clustering algorithm. The second part of this exercise is to use a kmeans provided by a software library of your choice. Compare the results of your implementation with kmeans library. What optimal value of K you get in both the cases. Which implementation takes longer i.e. time your program. [Hint: look at time or timeit library for the timing portion of your code. Scikit learn provides a function $sklearn.datasets.fetch\_20newsgroups$, which is not allowed to use for implementing Exercise 1 and 2].

1.1 ANNEX

- rcv1v2 Help: rcv1v2 (topics; subsets) D2: dataset provided at `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel.html#rcv1v2(topics;`

subsets) has multiple labels. Another online version is available at https://archive.ics.uci.edu/ml/datasets/Reuters+RCV1+RCV2+Multilingual,+Multiview+Text+Categorization+Test+collection. There are multiple files and folders you can pick $Index\_EN - EN$: Original English documents, inside EN folder.