Lab Course Machine Learning
Exercise 10

Prof. Dr. Dr. Lars Schmidt-Thieme,
Mofassir ul Islam Arif
Information Systems and Machine Learning Lab
University of Hildesheim
Submission: 3.02.2019 LearnWeb 3112

January 28, 2019

1 Exercise Sheet 10

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)

The K Means algorithm (cluster-kmeans) is given in the lecture `https://www.ismll.uni-hildesheim.de/lehre/ml-16w/script/ml-10-B1-cluster-analysis.pdf`. 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).

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. 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 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 results of your implementation with kmeans library. What optimal value of K you get in both the cases. Which implementation take longer i.e. time your program. [Hint: look at time or timeit library for timing portion of your code. Scikit learn provides a function $sklearn.datasets.fetch\_20newsgroups$, which is not allowed to use for implementing Exercise1 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.