Lab Course Machine Learning
Exercise Sheet 11

Prof. Dr. Dr. Lars Schmidt-Thieme, Mohsan Jameel
Information Systems and Machine Learning Lab
University of Hildesheim

February 1st, 2017
Submission on January 26th, 2017 at 11:55pm, (on moodle, course code 3112)

Instructions

Please read the lab related instructions, i.e. submission, report format and policies, at `https://www.ismll.uni-hildesheim.de/lehre/prakAIML-16w/exercises/ml_lab_instructions.pdf`

Datasets

1. **Document dataset:**
   (a) IRIS dataset $\mathcal{D}_1$: `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multiclass/iris.scale`
   (b) rcv1v2 (topics; subsets $\mathcal{D}_2$: `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/multilabel.html#rcv1v2(topics;subsets)`
   (c) 20Newsgroups dataset $\mathcal{D}_3$: `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 $\mathcal{D}_1$ or $\mathcal{D}_2$ datasets. Your algorithm should be able to handle sparse data ([note: $\mathcal{D}_2$ 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)

$\mathcal{D}_3$ 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].

Annex

rcv1v2 (topics; subsets) $\mathcal{D}_2$: 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.