Data Mining and Matrices
Assignment 3: Non-Negative Matrix Factorization

This assignment focuses on NMF. We provide implementations of all relevant methods; your task is to experiment with these methods and evaluate the results.

We will work with a dataset that consists of a subset of the 20-newsgroups dataset (`http://qwone.com/~jason/20Newsgroups/`). The subset contains 100 news articles from each of the following Usenet groups: `sci.crypt`, `sci.med`, `sci.space`, and `soc.religion.christian`. Terms have been stemmed, stop words removed, and only the 800 most popular terms have been retained. The data is given in form of an 400×800 document-term matrix; entry (d, w) denote the term frequency (tf) of word w in document d. Before working with the newsgroup data, normalize the entries of the data matrix such that they sum to 1.

The goals of this assignment are: (1) find latent topics in the newsgroup data using NMF, (2) cluster the documents (using NMF and other methods of your choice). The data is given in a CSV file named `news.csv`. The R script `util.R` contains functions relevant to NMF, the file `03-nmf-task.R` contains some examples on how to use these functions. Familiarize yourself with the dataset and the provided functions before working through the assignment.

1 Topic modelling with NMF

We will first apply NMF to the normalized newsgroup data to discover latent topics. Start by computing the NMF optimizing the KL divergence for $r = 4$ factors. The output of this step will be two unnormalized matrices $\tilde{\boldsymbol{L}}$ and $\tilde{\boldsymbol{R}}$.

a) Study the top-10 terms in the right-hand factor matrix $\tilde{\boldsymbol{R}}$ (see `03-nmf-task.R`). Are the results good? To evaluate the results, you must consider the terms associated with each topic and argue why you think they do or do not constitute a meaningful topic (and what that topic is)—remember, we understand the term "topic" very loosely, e.g., "words related to sports" is a valid topic. Also, the topics are not necessarily only those of the newsgroups. Remember also that the terms are stemmed.[1]

b) Study the reconstructed matrix. Does it look like you would have expected? Which aspects are covered well? Which are not?

[1] `http://en.wikipedia.org/wiki/Stemming`

c) Take the rank-4 truncated SVD of the data and study the decompositon along the lines mentioned above. Compare!

d) Now try different values of k (at least $r = 2$ and $r = 8$) and repeat the analysis (for NMF only). How do the results change? Can you name a single best rank?

e) Apply Gaussian NMF (i.e., using Euclidean norm). Do the results change? In your opinion, which NMF variant produces better results, if any? Argue!

2 PLSA

The R script provides two functions, which factor the unnormalized output $\tilde{\boldsymbol{L}}$ and $\tilde{\boldsymbol{R}}$ of the NMF into three non-negative matrices:

a) `nmf.lsr` produces an $m \times r$ matrix $\boldsymbol{L}'$, and $r \times r$ diagonal matrix $\boldsymbol{\Sigma}'$, and an $r \times n$ matrix $\boldsymbol{R}'$ such that $\tilde{\boldsymbol{L}}\tilde{\boldsymbol{R}} = \boldsymbol{L}'\boldsymbol{\Sigma}'\boldsymbol{R}'$, and the columns of $\boldsymbol{L}'$ as well as the rows of $\boldsymbol{R}'$ sum to one.

b) `nmf.slr` produces an $m \times m$ diagonal matrix $\boldsymbol{\Sigma}''$, an $m \times r$ matrix $\boldsymbol{L}''$, and an $r \times n$ matrix $\boldsymbol{R}''$ such that $\tilde{\boldsymbol{L}}\tilde{\boldsymbol{R}} = \boldsymbol{\Sigma}''\boldsymbol{L}''\boldsymbol{R}''$, and the rows of $\boldsymbol{L}''$ as well as the rows of $\boldsymbol{R}''$ sum to one.

Run NMF with KL divergence and $r = 4$ and factor the resulting decomposition using each of the two functions above. Study the result. Which information is contained in each of the three matrices? What can you say about the sum of the entries in each matrix? Can you give a probabilistic interpretation of the result (i.e., each entry (i, j) of each matrix)?

3 Clustering

The documents in the data came from four newsgroups. Your task is to cluster the documents in such a way that the clusters correspond to the newsgroups (which we can think of as topics). Note that you are not allowed to use the class labels during clustering, i.e., we pretend that we are in an unsupervised setting.

To evaluate the quality of the clustering, we treat cluster identifiers as predicted labels and consider the accuracy (fraction of correctly predicted labels) and the confusion matrix[2]. Examples can be found in the provided R code.

Cluster the normalized newsgroup data into 4 clusters using each of the methods below and study the results. Also look at the clusters manually. Which clustering(s) perform well, which do not? Why?

a) k-means,

b) k-means on $\boldsymbol{U}_4\boldsymbol{\Sigma}_4$ (i.e., the first two factor matrices of rank-4 truncated SVD),

c) k-means on the $\tilde{\boldsymbol{L}}$ matrix of the NMF (using KL divergence and $r = 4$),

[2] `https://en.wikipedia.org/wiki/Confusion_matrix`

d) k-means on the $\boldsymbol{L}'$ matrix of factorization $\boldsymbol{L}'\boldsymbol{\Sigma}'\boldsymbol{R}'$ obtained from the NMF, and

e) k-means on the $\boldsymbol{L}''$ matrix of factorization $\boldsymbol{\Sigma}''\boldsymbol{L}''\boldsymbol{R}''$ obtained from the NMF.

4 Beat the NMF (optional)

Experiment with preprocessing methods and alternative clustering methods to see if you can find a method that works better than the best result obtained in the previous task.

Handing in Your Solution

As before, your report should document what you did and which results you obtained. For example, give the top-10 terms per topic as well as the names you would assign to each of the topics for NMF (the latter at least for $r \leq 8$). Return a full transcript of the R commands that you used as a separate document.