Exercises
Computational Intelligence Lab
SS 2019

Machine Learning Institute
Dept. of Computer Science, ETH Zürich
Prof. Dr. Thomas Hofmann
Web http://cil.inf.ethz.ch/

Series 4, March 8, 2018
(Non-Negative Matrix Factorization)

Problem 1 (Implementing pLSA for Discovering Topics in a Corpus):

In this question, we are going to use probabilistic latent semantic models (pLSA) to discover topics in a corpus of documents. We will use a preprocessed dataset of documents from the Associated Press (courtesy of https://github.com/kzhai/PyLDA) which contains a collection of 2221 documents, "doc.dat," which has been conveniently split into train and test sets of sizes 2000 and 221 documents respectively.

Fortunately, thanks to the pre-processed data and some handy libraries, you do not need to do any additional data wrangling. Everything is ready for training models.

Setup:

- Download the Associated Press corpus, "associated-press.tar.gz" and the Jupyter notebook template "pLSA-for-the-AP.ipynb" from the lecture's github repository (link below).
- Install the dependencies listed in the "README.md" file.
- Follow instructions in the notebook. You will implement the Expectation-Maximization algorithm and do some basic analysis of the resulting model and learned topics.

`https://github.com/dalab/lecture_cil_public/tree/master/exercises/ex5`

Questions:

1. Why does the lower bound increase on each iteration?
2. Why does the log-likelihood increase on each iteration?
3. What are the learned topics?
4. What is the best choice for the number of topics? How do you know?

Problem 2 (Latent semantic models):

Recall the probability semantic model: for each word w and document d, we want to model $p(w|d)$. We use two different probabilistic models here:

i. For each pairs of word w and document d, we have

$$p(w|d) = u_{wd} \geq 0, \quad \sum_w u_{wd} = 1. \tag{1}$$

ii. In a probabilistic latent semantic model (pLSM), we consider a latent (a.k.a. hidden, or unobserved) variable which we call a "topic," denoted by $z \in \{1, \ldots, k\}$. This variable intermediates between documents and words precisely as follows:

$$p(w|d) = \sum_{i=1}^{k} p(w, z = i|d) = \sum_{i=1}^{k} p(w|z = i, d)p(z = i|d) \tag{2}$$

$$= \sum_{i=1}^{k} \underbrace{p(w|z = i)}_{u_{wi}} \underbrace{p(z = i|d)}_{v_{id}} \tag{3}$$

Where Equation (3) follows from a conditional independence assumption on w and d given z.

Suppose you are given a dataset containing D documents with a total vocabulary size of W words. Compare the number of parameters that need to be estimated for models i. with that of model ii. Which model requires fewer parameters and by what order of magnitude? Discuss the trade-offs between the two models not only in terms of memory-size, but also in terms of applicability.

Problem 3 (EM for LSM):

In this exercise, we consider maximum likelihood estimation of parameters of model ii. of problem 2. We consider a Bag-of-Words model in which we do not care how many times a word appears in a document. Namely, $X_{wd} \in \{0, 1\}$ and is equal to one if word w appeared in document d and zero otherwise. The log-likelihood of X is

$$\log \ell(X; U, V) = \sum_{w,d} X_{wd} \log \sum_{i=1}^{k} p(w|z = i)p(z = i|d) \tag{4}$$

$$= \sum_{w,d} X_{wd} \log \sum_{i=1}^{k} u_{wi} v_{id} \tag{5}$$

i. Consider the following maximum likelihood problem for the inference of u_{wi}s and v_{di}s.

$$\max_{U,V} \ell(X; U, V), \tag{6}$$

$$\text{subject to} \sum_{w} u_{wi} = 1, \sum_{i} v_{di} = 1, u_{wi} \geq 0, \text{ and } v_{id} \geq 0 \tag{7}$$

Explain why we can not solve the maximum likelihood estimation in a closed form.

ii. Suppose that the latent variables z were observed (for example assume that the topic of document d is z_d and word w is assigned to topic z_w). In this case, can we compute the solution in a closed form?

iii. Now we assume z is latent, but we know probability $p(z = i|wd) = q_{iwd}$. Then we can write the log-likelihood as

$$\log \ell(X; U, V) = \sum_{w,d} X_{wd} \log \sum_{i=1}^{k} q_{iwd} p(w|z = i)p(z = i|d)/q_{wid} \tag{8}$$

Show that

$$g(X; U, V) := \sum_{w,d} q_{iwd} X_{wd} \left(\log(p(w|z = i)) + \log(p(z = i|d)) - \log(q_{iwd})\right) \leq \log \ell(X; U, V) \tag{9}$$

holds. **Hint:** $\sum_i q_{wid} = 1$ (why?)

iv. To maximise the established lower-bound in the last section, we have to solve the following maximization:

$$\max_{U,V} g(X; U, V) \tag{10}$$

$$\text{subject to} \sum_{w} u_{wi} = 1, \sum_{i} v_{di} = 1, u_{wi} \geq 0, \text{ and } v_{id} \geq 0 \tag{11}$$

Why inequality constraints are redundant in the above maximisation?

v. Compute the solution of (10) using method of Lagrangian multipliers.