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 5 Solutions (Non-Negative Matrix Factorization)

Note that the order of questions is changed intentionally.

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 the last step 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.
Solution : Model (i) needs to estimate $D \times W$ parameters $\{u_{wd}\}$. But model (ii) has $D \times K + K \times W$ parameters. Considering K is relatively much smaller than D and W, model (ii) needs less parameters to estimation. However, inference of u_{wd} by minimizing negative log-likelihood function is a convex optimization program, while one need to solve a non-convex optimization problem to estimate marginal probabilities u_{wi} and v_{id} in model ii [1]. Model (ii) can group documents by some topics while model (i) does not provide such information. Despite these differences, two models share one common feature: they both skip the sequence(ordering) of words in a document.

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, $X_{wd} \in \{0, 1, \ldots\}$ and is equal to one if word w appeared in document d and zero otherwise. The log-likelihood of X is

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

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

Solution remarks: Let us explain more about the derivation of the above log-likelihood. Recall the probability of observing word w_i given the document d is

$$p(w_i|d) = \sum_{i=1}^{k} p(w_i|z = k)p(z = k) \tag{6}$$

[1] We will show this formally in the next exercise.

Let document d be a bag of words $w_1, w_2, \dots, w_m$. Assuming each word appears independently for the others[2], we have

$$P(w_1, \dots, w_m|d) = \prod_{j=1}^{m} p(w_j|d) = \prod_{w} p(w|d)^{X_{wd}} \tag{7}$$

Replacing Eq. (6) in the above equation and taking log concludes the $\log$-likelihood function.

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

$$\max_{U,V} \ell(X; U, V), \tag{8}$$

$$\text{subject to} \sum_{w} u_{wi} = 1, \sum_{i} v_{id} = 1, u_{wi} \geq 0, \text{ and } v_{id} \geq 0 \tag{9}$$

Explain why we can not solve the maximum likelihood estimation in a closed form.
Solution i. The function $-\log(\ell)$ is not convex in u_{wi} and v_{id} jointly. Let's start with an illustrative example. If there was only one topic $z = 0$, then the $-\log(\ell)$ becomes convex. For example, consider the simplified likelihood formulation for one document with one words as

$$-\log(\ell(u, v)) = -\log(uv) = -\log(u) - \log(v) \tag{10}$$

which is convex in u and v (since $-\log(.)$ is convex and sum of two convex functions is convex). But the above decomposition does not extend to more than one topics. To justify this, we consider a the likelihood over two topics for one document and one word as

$$h(x) = -\log(u_1 v_1 + u_2 v_2), \quad x = (u_1, v_1, u_2, v_2) \tag{11}$$

The above function is not convex in (u_1, v_1). To check the non-convexity, we pick the following two vectors

$$x = (1, 1, 0, 0), \quad y = (0, 0, 1, 1) \tag{12}$$

The you can see

$$h(x/2 + y/2) = -\log(1/2) > 0 = (h(x) + h(y))/2 \tag{13}$$

which shows that the convexity condition does not hold in (u_1, v_1, u_2, v_2) jointly.

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?
Solution ii. Let $Z_{wdi} \in \{0, 1\}$ is one if both of document d and word w are associated with topic i, otherwise $Z_{wdi} = 0$. Using observations Z_{wdi}, we derive the log-likelihood as

$$-\log(\ell) = -\sum_{w,d} X_{wd} Z_{wdi} \log u_{wi} v_{id} \tag{14}$$

$$= -\sum_{w,d} X_{wd} Z_{wdi} \left(\log(u_{wi}) + \log(v_{id})\right) \tag{15}$$

The above objective is convex in u_{wi} and v_{id}. The closed-form solution can be obtained by making the Lagrangian function and setting the gradient to zero. More details on this approach is provided part v.

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

$$\log \ell(X; U, V) = \sum_{w,d} X_{wd} \log \sum_{i=1}^{k} q_{iwd} u_{wi} v_{id} / q_{wid} \tag{16}$$

Show that

$$g(X; U, V) := \sum_{w,d,i} q_{iwd} X_{wd} \left(\log(u_{wi}) + \log(v_{id})\right) - \log(q_{iwd})) \leq \log \ell(X; U, V) \tag{17}$$

[2] In bag of words model for documents, each document is a collection of words and the sequencing does not matter.

holds.
Solution iii. Since q_{iwd} is a posterior the following holds:

$$\sum_i q_{iwd} = \sum_i p(z = i|wd) = 1 \tag{18}$$

Hence we can use Jensen's inequality to established the desired lower-bound

$$\sum_{w,d} X_{wd} \log \sum_{i=1}^{k} q_{iwd} u_{wi} v_{id} / q_{wid} \geq \sum_{w,d,i} q_{iwd} X_{wd} \left(\log(u_{wi}) + \log(v_{id})) - \log(q_{iwd})\right) \tag{19}$$

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

$$\max_{U,V} g(X; U, V) \tag{20}$$

$$\text{subject to} \sum_w u_{wi} = 1, \sum_i v_{id} = 1, u_{wi} \geq 0, \text{ and } v_{id} \geq 0 \tag{21}$$

Why inequality constraints are redundant in the above maximisation?
Solution iv. The domain $\log(u_{wi})$ is limited to non-negative u_{wi}. Note that if $u_{wi} = 0$, then g goes to $-\infty$. So u_{wi} can not be zero neither.

v. Compute the solution of (20) using method of Lagrangian multipliers.
Solution v. We first make the Lagrangian function. Let α, β to be the multipliers for u, v, respectively, then the Lagrangian function is (arguments except for α, β skipped for simplicity):

$$\mathcal{L}_{U,V}(\alpha, \beta) = -g(X; U, V) + \sum_i \alpha_i (\sum_w u_{iw} - 1) + \sum_d \beta_d (\sum_i v_{id} - 1) \tag{22}$$

Then the optimal U, V can be obtained by solving the following unconstraint problem

$$\min_{U,V} \max_{\alpha,\beta} \mathcal{L}_{U,V}(\alpha, \beta) \tag{23}$$

Remarkably, when the constraint is not met the maximisation over α and β makes $\mathcal{L}$ infinity. In this way, we reduce a constrained optimization problem to a non-smooth min-max problem. We set the gradient with respect to u_{wi} and α_i to zero to find the minimum of $\mathcal{L}$ with respect to u_{wi}.

$$\partial \mathcal{L} / \partial u_{wi} = 0 \Leftrightarrow -\sum_d X_{wd} q_{iwd} / u_{wi} + \alpha_i = 0 \Leftrightarrow u_{wi} = \sum_d X_{wd} q_{iwd} / \alpha_i \tag{24}$$

Setting $\partial \mathcal{L} / \partial \alpha_i$ to zero zero yields

$$\sum_w u_{wi} = 1 \Leftrightarrow \sum_{w,d} X_{wd} q_{iwd} / \alpha_i = 1 \Leftrightarrow \alpha_i^{-1} = \sum_{w,d} X_{wd} q_{iwd} \tag{25}$$

Replacing α_i in the formulation of u_{wi} concludes the derivation of u_{wi}:

$$u_{wi} = \sum_d X_{wd} q_{iwd} / (\sum_{dw} X_{wd} q_{iwd}) \tag{26}$$

Similarly, we derive the optimum v_{id} by setting the gradient to zero

$$\partial \mathcal{L} / \partial v_{id} = 0 \Leftrightarrow -\sum_w X_{wd} q_{iwd} / v_{id} + \beta_d \Leftrightarrow v_{id} = \sum_w X_{wd} q_{iwd} / \beta_d \tag{27}$$

The constraint $\sum_i v_{id} = 1$ determines the choice of β_d:

$$\sum_i v_{id} = 1 \Leftrightarrow \sum_{iw} X_{wd} q_{iwd} / \beta_d = 1 \Leftrightarrow \sum_w X_{wd} \left(\underbrace{\sum_i q_{iwd}}_{=1} \right) / \beta_d = 1 \Leftrightarrow \beta_d^{-1} = \sum_w X_{wd} \tag{28}$$

Replacing β_d into the derivation of v_{id} yields

$$v_{id} = \sum_w X_{wd} q_{iwd} / \left(\sum_w X_{wd} \right) \tag{29}$$

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?
Solution: Recall EM steps as

$$\text{E step:} \qquad q_{iwd} \leftarrow \frac{u_{wi}v_{id}}{\sum_i u_{wi}v_{id}} \tag{30}$$

$$\text{M step:} \qquad (U,V) \leftarrow \arg\max_{U,V} g(Q,U,V), \quad \text{subject to } \sum_w u_{wi} = 1, \sum_i v_{id} = 1 \tag{31}$$

Clearly the lower-bound does not decrease in M-step. We need to show that the g does not decrease in E-step neither. We claim that the E-step is derived from the following maximization problem

$$\max_Q g(Q,U,V), \quad \text{subject to } \sum_i q_{iwd} = 1 \tag{32}$$

To prove this, we need to construct the Lagrangian function and setting its gradient to zero:

$$L(\alpha, Q) = -g(Q,U,V) + \sum_{wd} \alpha_{wd}(\sum_{iwd} q_{iwd} - 1) \tag{33}$$

We first derive the optimality condition on q_{iwd}:

$$\partial L/\partial q_{iwd} = -\log(u_{wi}) - \log(v_{id}) + \log(q_{iwd}) + 1 + \alpha_{wd} = 0 \Leftrightarrow q_{iwd} = C u_{wi} v_{id} \tag{34}$$

Then the optimality condition on α_{wd} implies that $C^{-1} = \sum_i u_{wi} v_{id}$. This concludes the E-step in EM updates.

2. Why does the log-likelihood increase on each iteration?
Solution: Let's write the EM recurrence using more convenient notations

$$\text{E step:} \qquad Q_{n+1} = \arg\max_Q g(Q, U_n, V_n) \tag{35}$$

$$\text{M step:} \qquad (U_{n+1}, V_{n+1}) = \arg\max_{U,V} g(Q_{n+1}, U, V), \tag{36}$$

where we skipped constraints. One can readily check that $g(Q_{n+1}, U_n, V_n) = \log \ell(U_n, V_n)$, hence

$$\log \ell(U_n, V_n) = g(Q_{n+1}, U_n, V_n) \leq g(Q_{n+1}, U_{n+1}, V_{n+1}) \leq g(Q_{n+2}, U_{n+1}, V_{n+1}) = \log \ell(U_{n+1}, V_{n+1})$$

Indeed, EM is an alternating maximisation algorithm.

3. What are the learned topics? **Solution:** See `https://github.com/dalab/lecture_cil_public/tree/master/exercises/ex5`.

4. What is the best choice for the number of topics? How do you know?