

Nonnegative Matrix Factorization and Probabilistic Latent Semantic Indexing: Equivalence, Chi-square Statistic, and a Hybrid Method

Chris Ding^a, Tao Li^b and Wei Peng^b

^a Lawrence Berkeley National Laboratory, Berkeley, CA 94720

^b School of Computer Science, Florida International University, Miami, FL 33199

Abstract

Non-negative Matrix Factorization (NMF) and Probabilistic Latent Semantic Indexing (PLSI) have been successfully applied to document clustering recently. In this paper, we show that PLSI and NMF optimize the same objective function, although PLSI and NMF are different algorithms as verified by experiments. This provides a theoretical basis for a new hybrid method that runs PLSI and NMF alternatively, each jumping out of local minima of the other method successively, thus achieving better final solution. Extensive experiments on 5 real-life datasets show relations between NMF and PLSI, and indicate the hybrid method lead to significant improvements over NMF-only or PLSI-only methods. We also show that at first order approximation, NMF is identical to χ^2 -statistic.

Introduction

Document clustering has been widely used as a fundamental and effective tool for efficient document organization, summarization, navigation and retrieval of large amount of documents. Generally document clustering problems are determined by the three basic tightly-coupled components: a) the (physical) representation of the given data set; b) The criterion/objective function which the clustering solutions should aim to optimize; c) The optimization procedure (Li 2005).

Among clustering methods, the K-means algorithm has been the most popularly used. A recent development is the Probabilistic Latent Semantic Indexing (PLSI). PLSI is a unsupervised learning method based on statistical latent class models and has been successfully applied to document clustering (Hofmann 1999). (PLSI is further developed into a more comprehensive Latent Dirichlet Allocation model (Blei, Ng, & Jordan 2003).)

Nonnegative Matrix Factorization (NMF) is another recent development for document clustering. Initial work on NMF (Lee & Seung 1999; 2001) emphasizes the contain coherent parts of the original data (images). Later work (Xu, Liu, & Gong 2003; Pauca *et al.* 2004) show the usefulness of NMF for clustering with in experiments on documents

Copyright © 2006, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

collections, and a recent theoretical analysis (Ding, He, & Simon 2005) shows the equivalence between NMF and K -means / spectral clustering.

Despite significant research on both NMF and PLSI, few attempts have been made to establish the connections between them while highlighting their differences in the clustering framework. Gaussier and Goutte (Gaussier & Goutte 2005) made the first connection between NMF and PLSI, by showing that the local fixed point solutions of the iterative procedures of PLSI and NMF are the same. Their proof is, however, incorrect. NMF and PLSI are different algorithms. They converge to different solutions while starting from the same initial condition, as verified by experiments (see later sections).

In this paper, we first show that both NMF and PLSI optimize the same objective function. This fundamental fact and the L_1 normalization NMF ensures that NMF and PLSI are equivalent.

Second, we show, by an example and extensive experiments, that NMF and PLSI are different algorithms and they converge to different local minima. This leads to a new insight: NMF and PLSI are different algorithms for optimizing the same objective function.

Third, we give a detailed analysis about the NMF and PLSI solutions. They are local minima of the same landscape in a very high dimensional space. We show that PLSI can jump out of the local minima where NMF converges to and vice versa. Based on this, we further propose a hybrid algorithm to run NMF and PLSI alternatively to jump out a series of local minima and finally reach to a much better minimum. Extensive experiments show this hybrid algorithm improves significantly over the standard NMF-only or PLSI-only algorithms.

Data Representations of NMF and PLSI

Suppose we have n documents and m words (terms). Let $F = (F_{ij})$ be the word-to-document matrix: $F_{ij} = F(w_i, d_j)$ is the frequency of word w_i in document d_j .

In this paper, we re-scale the term frequency F_{ij} by $F_{ij} \leftarrow$

F_{ij}/T_w , where $T_w = \sum_{ij} F_{ij}$ is the total number of words. With this stochastic normalization, $\sum_{ij} F_{ij} = 1$. The joint occurrence probability $p(w_i, d_j) = F_{ij}$.

The general form of NMF is

$$F = CH^T, \quad (1)$$

where the matrices $C = (C_{ik}), H = (H_{jk})$ are nonnegative matrices. They are determined by minimizing

$$J_{\text{NMF}} = \sum_{i=1}^m \sum_{j=1}^n F_{ij} \log \frac{F_{ij}}{(CH^T)_{ij}} - F_{ij} + (CH^T)_{ij} \quad (2)$$

PLSI maximize the likelihood

$$\max J_{\text{PLSI}}, \quad J_{\text{PLSI}} = \sum_{i=1}^m \sum_{j=1}^n F_{ij} \log P(w_i, d_j) \quad (3)$$

where $P(w_i, d_j)$ is the factorized (i.e., parameterized or approximated) joint occurrence probability

$$P(w_i, d_j) = \sum_k p(w_i|z_k)p(z_k)p(d_j|z_k), \quad (4)$$

where the probability factors follow the normalization of probabilities

$$\sum_{i=1}^m p(w_i|z_k) = 1, \sum_{j=1}^n p(d_j|z_k) = 1, \sum_{k=1}^K p(z_k) = 1. \quad (5)$$

Equivalence of NMF and PLSI

In this section, we present our main results:

Theorem 1. PLSI and NMF are equivalent.

The proof is better described by the following

Proposition 1. The objective function of PLSI is identical to the objective function of NMF, i.e.,

$$\max J_{\text{PLSI}} \iff \min J_{\text{NMF}} \quad (6)$$

Proposition 2. Column normalized NMF of Eq.(1) is equivalent to the probability factorization of Eq.(4), i.e., $(CH^T)_{ij} = P(w_i, d_j)$.

Proof of Theorem 1: By Proposition 2, NMF (with L_1 -normalization, see §3) is identical to PLSI factorization. By Proposition 1, they minimize the same objective function. Therefore, NMF is identical to PLSI. \square

We proceed to prove Proposition 1 in this section.

Proof of Proposition 1:

First, we note that the PLSI optimization Eq.(3) can be written as $\min \sum_{i=1}^m \sum_{j=1}^n -F_{ij} \log P(w_i, d_j)$. Adding a constant, $\sum_{i=1}^m \sum_{j=1}^n F_{ij} \log F_{ij}$, PLSI is equivalent to solve

$$\min \sum_{i=1}^m \sum_{j=1}^n F_{ij} \log \frac{F_{ij}}{P(w_i, d_j)}.$$

Now since

$$\sum_{i=1}^m \sum_{j=1}^n [P(w_i, d_j) - F_{ij}] = [1 - 1] = 0,$$

we can add this constant to the summation; PLSI is equivalent to minimize

$$\sum_{i=1}^m \sum_{j=1}^n F_{ij} \log \frac{F_{ij}}{P(w_i, d_j)} - F_{ij} + P(w_i, d_j) \quad (7)$$

This is precisely the objective function for NMF. \square

NMF and χ^2 -statistic.

J_{NMF} of Eq.(2) has a somewhat complicated expression. It is related to the Kullback-Leibler divergence. We give a better understanding by relating it to the familiar χ^2 test in statistics. Assume $\frac{|(CH^T)_{ij} - F_{ij}|}{F_{ij}}$ is small. We can write

$$J_{\text{NMF}} = \sum_{i=1}^m \sum_{j=1}^n \frac{[(CH^T)_{ij} - F_{ij}]^2}{2F_{ij}} - \frac{[(CH^T)_{ij} - F_{ij}]^3}{3F_{ij}^2} + \dots \quad (8)$$

This is obtained by setting $\delta_{ij} = (CH^T)_{ij} - F_{ij}$, $z = \delta_{ij}/F_{ij}$, and $\log(1+z) = z - z^2/2 + z^3/3 \dots$; then the ij -th term in J_{NMF} becomes

$$\delta_{ij} - F_{ij} \log \left(1 + \frac{\delta_{ij}}{F_{ij}} \right) = \frac{1}{2} \frac{\delta_{ij}^2}{F_{ij}} - \frac{1}{3} \frac{\delta_{ij}^3}{F_{ij}^2} + \dots$$

Clearly, the first term in J_{NMF} is the χ^2 statistic,

$$\chi^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{[(CH^T)_{ij} - F_{ij}]^2}{F_{ij}}, \quad (9)$$

since F_{ij} is the data and $(CH^T)_{ij}$ is the model fit to it. Therefore, to first order approximation, NMF objective function is a χ^2 statistic. As a consequence, we can associate a confidence to NMF factorization.

The χ^2 form of NMF naturally relates to another NMF cost function, i.e., the sum of squared errors

$$J'_{\text{NMF}} = \sum_{i=1}^m \sum_{j=1}^n [(CH^T)_{ij} - F_{ij}]^2. \quad (10)$$

A comprehensive comparison among J_{NMF}, χ^2 and J'_{NMF} forms of NMF would be useful, but goes beyond the scope of this paper.

Normalizations of NMF

For any given NMF solution (C, H) , there exist a large number of matrices (A, B) such that $AB^T = I$, $CA \geq 0$, $HB \geq 0$. Thus (CA, HB) is also a solution with the same cost function value. Normalization is a way to eliminate this uncertainty. We mostly consider the normalization of columns of C, H . Specifically, let the columns be expressed explicitly, $C = (\mathbf{c}_1, \dots, \mathbf{c}_k)$, $H = (\mathbf{h}_1, \dots, \mathbf{h}_k)$.

¹In this column form, for clustering interpretation (Ding, He, & Simon 2005), \mathbf{c}_k is the centroid for k -th cluster, while \mathbf{h}_k is the posterior probability for k -th cluster. For hard clustering, on each row of H , set the largest element to 1 and the rest to 0.

