Revised Topic and Bibliography, second draft

Ziv Epstein

`ziv.epstein@pomona.edu`

November 5, 2016

1 Introduction to Topic Modeling

With the vast amount of digital text being generated across the internet, methods for understanding and processing corpora of human language become necessary. Across mathematics and computer science, many techniques have been put forward that allow one to understand a body of text far too large to read herself. A successful method in this domain is *topic modelling*, whereby semantically cohesive subgroups of words can be identified. In particular, let $\mathcal{C} = \{d_1, d_2, \cdots, d_n\}$ be a collection of documents with a vocabulary $\mathcal{V}$. A *topic* t_i is a vector over the words in the vocabulary that represents a coherent high level notion in the corpus:

$$t_i = \{v_1^i, v_2^i, \cdots, v_m^i\}$$

where m is the size of the vocabulary. Topic modelling offers a powerful tool for understanding large amounts of text because they can discover latent semantic structure within text.

There are two primary techniques for learning these topics t_i. The first is LDA, a generative Bayesian statistical model which views each document d_j as a mixture of various topics. The second is non-negative matrix factorization, which aims to factor the document/word matrix into a document/topic and a topic/word matrix [6]. The focus of this thesis will be NNMF, because of its relation to linear algebra, and its deep visual and conceptual intuition.

Non-negative matrix factorization was first employed by Paatero and Tapper [8] but was made popular by Lee and Seung [6] who suggested the importance of non-negative in human perception and first linked it to topic modeling.

Given our n documents with vocabulary $\mathcal{V}$ of size m, we construct a matrix $X \in \mathbb{R}^{n \times m}$ where $X_{i,j}$ is the number of occurences of word j in document i. For a given inner dimension k, we seek to factor X into two matrices A and S such that

$$X \approx AS$$

where $A \in \mathbb{R}^{n \times k}$ is the document/topic matrix and $S \in \mathbb{R}^{k \times m}$ is the topic/word matrix. When we impose that A and S must be non-negative, a strong intuition emerges. In particular, the (i, j)th entry of A corresponds to the proportion of topic j in document i and the (i, j)th entry of S corresponds to the relevance of word j in topic i. The problem of finding
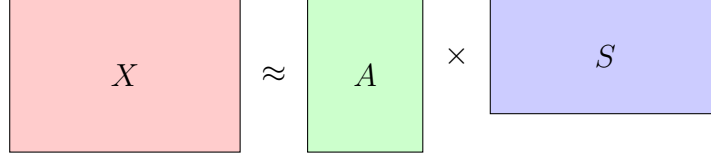
Figure 1: Figure 1: A visual representation of the non-negative matrix factorization

such a factorization can be formulated as finding a non-negative A and S that minimize the error

$$F = ||X - AS||^2 \tag{1}$$

While this optization problem is not convex in both A and S, it is convex in one of them. So for a given, fixed S, we can find the optimal A by setting the gradient equal to zero. Since $||X - AS||^2 = \langle X - AS, X - AS \rangle = X^T X - 2X^T AS + (AS)^T (AS)$ we have

$$\frac{\partial F}{\partial A}(X^T X - 2X^T AS + (AS)^T (AS)) = 0$$

$$\text{implies } S^T AS = 2X^T S$$

which is to say $\frac{X^T S}{S^T AS} = 1$ at the optimal A. This equality gives us the below multiplicative update algorithm.

Input: k=0; Initialize A^0, S^0

repeat

$$A^{k+1} = A^k \circ \frac{XS^k}{A^k(S^k)^T A^k}$$

$$S^{k+1} = S^k \circ \frac{XA^{k+1}}{S^k(A^{k+1})^T A^{k+1}}$$

$$k = k + 1$$

until *Stopping condition*;

Algorithm 1: Multiplicative Update

This optimization scheme naturally leads to a convex optimization function, so the above algorithm can simply be iteratively applied (until for given T we have $k > T$ or for a given $\epsilon > 0$ we have $||X - AS||^2 \leq \epsilon$).

Theorem 1. *The Euclidean distance $||X - AS||^2$ is non-increasing under the updating rules of Algorithm 1.*

2 Annotated Bibliography

We begin with the two foundation papers introducing the NNMF concept and standard algorithms. Lee and Seung [6] were the first to introduce this idea, and to propose that

topic modelling could be thought of as a matrix factorization problem. Ho [4] expands on this notion by elaborating on optimization schemas and corresponding algorithms that we will be taking advantage of. We then consider the work of Griffiths and Tenenbaum [3], who extened the notion of topic models to a hierarchical domain. We aim to replicate this structure but using an NNMF implementation instead of Latent Dirlichet Allocation (LDA).

Lee et al [7] introduces the idea of semi-supervised non negative matrix factorization, which is a key element of extending NNMF to labeled groups. The core of the thesis will be extending the work for Lee et al to hierarchical topic modeling.

Another foundational approach to hierarchical topic modelling is representing the topics not as a tree but rather as a deep matrix equation. Two papers explore this idea in the domain of NNMF [9, 1]. Trigerogis et al. consider semi-negative matrices for the domain to images and graph clustering. They are able to learn a "hierarchy" of features for a dataset of faces. Flenner and Hunter treat the deep topic representation as an analogized neural network, and use backpropogation as an optimization scheme to learn appropriate weights.

References

[1] Jennifer Flenner and Blake Hunter. A deep nonnegative matrix factorization.

[2] Santo Fortunato. Community detection in graphs. *Physics reports*, 486(3):75–174, 2010.

[3] DMBTL Griffiths and MIJJB Tenenbaum. Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16:17, 2004.

[4] Ngoc-Diep Ho. *Nonnegative matrix factorization algorithms and applications*. PhD thesis, ÉCOLE POLYTECHNIQUE, 2008.

[5] Andrea Lancichinetti, Santo Fortunato, and János Kertész. Detecting the overlapping and hierarchical community structure in complex networks. *New Journal of Physics*, 11(3):033015, 2009.

[6] Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[7] Hyekyoung Lee, Jiho Yoo, and Seungjin Choi. Semi-supervised nonnegative matrix factorization. *IEEE Signal Processing Letters*, 17(1):4–7, 2010.

[8] Pentti Paatero and Unto Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[9] George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, and Bjoern Schuller. A deep semi-nmf model for learning hidden representations. In *ICML*, pages 1692–1700, 2014.