Week 13

From this week, I will read some papers closely associated with my course project: Semi-supervised Network Representation Learning based on Multiview Dataset.
I plan to use Non-negative Matrix Factorization techniques to optimize the objective function that I propose in my project. Thus, here are three papers which help me understand more details about NMF:

- Paper One: Learning the parts of objects by non-negative matrix factorization
- Paper Two: Algorithms for Non-negative Matrix Factorization
- Paper Three: Non-negative Matrix Factorization in Multimodality Data for Segmentation and Label Prediction

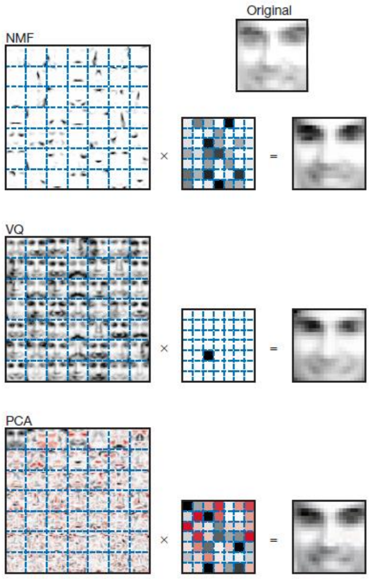
Paper One

Learning the parts of objects by non-negative matrix factorization Motivation

Main Theory
This paper is one of the earliest papers to study about NMF methods.
Authors show us the powerful effect of NMF according to the results on Facial representation learning. Without the mathematical details on how to solve NMF models, authors utilize it to do the experiments and give us their analysis.

From the paper, we can know that NMF is a little similar to but more different from VQ and PCA methods. PCA, VQ learn holistic, not part-based representations, but NMF is different from them via non-negative constrains, learning partial expressions.

As the following picture shows,



The 3 methods learn to represent a face as a linear combination of basis images. But they are different.

- VQ: discovers a basis consisting of whole-face prototypes.
- PCA: discovers a basis of ‘eigenfaces’ ,some of which resemble distorted versions of whole faces.
- NMF: discovers a basis consisting localized features that correspond better with intuitive notions of the parts of faces.

$$V_{i\mu} \approx (WH)_{i\mu} = \sum_{a=1}^{r} W_{ia} H_{a\mu}$$

Their goals are all learning the representations in H as well as the basic matric W. Authors also analysis more deep:

- In VQ, each column in H is constrained to be a unary vector, with one element equal to unity and the other elements equal to zero. Therefore, each face (column of V) is approximated by only single basis image (column of W).
- In PCA, it does some relaxation on VQ’s model. It constrains the columns of W to be orthonormal and the rows of H to be orthogonal to each other. It can allow a distributed representation in which each face is approximated by a linear combination of all the basis images, or eigenfaces. Although eigenfaces have a statistical interpretation as the directions of largest variance, PCA assign W and H randomly, thus making visual interpretation nonsense.

Unlike the unary constraint of VQ, NMF’s non-negativity constraints permit the combination of multiple basis images to represent a face. But only additive combinations are allowed, because the non-zero elements of Wand H are all positive.

To sum up, PCA enforces only a weak orthogonality constraint, resulting in a very distributed representation while VQ enforces a too hard constraint, resulting in clustering the data into mutually exclusive prototypes. NMF has a just proper constraint for matrix factorization that can learn a parts representation of data.

And authors also show their NMF’s results on representations of facial images and semantic analysis in the paper, which were quite better than the state-of-the-art methods at that time.

But there is no proof of NMF’s updating rules in this paper. Authors show the way how to solve the NMF problem in another paper as the following one, *Algorithms for Non-negative Matrix Factorization*.

Paper Two

Algorithms for Non-negative Matrix Factorization

Motivation

Authors have previously shown that NMF is a useful decomposition for a multivariate data. In this paper, they want to give us how NMF algorithms work and why they can work in this way with the detailed mathematical deduction.

Main theory

Unlike the last paper, authors did not put their attentions on NMF's advantages and the analysis why NMF has such advantages, but focused on the numerical algorithms on how to solve NMF problem for learning the optimal non-negative factors from data.

NMF: $V \approx WH$

There are two kinds of the Cost Functions:

$$||A-B||^2 = \sum_{ij}(A_{ij}-B_{ij})^2 \qquad D(A||B) = \sum_{ij}\left(A_{ij}\log\frac{A_{ij}}{B_{ij}} - A_{ij} + B_{ij}\right)$$

Thus we have two kinds of objectives:

$\min_{W,H} \|\mathrm{V} - WH\|^2$ ------- (1)

$\min_{W,H} D\left(V \,\|\, WH\right)$ ------- (2)

Obviously, they have different updating rules:

For (1),

$$H_{a\mu} \leftarrow H_{a\mu}\frac{(W^TV)_{a\mu}}{(W^TWH)_{a\mu}} \qquad W_{ia} \leftarrow W_{ia}\frac{(VH^T)_{ia}}{(WHH^T)_{ia}}$$

For (2),

$$H_{a\mu} \leftarrow H_{a\mu}\frac{\sum_i W_{ia}V_{i\mu}/(WH)_{i\mu}}{\sum_k W_{ka}} \qquad W_{ia} \leftarrow W_{ia}\frac{\sum_\mu H_{a\mu}V_{i\mu}/(WH)_{i\mu}}{\sum_\nu H_{a\nu}}$$

Since my course project is only related with the first kind of the objective function – (1), which is least-square-measure, I focus on the first kind as well here.

There is more than one mathematical proof of the updating rules. I show the one with Lagrange term:

Objective Function: $J = \frac{1}{2}\|V - WH\|$

$$\begin{aligned} J &= \frac{1}{2}tr((V-WH)(V-WH)^T) \\ &= \frac{1}{2}tr(VV^T - 2VH^TW^T + WHH^TW^T) \\ &= \frac{1}{2}(tr(VV^T) - 2tr(VH^TW^T) + tr(WHH^TW^T)) \end{aligned}$$

Let α_{ij} and β_{ij} be the Lagrange multiplier for constraint $w_{ij} \geq 0$ and $h_{ij} \geq 0$. we can construct L as: $L = \mathrm{J} + tr(\alpha W^T) + tr(\beta H)$

The derivatives of L with respect to W and H are:

$$\frac{\partial L}{\partial W} = -VH^T + WHH^T + \alpha$$

$$\frac{\partial L}{\partial H} = -V^T W + H^T W^T W + \beta$$

Using KTT condition $\alpha_{ij} w_{ij} = 0$ and $\beta_{ij} h_{ij} = 0$, we can get

$$W_{ij} \leftarrow W_{ij} \frac{(VH^T)_{ij}}{(WHH^T)_{ij}} \text{ and } H_{ij} \leftarrow H_{ij} \frac{(W^T V)_{ij}}{(W^T WH)_{ij}}$$

Paper Three

Non-negative Matrix Factorization in Multimodality Data for Segmentation and Label Prediction

Main Theory

In this paper, authors utilize NMF in image segmentation and multiview clustering of image features and image labels. This paper proposes the joint NMF updating rules based on the standard NMF updating rules in the last paper:

Original: $X \approx WH$ Goal: $\min_{W,H} \|X - WH\|^2$

Updating Rules: $W \leftarrow W \odot \frac{XH^T}{WHH^T}$ $H \leftarrow H \odot \frac{W^T X}{W^T WH}$

New: $X \approx WH \text{ and } Y \approx VH$

Authors want to utilize more information of different views from X and Y, thus constructing a new objective function:

$$\min_{W,V,H} (1-\lambda)\|X - WH\|^2 + \lambda\|Y - VH\|^2$$

Therefore, authors introduce joint NMF updating rules based on standard ones:

Standard: $W = W \odot \frac{XH^T}{WHH^T}$ $V = V \odot \frac{YH^T}{VHH^T}$

Joint: $H = H \odot \frac{(1-\lambda)W^T X + \lambda V^T Y}{((1-\lambda)W^T W + \lambda V^T V)H}.$

Thus, the objective function can even be proposed to more views or more terms like:

$$\min_{W^i,H} \sum_{i=1}^{p} \lambda_i \|X^i - W^i H\|^2$$

which means NMF can be utilize to embed different information matrices into same representations. That’s one interesting and useful thing of which I want to make full use in my project design.