MATP6600 Project Report

Saurabh Sihag (RIN: 661679538)

1 Introduction

Non-negative matrix factorization(NMF) is the factorization of a non-negative matrix $X \in \mathbb{R}^{m\times n}$ into a product of two non-negative matrices $W \in \mathbb{R}_+^{m\times r}$ and $H \in \mathbb{R}_+^{n\times r}$, such that

$$WH^\mathsf{T} = A\,. \tag{1}$$

In general, k can be set such that $k << \{m, n\}$, which allows for the use of NMF in applications that use high dimension of data, such as supervised learning applications, feature extraction, dictionary learning etc [1, 2, 3]. Without any constraints or requirements on the structure of W and H, the NMF can be represented by a simple optimization problem

$$\begin{aligned} &\underset{W,H}{\text{minimize}}\, \frac{1}{2}\|WH^\mathsf{T} - X\|_F^2 \\ &\text{s.t. } W \in \mathbb{R}_+^{m\times r}, H \in \mathbb{R}_+^{n\times r} \end{aligned} \tag{2}$$

Note that the NMF obtained by solving the above minimization problem is not unique. For example, for any diagonal matrix $D \in \mathbb{R}_+^{k\times k}$,

$$WH^\mathsf{T} = WDD^{-1}H^\mathsf{T} = (WD)(D^{-1}H)^\mathsf{T} = W'H'^\mathsf{T} \tag{3}$$

In order to incorporate certain properties (like sparsity) in the matrices W and H, the minimization problem in (2) can be modified to

$$\begin{aligned} &\underset{W,H}{\text{minimize}}\, \frac{1}{2}\|WH^\mathsf{T} - X\|_F^2 + \phi_1(W) + \phi_2(H) \\ &\text{s.t. } W \in \mathbb{R}_+^{m\times r}, H \in \mathbb{R}_+^{n\times r} \end{aligned} \tag{4}$$

where ϕ_1 and ϕ_2 are regularization functions. For example, using $\phi_1(W) = \alpha\|W\|_F^2$ and $\phi_2(H) = \beta\sum_{j=1}^{n}\|H(j,:)\|_1^2$ encourages sparsity in H and α preserves the accuracy of approximation [4]. Finding an exact solution to the NMF problem is NP-hard. Therefore, a heuristic approach is adopted in this project that consists of alternatively optimizing W for

a fixed H and H for a fixed W. This approach is formalized by the following:
Start with an initial guess W^0, H^0. Iteratively update W and H by

$$W^k = \underset{W \in \mathbb{R}_+^{m \times r}}{\arg\min} \frac{1}{2}\|W(H^{k-1})^\mathsf{T} - A\|_F^2 + \phi_1(W) \tag{5}$$

$$H^k = \underset{H \in \mathbb{R}_+^{n \times r}}{\arg\min} \frac{1}{2}\|(W^k)H^\mathsf{T} - A\|_F^2 + \phi_2(H) \tag{6}$$

for $k = 1, 2, \ldots$ until some stopping criterion is satisfied. In order to build an algorithm that can solve for W and H using the iterative algorithm above, the solutions to the individual optimization problems in (5) and (6) are provided first.

2 Task 1: Algorithm Problem (5)

In this project, $\phi_1(W)$ is chosen to be $\alpha\|W\|_1$, where $\|W\|_1 = \sum_{i,j} |w_{ij}|$ and $\alpha > 0$. Therefore, the algorithm to solve the following optimization problem is provided in this section. Let $A \in \mathbb{R}_+^{m \times p}$ and $B \in \mathbb{R}_+^{p \times n}$. Then, the least squares optimization problem is given by

$$\min_Z \|AZ - B\|_F^2 + \alpha\|Z\|_1, \quad \text{s.t.} \quad Z \in \mathbb{R}_+^{p \times n} \tag{7}$$

In this project, conjugate gradient method is applied to solve (7). Let $f(Z) = \|AZ - B\|_F^2 + \alpha\|Z\|_1$. The steps for conjugate gradient method are given below: The implementation of conjugate gradient method in this project is based on Algorithm 1. Note that the restart criterion in this algorithm is modified to $\arg\max_{i.j}\{|(\lambda d)_{ij})|\} < \epsilon_2$ where ϵ_2 is a small non zero constant and $|(\lambda d)_{ij})|$ is the absolute value of $(i,j)^{\text{th}}$ element of λd. Conjugate gradient method using $\beta = \frac{\|\nabla f(Z^1)\|^2}{\|\nabla f(Z^0)\|^2}$ is susceptible to jamming [5], i.e. the algorithm proceeds towards the minimum in very small steps. The non-zero constraint on elements of Z is incorporated by setting the non positive elements of Z to 0 at every iteration. However, this approximation to non negative least squares method may not always be optimal [6].

3 Task 2: Algorithm for Problem (6)

In this project, H is constrained by $\sum_{j=1}^r h_{ij} = 1, \forall i = \{1, \ldots, n\}$. Therefore, ϕ_2 can be chosen to be an indicator function $\iota_{\mathcal{H}}$, where

$$\mathcal{H} = \left\{ H : \sum_{j=1}^r h_{ij} = 1, \forall i = \{1, \ldots, n\} \right\}, \tag{8}$$

and $\iota_{\mathcal{H}} = 0$ when $H \in \mathcal{H}$ and $\iota_{\mathcal{H}} = \inf$ when $H \notin \mathcal{H}$. The algorithm to solve this optimization problem is provided in this section. The sum to one constrained least squares optimization

Algorithm 1 Conjugate Gradient Optimization

Initialize Z^0
Set $d = -\nabla f(Z_0)$
find $\lambda = \min_\lambda f(Z_0 + \lambda d)$
Set $Z^1 = Z^0 + \lambda d$
while $err > \epsilon$ **do**
    $\beta = \frac{\|\nabla f(Z^1)\|^2}{\|\nabla f(Z^0)\|^2}$
    $d = -\nabla f(Z^1) - \beta \nabla f(Z^0)$
    $\lambda = \min_\lambda f(Z^0 + \lambda d)$
    if $\arg\max_{i.j}\{|(\lambda d)_{ij})|\} < \epsilon_2$ **then**
        Jump to step 2
    end if
    $Z^1 = Z^0 + \lambda d$
    $Z^1(Z^1 < 0) = 0$ Setting non positive elements of Z to 0
    $err = \|f(Z^1) - f(Z^0)\|$
    $Z_0 = Z_1$
end while

problem is given by

$$\min_Z \frac{1}{2}\|AZ - B\|^2 \quad \text{s.t.} A \in \mathbb{R}_+^{m\times p}, B \in \mathbb{R}_+^{m\times n}, Z \in \mathbb{R}_+^{p\times n} \text{ and } \sum_{j=1}^{r} z_{ij} = 1, \forall i. \tag{9}$$

The optimization problem defined in (9) is solved using a projected gradient descent method. Let $f(Z) = \frac{1}{2}\|AZ - B\|^2$. The function $\text{proj}_{\mathcal{Z}}(Z)$ finds the projection of matrix Z in the

Algorithm 2 Projected Gradient Descent

Initialize Z^0
while $err > \epsilon$ **do**
    $d = -\nabla f(Z^0)$
    $\lambda = \arg\min_\lambda f(Z^0 + \lambda d)$
    $Z^1 = Z^0 + \lambda d$
    $Z^1 = \text{proj}_{\mathcal{Z}}(Z^1)$
    $err = \|f(Z^1) - f(Z^0)\|$
    $Z^0 = Z^1$
end while

space $\mathcal{Z}$ (defined similarly to $\mathcal{H}$). The projection function is designed on the basis of theory provided in [7].

4 Task 3: Algorithm for NMF

NMF for a matrix is performed by solving (5) and (6) iteratively. Since NMF is an approximation algorithm, the convergence of this heuristic approach to the optimal value is quite slow. Based on the evaluation results in [8, 9], it is observed that the normalized residuals converge as NMF converges. Therefore, in this project, the stopping criterion for the NMF is based on the convergence of normalized residuals. For given W and H, the normalized residual is defined as

$$r \triangleq \frac{0.5\|WH^{\mathsf{T}} - X\|_F^2 + \phi_1(W) + \phi_2(H)}{\|X\|^2} \ . \tag{10}$$

When the difference between the normalized residuals from two consecutive iterations drops below a certain tolerance value, the algorithm is stopped. This stopping criteria for NMF makes the algorithm to converge faster and provides good results for the application to a classification problem in the subsequent section.
For the results on Task 1, Task 2 and Task 3, the details are included in the Readme.txt file.

5 Application

The NMF algorthm developed in the previous sections is applied to the handwritten digit classification problem on USPS dataset (downloaded from [10]) . The algorithm used in this project is similar to the one used in [4].

5.1 Dataset Description

The dataset consists of 7291 training images and 2007 test images from 10 classes, each class representing the digits 0,1,. . . ,9. Each image is grayscale with size 16×16.

5.2 Classification Algorithm using NMF

Each digit is resized to form a 1×256 matrix. These 1×256 matrices from images corresponding to a class $i \in \{0, \ldots, 9\}$ are stacked together to form a matrix D_i, where D_i is a $n_i \times 256$ matrix and n_i is the number of images belonging to class i in the training dataset. For a given test digit d, we can find

$$q_i = \min_y \|D_i^{\mathsf{T}} y - d^{\mathsf{T}}\|^2 \ , \tag{11}$$

and the predicted class is given by $j = \arg\min_{i \in \{0,\ldots 9\}} \{q_i\}$. The matrices D_i can be factorized into $W_i H_i$, where the rows of H_i form the basis of the row space of D_i. Therefore,

$$D_i^{\mathsf{T}} y \approx H_i^{\mathsf{T}} W_i^{\mathsf{T}} y = H_i^{\mathsf{T}} x, x = W_i^{\mathsf{T}} y \ . \tag{12}$$

This implies that

$$\min_y \|D_i^\mathsf{T} y - d^\mathsf{T}\|^2 \approx \min_x \|H_i^\mathsf{T} x - d^\mathsf{T}\|^2 \ . \tag{13}$$

H_i is ensured to be sparse by the sum to one constraint and the accuracy in W_i is preserved by α. A summary of the algorithm is given below.

---

1. For $i \in \{0, \ldots, 9\}$, form D_i from training dataset.
2. Compute H_i by NMF of D_i.
3. For every sample d in test set, compute $q_i = \min_x \|H_i^\mathsf{T} x - d^\mathsf{T}\|^2$.
4. Predicted class for sample d is $j = \arg\min_{i \in \{0,\ldots 9\}} \{q_i\}$.

5.3 Results

The class-wise classification results are summarized below for $\alpha = 0.1$ and a rank-10 approximation using NMF.

Table 1: Error Rate for Different Classes

Digit	Training Sample Size	Test Sample Size	Number of Errors on Test set	Percentage Error
0	1194	391	6	1.53
1	1005	325	1	0.31
2	731	221	17	7.69
3	658	149	12	8.05
4	652	143	10	6.99
5	556	103	9	8.74
6	664	166	5	3.01
7	645	182	18	9.89
8	542	158	18	11.39
9	644	169	7	4.14

Average error rate for the setting in Table 1 is 5.13%. The error rates for different values of α are summarized below.

Table 2: Error Rate for Different α

α	Number of Errors	Error Rate (%)
0.5	102	5.08
1	139	6.93
2	190	9.47
3	205	10.21
4	221	11.01
5	222	11.06

The classification error is observed to uniformly increase with the increase in α and H seems to become less sparse.

6 Discussion

In this project, a heuristic approach to non-negative matrix factorization is implemented. Conjugate gradient optimization method is used for solving the non-negative least squares problem to find optimal W for a fixed H and a projected gradient method is used for optimizing H for a fixed W by solving another non-negative least squares problem. The constraint of having non-negative elements in W and H is incorporated by setting the respective negative elements to zero. This is an approximation to the non-negative least squares problem. Varying the parameter α and sum-to-one constraint incorporates sparsity structure in the factorization, which is particularly relevant in applications like classification. Since NMF provides an approximation to a matrix, using the convergence in the objective value directly in the stopping criterion proves to be time consuming. For this reason, convergence in normalized residuals is used as a stopping criterion.
The NMF implementation in this project is used for hand-written digits classification for USPS dataset. The classification error is observed to increase with increase in α, while the factor H seems to become less sparse with increase in α.

References

[1] D. Guillamet and J. Vitria, "Classifying faces with nonnegative matrix factorization," in *Proc. 5th Catalan conference for artificial intelligence*, 2002, pp. 24–31. 1

[2] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999. 1

[3] Z. Yuan and E. Oja, "Projective nonnegative matrix factorization for image compression and feature extraction." *Image analysis*, pp. 333–342, 2005. 1

[4] M. Mazack, "Non-negative matrix factorization with applications to handwritten digit recognition," Department of Scientific Computation, University of Minnesota, Tech. Rep., 2009. 1, 4

[5] W. W. Hager and H. Zhang, "A survey of nonlinear conjugate gradient methods," *Pacific journal of Optimization*, vol. 2, no. 1, pp. 35–58, 2006. 2

[6] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares," Georgia Institute of Technology, Tech. Rep., 2006. 2

[7] Y. Chen and X. Ye, "Projection onto a simplex." *arXiv preprint arXiv:1101.6081*, 2011. 3

[8] R. Zdunek, "Approximation of feature vectors in nonnegative matrix factorization with gaussian radial basis functions," *Neural Information Processing*, pp. 616–623, 2012. 4

[9] R. Gaujoux and C. Seoighe, "A flexible r package for nonnegative matrix factorization." *BMC bioinformatics*, vol. 11, no. 1, 2010. 4

[10] "http://www.cad.zju.edu.cn/home/dengcai/data/mldata.html." 4