Multiplicative Updates for Unsupervised and Contrastive Learning in Vision

Daniel D. Lee*, H. Sebastian Seung†, and Lawrence K. Saul*
*Univ. of Pennsylvania, Philaldelphia, PA 19104
†Mass. Inst. of Technology, Cambridge, MA 02138

Abstract. We describe two learning algorithms for unsupervised and supervised learning in image data. Both algorithms are distinguished by the use of nonnegative data and features in the models. In unsupervised learning, the likelihood is maximized under the appropriate nonnegativity constraints. For supervised learning, the conditional likelihood is maximized resulting in a constrastive objective function that directly optimizes discriminative performance. In both cases, multiplicative update rules are derived that have a simple closed form and interpretation. These update rules can also be shown to guarantee monotonic improvement in the appropriate objective functions without any adjustable tuning parameters. We illustrate the application of these algorithms on images of human faces and handwritten digits.

1 Introduction

In vision, the problem of learning from experience typically appears in two distinct scenarios. The first scenario arises when the learning algorithm is presented with image data that is unlabelled. In this "unsupervised" mode, the objective of the learning algorithm is to come up with some sort of compact description for the aggregate set of images. Methods that perform this type of unsupervised learning typically involve clustering and/or dimensionality reduction. Examples of such algorithms include vector quantization [4], self-organized maps [6], principal components analysis [5], independent components analysis [1], multidimensional scaling [2], local linear embedding [11], and others.

The second mode of learning occurs when image data is provided with categorical labels. By learning from these labelled examples, a "supervised" learning algorithm should be able to provide the correct labelling when presented with a new image. Pattern recognition is a traditional form of this problem, and specialized supervised learning algorithms have been developed to address this area of vision. Examples of methods that have been applied to these problems include backpropagation neural networks [7], radial basis functions [10], and support vector machines [13].

In these proceedings, we discuss novel learning rules that apply to hidden variable models for both unsupervised and supervised learning. These models incorporate nonnegativity constraints, which allow the derivation of simple multiplicative update rules for the parameters of the models. The update rules for unsupervised learning maximize the likelihood of the model, while the update rules for the supervised learning model directly optimize a discriminative objective function in a contrastive manner. Both learning algorithms have the virtue of guaranteed monotonic convergence to an optimum of the appropriate objective functions, without the need for any adjustable tuning parameters. We illustrate the application of these algorithms to learning from various image databases.

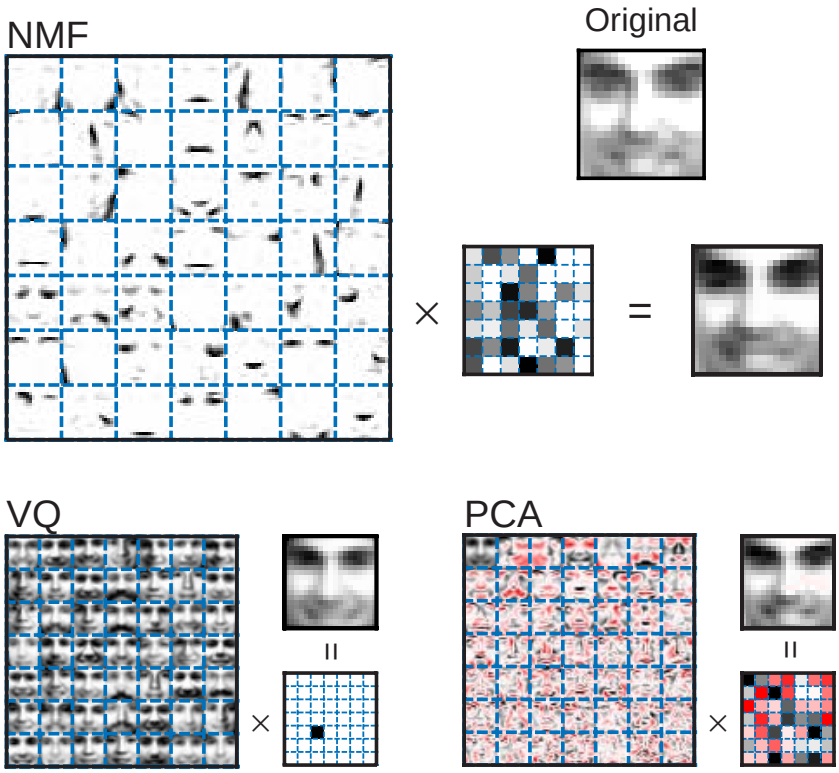

Figure 1: Basis vectors and feature coefficients for face images discovered by nonnegative matrix factorization (NMF), vector quantization (VQ) and principal components analysis (PCA).

2 Unsupervised Learning

We first describe the use of nonnegativity constraints in an architecture that models the variability of unlabelled images. A simple way to describe the algorithm is as a matrix factorization. A collection of grayscale images can formally be viewed as a large $D \times M$ matrix X of pixel values. Each of the M images is represented by a column vector of D pixel values within the matrix. Given this representation, many unsupervised learning algorithms can be viewed as an approximate factorization of the data matrix $X \approx BH$ where B is a $D \times N$ matrix and H is a $N \times M$ matrix. In this factorization, the columns of B are interpreted as basis vectors, and the columns of H as coefficients of features which represent the activations of hidden variables.

For example, in Figure 1, a database of $M = 2429$ facial images each consisting of $D = 19 \times 19$ pixels, is factorized with two standard unsupervised learning methods as well as our method. In all three cases, the number of learned bases is equal to $N = 49$. In principal components analysis (PCA) [5, 12], the factorization uses an orthogonal basis set. The PCA representation is able to efficiently capture much of the variability in the images using cancelling positive and negative linear combinations of the basis images. On the other hand, vector quantization (VQ), the other standard method shown in Fig. 1, involves a winner-take-all constraint that limits the representation to a set of prototypes that are individually replicated to model the image distribution.

In contrast to these standard techniques, nonnegative matrix factorization (NMF) uses nonnegativity constraints on the elements of the matrix factors B and H [8]. These constraints allow the representation to use additive combinations to model the variability of the face images. The nonnegativity constraint also forces a majority of the pixels in the basis set in B as well as the coefficients in H to zero. Thus, the nonnegativity constraint automatically

gives rise to a representation which is sparse and distributed [3].

The NMF algorithm involves iterating the following update rules:

$$B_{ij} \leftarrow B_{ij} \sum_k \frac{X_{ik}}{(BH)_{ik}} H_{jk}$$

$$H_{ij} \leftarrow H_{ij} \sum_k \frac{X_{kj}}{(BH)_{kj}} B_{ki}$$

These rules guarantee that the log likelihood $\mathcal{L} = \sum_{i=1}^{D} \sum_{k=1}^{M} [X_{ik} \log (BH)_{ik} - (BH)_{ik}]$ which describes the fidelity of the approximation $X \approx BH$ is monotonically optimized. The proof of convergence involves construction of an auxiliary function [9], but it is interesting to note that the update rules can also be viewed as rescaled gradient ascent. The updates multiply the current estimates for the parameters B and H by a quotient formed by taking the ratio between the positive and negative terms of the gradient of the objective function $\mathcal{L}$. When the gradient goes to zero, this multiplicative ratio goes to one indicating that the fixed point of the algorithm occurs at an optimium of the objective function. Thus, these simple multiplicative updates are able to efficiently optimize the likelihood while automatically preserving the nonnegativity constraints.

3 Contrastive Learning

Given a set of labels for images, a supervised learning algorithm should be able to learn an accurate mapping between the images and the labels. We show how the sparse, nonnegative features learned in NMF can be exploited by a discriminative mixture model. In this case, given the nonnegative features of an image, $\vec{h}$, the posterior distribution of the labels $\Pr[y|\vec{h}]$ is written as the following form:

$$\Pr[y = i|\vec{h}] = \frac{\sum_j W_{ij} \Phi_j(\vec{h})}{\sum_{k\ell} W_{k\ell} \Phi_\ell(\vec{h})}. \tag{1}$$

The right hand side of this equation defines a valid posterior distribution when the mixture weights W_{ij} and functions $\Phi_j(\vec{h}) = e^{\vec{\theta}_j \cdot \vec{h}}$ are nonnegative.

Our supervised learning algorithm directly optimizes the performance of Eq. (1) as a classifier. The objective function for discriminative training is the conditional log likelihood, $\mathcal{L}_C = \sum_k \log \Pr[y_k|\vec{h}_k]$ summed over all the training examples. If Y_{ki} is the binary matrix whose ki-th element denotes whether the k-th training example belongs to the i-th class, this objective function can be written as the difference of two terms, $\mathcal{L}_C = \mathcal{L}_+ - \mathcal{L}_-$, where:

$$\mathcal{L}_+ = \sum_k \log \sum_{ij} Y_{ki} W_{ij} e^{\vec{\theta}_j \cdot \vec{h}_k} \tag{2}$$

$$\mathcal{L}_- = \sum_k \log \sum_{ij} W_{ij} e^{\vec{\theta}_j \cdot \vec{h}_k}. \tag{3}$$

The competition between these two terms gives rise to contrastive learning, and distinguishes the algorithm from other algorithms such as Expectation-Maximization that maximize only the joint log likelihood.

The update rules for estimating the mixture coefficients W as well as the exponential parameters θ take the simple multiplicative form:

$$W_{ij} \leftarrow W_{ij} \left\{ \left(\frac{\partial \mathcal{L}_+}{\partial W_{ij}} \right) \Big/ \left(\frac{\partial \mathcal{L}_-}{\partial W_{ij}} \right) \right\}, \tag{4}$$

$$e^{\theta_{j\mu}} \leftarrow e^{\theta_{j\mu}} \left\{ \left(\frac{\partial \mathcal{L}_+}{\partial \theta_{j\mu}} \right) \Big/ \left(\frac{\partial \mathcal{L}_-}{\partial \theta_{j\mu}} \right) \right\}^{\frac{1}{\eta}} \text{ where } \eta = \max_n \sum_\mu H_{n\mu}. \tag{5}$$

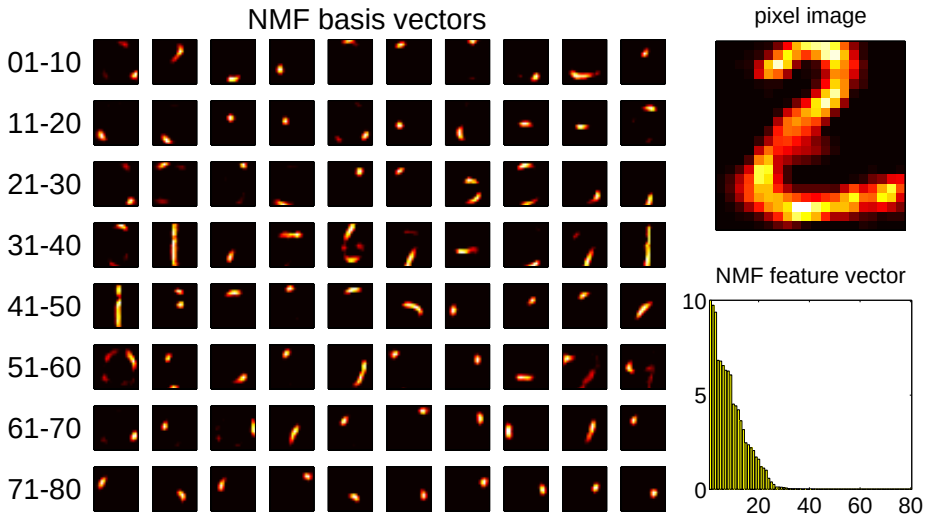

Figure 2: Nonnegative basis vectors for handwritten digits discovered by NMF, and the sparse feature vector for a handwritten "2".

Again, these updates have the simple intuition of forming a multiplicative ratio based upon opposing terms of the gradient of the conditional log likelihood. These multiplicative updates automatically enforce the required nonnegativity constraints, and are also guaranteed to converge monotonically to a maximum of $\mathcal{L}_C$. The rate of convergence for the exponential parameters θ is governed by the exponent $1/\eta$, which measures the sparseness of the features in H. Thus, using the nonnegative, sparse features discovered by NMF leads to faster learning in this discriminative model.

We illustrate the application of this algorithm in classifying handwritten digits. As shown in Fig. 2, we first use NMF to discover a sparse distributed representation of the MNIST data set of handwritten digits [7]. The data set contains 60000 training and 10000 test examples that were deslanted and cropped to form 20×20 grayscale pixel images. The left plot shows the $N = 80$ basis vectors discovered by NMF, and the right shows the sparseness of the representation in reconstructing a handwritten "2".

model	**EM-PCA40**		**EM-NMF80**		**CL-NMF80**	
K	$\mathcal{E}_t$	$\mathcal{E}_g$	$\mathcal{E}_t$	$\mathcal{E}_g$	$\mathcal{E}_t$	$\mathcal{E}_g$
1	10.2	10.1	15.7	14.7	5.5	5.8
2	8.5	8.3	12.3	10.7	4.0	4.4
4	6.8	6.4	9.3	8.2	2.8	3.5
8	5.3	5.1	7.8	7.0	1.7	3.2
16	4.0	4.4	6.2	5.7	1.0	3.4
32	3.1	3.6	5.0	5.1		
64	1.9	3.1	3.9	4.2		

Discriminative mixture models were then trained on these NMF features by the multiplicative updates in Eqs. (4–5). The results are shown in the table above, with baseline comparisons to mixture models trained using EM on NMF feature vectors as well as on PCA features. With only $K = 8$ mixture components per digit, the contrastive learning algorithm (CL-NMF80) achieves a training error of 1.7%, and a test error of 3.2%. This compares very favorably to comparably sized models trained by EM. The ability to learn a very compact classifier appears to be the major advantage of the contrastive learning algorithm. A slight

disadvantage is that the resulting classifiers are more susceptible to overtraining, as would be expected from discriminative training.

4 Discussion

It should be noted that better errors rates on the handwritten digit set have been obtained by support vector machines ($\mathcal{E}_g = 1.1\%$), k-nearest neighbor ($\mathcal{E}_g = 2.4\%$), and fully connected multilayer neural networks ($\mathcal{E}_g = 1.6\%$) [7]. However, these methods required storing large numbers of training examples or training significantly larger models. For example, the neural network classifier had over 120,000 weights. By contrast, the $K = 8$ discriminatively trained mixture model has less than 6500 iteratively adjusted parameters, most of which were learned in the unsupervised learning step of nonnegative matrix factorization.

The ability to learn such a compact discriminative model depends crucially on having nonnegative features. The nonnegativity constraints automatically give rise to a sparse, distributed representation that can be learned using simple update rules. These updates are guaranteed to converge monotonically to the appropriate objective functions, making the simplicity of the learning algorithms especially attractive.

References

[1] A. J. Bell and T. J. Sejnowski. An information maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[2] T. Cox and M. Cox. *Multidimensional Scaling*. Chapman & Hall, London, 1994.

[3] P. Foldiak and M. Young. *Sparse coding in the primate cortex*, pages 895–898. MIT Press, Cambridge, MA, 1995.

[4] A. Gersho and R. M. Gray. *Vector Quanization and Signal Compression*. Kluwer Academic, 1992.

[5] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.

[6] T. Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78:1464–1480, 1990.

[7] Y. LeCun, L. Jackel, L. Bottou, A. Brunot, C. Cortes, J. Denker, H. Drucker, I. Guyon, U. Muller, E. Sackinger, P. Simard, and V. Vapnik. A comparison of learning algorithms for handwritten digit recognition. In F. Fogelman and P. Gallinari, editors, *International Conference on Artificial Neural Networks*, pages 53–60, 1995.

[8] D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

[9] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural and Information Processing Systems*, volume 13, 2001.

[10] T. Poggio and F. Girosi. Regularization algorithms for learning that are equivalent to multilayer networks. *Science*, 247:978–982, 1990.

[11] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[12] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cogn. Neurosci.*, 3:71–86, 1991.

[13] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, 1999.