THE UNIVERSITY OF TENNESSEE KNOXVILLE

AICIP RESEARCH

ECE 599/692 – Deep Learning

Lecture 10 – Regularized AE and Case Studies

Hairong Qi, Gonzalez Family Professor
Electrical Engineering and Computer Science
University of Tennessee, Knoxville
http://www.eecs.utk.edu/faculty/qi
Email: hqi@utk.edu

AICIP RESEARCH

Outline

- Lecture 9: Points crossed
  - General structure of AE
  - Unsupervised
  - Generative model?
  - The representative power
    - Basic structure of a linear autoencoder
    - Denoising autoencoder (DAE)
  - AE in solving overfitting problem
- Lecture 10: Regularized AE and case studies
- Lecture 11: VAE leading to GAN

AICIP RESEARCH

The performance of machine learning methods is heavily dependent on the choice of data representation (or features) on which they are applied. … much of the actual effort in deploying machine learning algorithms goes into the design of preprocessing pipelines and data transformations that result in a representation of the data that can support effective machine learning. … Such feature engineering is important but labor-intensive and highlights the weakness of current learning algorithms: their inability to extract and organize the discriminative information from the data ….

[Bengio:2014]

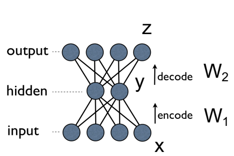
Different approaches

- Probabilistic models
  - Captures the posterior distribution of the underlying explanatory factors given observations
- Reconstruction-based algorithms
  - AE
- Geometrically motivated manifold-learning approaches

Issues

- Require strong assumptions of the structure in the data
- Make severe approximations, leading to suboptimal models
- Rely on computationally expensive inference procedures like MCMC

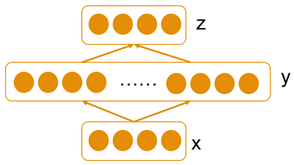
Basic structure of AE



$$\mathbf{y} = f_{\theta_1}(\mathbf{W}_1\mathbf{x} + \mathbf{b}_1)$$
$$\mathbf{z} = g_{\theta_2}(\mathbf{W}_2\mathbf{y} + \mathbf{b}_2)$$
$$\theta_1 = \{\mathbf{W}_1, \mathbf{b}_1\}, \theta_2 = \{\mathbf{W}_2, \mathbf{b}_2\}$$

$$\theta_1^*, \theta_2^* = \arg\min_{\theta_1,\theta_2} E_{q(X)}[L_H(X, g_{\theta_2}(f_{\theta_1}(X)))]$$

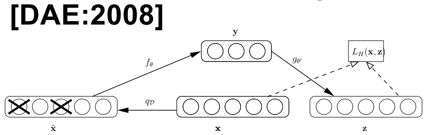
Regularized AE – Sparse AE



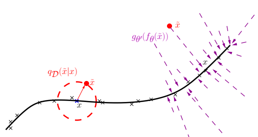
$$L(\theta_1, \theta_2) = \|X - g_{\theta_2}(f_{\theta_1}(X))\|_2^2 + \alpha_s h(f_{\theta_1}(X)) + \alpha_r \|W_1\|_1$$

7

Regularized AE – Denoising AE? [DAE:2008]



$$\theta_1^*, \theta_2^* = \arg\min_{\theta_1, \theta_2} E_{q(X,\tilde{X})}[L_H(X, g_{\theta_2}(f_{\theta_1}(\tilde{X})))]$$



The well-known link between "training with noise" and regularization

8

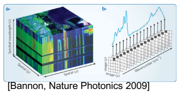
Application Example: AE in Spectral Unmixing

9

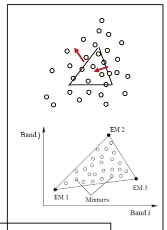
AICIP RESEARCH

Spectral Unmixing

- Mathematic formulation of HSI mixing: $x = As + n$
  - x: sensor readout
  - A: source matrix
  - s: abundance vector
  - n: measuring noise
- Constraints
  - Sum-to-one
  - Nonnegative
- Unsupervised unmixing
  - MVC-NMF
  - GDME



[Bannon, Nature Photonics 2009]



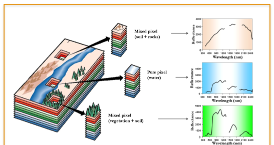
$$\text{minimize} \quad f(\mathbf{A},\mathbf{S}) = \frac{1}{2}\|\mathbf{X}-\mathbf{AS}\|_F^2 + \lambda J(\mathbf{A})$$
$$\text{subject to} \quad \mathbf{A} \geq 0,\ \mathbf{S} \geq 0,\ \mathbf{1}^T\mathbf{S} = \mathbf{1}^T$$

$$\text{minimize} \quad f_0(\mathbf{s}) = \sum_{j=1}^{c} s_j \ln s_j$$
$$\text{subject to} \quad h_0(\mathbf{s}) = \mathbf{1}^T s - 1 = 0,\ h_i(\mathbf{s}) = \sum_{j=1}^{c} A_{ij} s_j - x_i = 0,\ i = 1,\ldots,N$$

[1] L. Miao, H. Qi, "Endmember extraction from highly mixed data using minimum volume constrained non-negative matrix factorization," *IEEE Transactions on Geoscience and Remote Sensing*, 45(3):765-777, March 2007. (Highest Impact Paper Award)
[2] L. Miao, H. Qi, H. Szu, "A maximum entropyapproach to unsupervised mixed pixel decomposition," *IEEE Transactions on Image Processing*, 16(4):1008-1021, April 2007.

AICIP RESEARCH

Unmixing and DL?

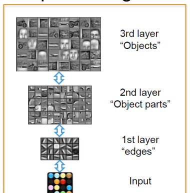


[Ciznicki,SPIE 2012]

Spectral Unmixing

A good marriage?

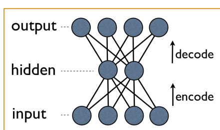
Deep Learning

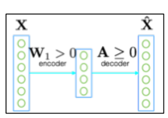


[CVPR 2012 Tutorial by Honglak Lee]

AICIP RESEARCH

AutoEncoder (AE)





How to formulate the unmixing problem in the framework of AutoEncoder?

1. What is endmember?
2. What is abundance?
3. How to incorporate the constraints?
4. What are the constraints?

X = **AS** vs. ^**X** = **A**σ(**W**$_1$**X**)

Part-based AE assumes

- Tied weight: $\mathbf{W}_1 = \mathbf{A}^T$
- Nonnegative weight:

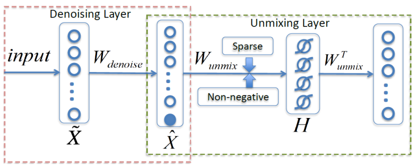
$$\mathbf{A} \geq 0 \quad \mathbf{W}_1 \geq 0$$

[1]Chen, Minmin, et al. "Marginalized denoising autoencoders for domain adaptation." arXiv preprint arXiv:1206.4683 (2012).
[2] Lemme, Andre, René Felix Reinhart, and Jochen Jakob Steil. "Efficient online learning of a non-negative sparse autoencoder." ESANN. 2010.

AICIP RESEARCH

AutoEncoder Cascade

- Marginalized Denoising AutoEncoder (mDA)
- Non-negative Sparse AutoEncoder (NNSA)



Rui Guo, Wei Wang, Hairong Qi, "Hyperspectral image unmixing using cascaded autoencoder," *IEEE Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensor (WHISPERS)*, Tokyo, Japan, June 2-5, 2015. (Best Paper Award)

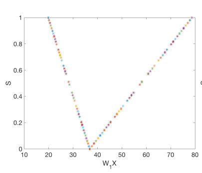
THE UNIVERSITY OF TENNESSEE

AICIP RESEARCH

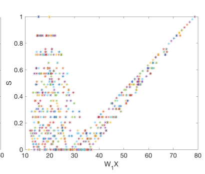
Discussion on the Network Structure

- On tied weight

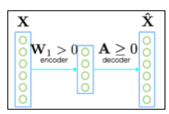
X = **AS** = **A**σ(**W**$_1$**X**) = **W**$_1^T$σ(**W**$_1$**X**)
That is, **S** = σ(**W**$_1$**X**)



$\mathbf{W}_1 \in \mathbb{R}^{2\times188}$



$\mathbf{W}_1 \in \mathbb{R}^{3\times188}$



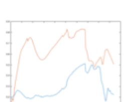
THE UNIVERSITY OF TENNESSEE

AICIP RESEARCH

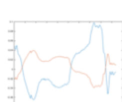
Discussion on the Network Structure (cont'd)

- On the nonnegative constraint

X = **AS** = **A**σ(**W**$_1$**X**)
Assume σ is linear: **AW**$_1$=**I**



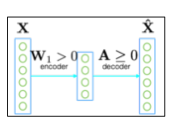
$\mathbf{A} \in \mathbb{R}^{188\times2}$



$\mathbf{W}_1 \in \mathbb{R}^{2\times188}$

$\mathbf{A} \in \mathbb{R}^{188\times3}$

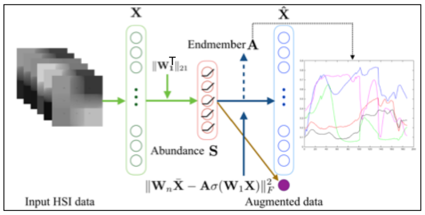
$\mathbf{W}_1 \in \mathbb{R}^{3\times188}$



THE UNIVERSITY OF TENNESSEE

Proposed Algorithm

AICIP RESEARCH



$$J(\mathbf{W}_1, \mathbf{A}) = \frac{1}{2}\|\mathbf{A}\sigma(\mathbf{W}_1\mathbf{X}) - \mathbf{X}\|_F^2 + \alpha\|\mathbf{W}_n\bar{\mathbf{X}} - \mathbf{A}\sigma(\mathbf{W}_1\mathbf{X})\|_F^2 + \beta\|\mathbf{W}_1\|_{21} \quad \text{subject to} \quad \mathbf{A} \geq 0$$

THE UNIVERSITY OF TENNESSEE

Proposed Algorithm

AICIP RESEARCH

- Objective function: $$J(\mathbf{W}_1, \mathbf{A}) = \frac{1}{2}\|\mathbf{A}\sigma(\mathbf{W}_1\mathbf{X}) - \mathbf{X}\|_F^2 + \alpha\|\mathbf{W}_n\bar{\mathbf{X}} - \mathbf{A}\sigma(\mathbf{W}_1\mathbf{X})\|_F^2 + \beta\|\mathbf{W}_1\|_{21} \quad \text{subject to} \quad \mathbf{A} \geq 0$$
- The denoising constraint:
  - multilayer scaled marginalized denosing autoencoder (mDA): $\mathbf{W}_n$
- Nonnegativity: $\mathbf{A} = \max(\mathbf{A} - \alpha\nabla\mathbf{A}, \mathbf{0})$
- Sum-to-one: $\bar{\mathbf{X}} = \begin{bmatrix} \hat{\mathbf{X}} \\ \delta\mathbf{1}_n^T \end{bmatrix} \quad \bar{\mathbf{A}} = \begin{bmatrix} \mathbf{A} \\ \delta\mathbf{1}_c^T \end{bmatrix}$

THE UNIVERSITY OF TENNESSEE