The neural autoregressive distribution estimator

Hugo Larochelle and Iain Murray

CS 598 presentation by Varun Kelkar

Outline

- Motivation
  - What is the paper trying to do?
  - Why is the problem important?
  - Why is the problem difficult?
- Prior approaches
  - Mixture of Bernoullis
  - Restricted Boltzmann machines
  - Bayesian Networks
- Approach
- Numerical studies
- Summary

Motivation

- *What the paper wishes to achieve:*

  Distribution estimation of high dimensional discrete/binary vectors.

Motivation

- *What the paper wishes to achieve:*

  Distribution estimation of high dimensional discrete/binary vectors.

- *Why is this problem important:*

  If one knows the joint distribution of objects, one can potentially begin to answer any question about the dependencies between them, including all of supervised learning.

Motivation

- *What the paper wishes to achieve:*

  Distribution estimation of high dimensional discrete/binary vectors.

- *Why is this problem important:*

  If one knows the joint distribution of objects, one can potentially begin to answer any question about the dependencies between them, including all of supervised learning.

- *Why is this problem difficult:*

  Curse of dimensionality, the PMF is a vector in a d^n-dimensional space, where d is the number of discrete levels, n is the dimensionality of the vector.

Previous approaches

- Mixture of Bernoullis (MoB)
- Restricted Boltzmann Machines (RBMs)
- Bayesian Networks
- Fully visible sigmoid belief network (FVSBN)

Restricted Boltzmann Machines (RBMs)

$$\mathbf{h} = \mathbf{W}\mathbf{v} + \mathbf{b}$$

Probabilities evaluated using the energy function:

$$E(\mathbf{v}, \mathbf{h}) = -\mathbf{h}^\top \mathbf{W}\mathbf{v} - \mathbf{b}^\top \mathbf{v} - \mathbf{c}^\top \mathbf{h} \tag{1}$$

probabilities are assigned to any observation $\mathbf{v}$ as follows:

$$p(\mathbf{v}) = \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}))/Z \,, \tag{2}$$

Problems:

- Computing partition function *Z* is intractable for all except the small networks. Approximations needed.
- Hence, RBMs cannot be used to model parts of a probabilistic system.
- Difficulty in evaluating the learned distribution

Bayesian networks

Strategy: Decompose the distribution using its conditionals

$$p(\mathbf{v}) = \prod_{i=1}^{D} p(v_i | \mathbf{v}_{\mathbf{parents}(i)}) \,, \tag{3}$$

Example: Fully visible sigmoid belief networks

$$p(v_i | \mathbf{v}_{\mathbf{parents}(i)}) = \mathrm{sigm}\Big(b_i + \sum_{j<i} W_{ij} v_j\Big) \,, \tag{4}$$

Converting RBMs into Bayesian networks

Rewrite the RBM PDF in terms of conditionals

$$\begin{aligned} p(\mathbf{v}) &= \prod_{i=1}^{D} p(v_i|\mathbf{v}_{<i}) \\ &= \prod_{i=1}^{D} p(v_i, \mathbf{v}_{<i})/p(\mathbf{v}_{<i}) \\ &= \prod_{i=1}^{D} \frac{\sum_{\mathbf{v}_{>i}} \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}))}{\sum_{\mathbf{v}_{j \geq i}} \sum_{\mathbf{h}} \exp(-E(\mathbf{v}, \mathbf{h}))}\,, \end{aligned} \tag{5}$$

Use a simplified model for the still intractable conditionals

$$\begin{aligned} q(v_i, \mathbf{v}_{>i}, \mathbf{h}|\mathbf{v}_{<i}) = \mu_i(i)^{v_i}(1-\mu_i(i))^{1-v_i} \\ \prod_{j>i} \mu_j(i)^{v_j}(1-\mu_j(i))^{1-v_j} \\ \prod_{k} \tau_k(i)^{h_k}(1-\tau_k(i))^{1-h_k}\,, \end{aligned}$$

Converting RBMs into Bayesian networks

Minimize the KL divergence by setting its gradients to 0, which gives the following. Use fixed-point iterations to find the parameters of the distribution:

$$\tau_k(i) = \text{sigm}\left(c_k + \sum_{j \geq i} W_{kj}\mu_j(i) + \sum_{j<i} W_{kj} v_j \right) \quad (7)$$

$$\mu_j(i) = \text{sigm}\left(b_j + \sum_{k} W_{kj}\tau_k(i) \right) \quad \forall j \geq i \,. \quad (8)$$

Problems:

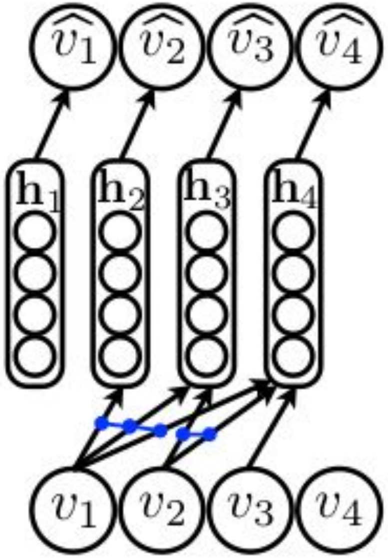
- Can be slow to converge.
- Needs to be repeated for each component v_i

Neural autoregressive distribution estimators (NADE)

Taking inspiration from the first fixed-point iteration above, formulate the network architecture.

$$p(v_i = 1|\mathbf{v}_{<i}) = \mathrm{sigm}\left(b_i + (\mathbf{W}^\top)_{i,\cdot}\mathbf{h}_i\right) \qquad (9)$$

$$\mathbf{h}_i = \mathrm{sigm}\left(\mathbf{c} + \mathbf{W}_{\cdot,<i}\mathbf{v}_{<i}\right), \qquad (10)$$



NADE

Training: Minimize the log-likelihood averaged over a training dataset

$$\frac{1}{T}\sum_{t=1}^{T} -\log p(\mathbf{v}_t) = \frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{D} -\log p(v_i|\mathbf{v}_{<i}), \quad (11)$$

Results of experiments

Experiments comparing various baselines using the average test log-likelihood (ALL) relative to the MoB baseline

Model	ADULT	CONNECT-4	DNA	MUSHROOMS	NIPS-0-12	OCR-LETTERS	RCV1	WEB
MoB	0.00 ± 0.10	0.00 ± 0.04	0.00 ± 0.53	0.00 ± 0.10	0.00 ± 1.12	0.00 ± 0.32	0.00 ± 0.11	0.00 ± 0.23
RBM	4.18 ± 0.06	0.75 ± 0.02	1.29 ± 0.48	-0.69 ± 0.09	12.65 ± 1.07	-2.49 ± 0.30	-1.29 ± 0.11	0.78 ± 0.20
RBM mult.	4.15 ± 0.06	-1.72 ± 0.03	1.45 ± 0.40	-0.69 ± 0.05	11.25 ± 1.06	0.99 ± 0.29	-0.04 ± 0.11	0.02 ± 0.21
RBForest	4.12 ± 0.06	0.59 ± 0.02	1.39 ± 0.49	0.04 ± 0.07	12.61 ± 1.07	3.78 ± 0.28	0.56 ± 0.11	-0.15 ± 0.21
FVSBN	**7.27 ± 0.04**	11.02 ± 0.01	**14.55 ± 0.50**	4.19 ± 0.05	13.14 ± 0.98	1.26 ± 0.23	-2.24 ± 0.11	0.81 ± 0.20
NADE	**7.25 ± 0.05**	**11.42 ± 0.01**	13.38 ± 0.57	**4.65 ± 0.04**	**16.94 ± 1.11**	**13.34 ± 0.21**	**0.93 ± 0.11**	**1.77 ± 0.20**
Normalization	-20.44	-23.41	-98.19	-14.46	-290.02	-40.56	-47.59	-30.16

Generative performance for binarized images

Figure 2: **(Left)**: samples from NADE trained on a binary version of MNIST. **(Middle)**: probabilities from which each pixel was sampled. **(Right)**: visualization of some of the rows of **W**. This figure is better seen on a computer screen.

Summary

- MoBs not sufficiently expressive to model the complex dependencies in high dimensional distributions
- RBMs can have an intractable computation of the partition function, rendering it difficult to use for downstream applications
- Bayesian networks, such as fully visible sigmoid belief networks may still be less expressive than desired.
- Approaches that convert RBMs to Bayesian networks are slow to converge.
- The proposed approach outperforms other approaches because it is able to utilize the recursiveness in Bayesian network architectures to decompose a complex probability distribution to a tractable form, while still having sufficient generality in the form of the network architecture chosen.

References

1. Larochelle, Hugo, and Iain Murray. "The neural autoregressive distribution estimator." *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011.
2. Fischer, Asja, and Christian Igel. "An introduction to restricted Boltzmann machines." *Iberoamerican congress on pattern recognition*. Springer, Berlin, Heidelberg, 2012.