Deep Generative Models

Sargur N. Srihari
srihari@cedar.buffalo.edu

Topics

1. Boltzmann machines
    2. Restricted Boltzmann machines
    3. Deep Belief Networks
    4. Deep Boltzmann machines
    5. Boltzmann machines for continuous data
    6. Convolutional Boltzmann machines
    7. Boltzmann machines for structured and sequential outputs
    8. Other Boltzmann machines
9. Backpropagation through random operations
10. Directed generative nets
11. Drawing samples from autoencoders
12. Generative stochastic networks
13. Other generative schemes
14. Evaluating generative models
15. Conclusion

Overview

- We describe several specific generative models
  - That can be built and trained using techniques of :
    - PGMs, Monte Carlo methods, Partition Functions, Approximate Inference
  - All models represent probability distributions in some way
    - Some allow probability distribution to be evaluated explicitly
    - Others do not allow distribution to be evaluated but allow operations such as sampling
    - Some are described by graphs and factors, others not

1. Boltzmann Machines

- Introduced for learning arbitrary probability distributions over binary vectors
- Variants include other kinds of variables
  - Surpassed popularity of the original
- First we consider binary Boltzmann machines and discuss their training and inference

Binary Boltzmann Machine

- We define a Boltzmann machine over a d-dimensional binary vector $\boldsymbol{x} \in \{0,1\}^d$
- Boltzmann machine is an energy-based model that defines the joint probability distribution

$$\boxed{P(\boldsymbol{x}) = \frac{\exp(-E(\boldsymbol{x}))}{Z}}$$

  – Energy function $E(\boldsymbol{x})$ is defined by $E(\boldsymbol{x}) = \boldsymbol{x}^T U \boldsymbol{x} - b^T \boldsymbol{x}$
    - where U is the weight matrix of model parameters and b is the vector of bias parameters

  – Z is the partition function that ensures

$$\sum_{x} P(\boldsymbol{x}) = 1$$

Boltzmann Machine as Linear Model

- Boltzmann machine is a joint probability distribution over observed variables

$$\boxed{P(\boldsymbol{x}) = \frac{\exp(-E(\boldsymbol{x}))}{Z}}$$

$$E(\boldsymbol{x}) = \boldsymbol{x}^T U \boldsymbol{x} - b^T \boldsymbol{x}$$

- It defines a distribution where the probability of a given unit being on is determined by a linear model (logistic regression) of the other variables
- In the general setting we are given training examples over all the variables

Boltzmann with hidden units

- Boltzmann becomes more powerful when not all variables are observed
- Just as hidden units convert logistic regression to MLP (a universal approximater of functions)
  - No longer limited to modeling linear relationships between variables
  - Boltzmann becomes universal approximater of probability mass functions over discrete variables
  - Units $\boldsymbol{x}$ are decomposed into visible units $\boldsymbol{v}$ and hidden units $\boldsymbol{h}$. Energy function becomes

$$E(\boldsymbol{v},\boldsymbol{h}) = -\boldsymbol{v}^T R \boldsymbol{v} - \boldsymbol{v}^T W \boldsymbol{h} - \boldsymbol{h}^T S \boldsymbol{h} - \boldsymbol{b}^T V - \boldsymbol{c}^T H$$

Boltzmann Machine Learning

- Usually based on maximum likelihood
- All Boltzmann machines have an intractable partition function
  - So max. likelihood gradient has to be approximated
- Interesting property:
  - Update for a particular weight connecting two units depends only on the statistics of the two units collected under different distributions
    - $P_{\text{model}}(\boldsymbol{v}),\ P_{\text{data}}(\boldsymbol{v}),\ P_{\text{model}}(\boldsymbol{h}|\boldsymbol{v})$
    - The rest of model shapes those statistics
    - This means learning rule is “local” which makes Boltzmann learning biologically plausible

Biological Plausibility

- If each neuron were a random variable in a Boltzmann machine
  - Then axons and dendrites connecting two variables could learn only by observing the firing patterns of variables that they touch
  - Positive phase: two units that frequently fire together have their connection strengthen
  - This is an example of Hebbian learning rule
    - Oldest hypothesized biological learning