Inductive Principles for Restricted Boltzmann Machine Learning
Benjamin Marlin, Kevin Swersky, Bo Chen and Nando de Freitas

Inductive Principles for Restricted Boltzmann Machine Learning

Benjamin Marlin
Department of Computer Science
University of British Columbia

Joint work with Kevin Swersky, Bo Chen and Nando de Freitas
Introduction: The Big Picture

Some facts about maximum likelihood estimation:

- ML is consistent (asymptotically unbiased)
- ML is statistically efficient (asymptotically lowest error)
- For certain model classes, computing the likelihood function can be computationally intractable.

This work studies alternative inductive principles for restricted Boltzmann machines that circumvent the computational intractability of the likelihood function at the expense of statistical consistency and/or efficiency.
Outline:

• Boltzmann Machines and RBMs
• Inductive Principles
  • Maximum Likelihood
  • Contrastive Divergence
  • Pseudo-Likelihood
  • Ratio Matching
  • Generalized Score Matching
• Experiments
• Demo
Introduction: Boltzmann Machines

A Boltzmann Machine is a Markov Random Field on D binary variables defined through a quadratic energy function.

$$E_\theta(x) = -(x^T W x + x^T b)$$

$$P_\theta(x) = \frac{1}{\mathcal{Z}} \exp(-E_\theta(x))$$

$$\mathcal{Z} = \sum_{x' \in \mathcal{X}} \exp(-E_\theta(x'))$$
Introduction: Restricted Boltzmann Machines

- A Restricted Boltzmann Machine (RBM) is a Boltzmann Machine with a bipartite graph structure.

- Typically one layer of nodes are fully observed variables (the visible layer), while the other consists of latent variables (the hidden layer).
Introduction: Restricted Boltzmann Machines

- The joint probability of the visible and hidden variables is defined through a bilinear energy function.

\[
E_\theta(x, h) = -(x^T W h + x^T b + h^T c)
\]

\[
P_\theta(x, h) = \frac{1}{\mathcal{Z}} \exp \left( -E_\theta(x, h) \right)
\]

\[
\mathcal{Z} = \sum_{x' \in \mathcal{X}} \sum_{h' \in \mathcal{H}} \exp \left( -E_\theta(x', h') \right)
\]
Introduction: Restricted Boltzmann Machines

- The bipartite graph structure gives the RBM a special property: the visible variables are conditionally independent given the hidden variables and vice versa.

\[
P_\theta(x_d = 1|h) = \frac{1}{1 + \exp\left(-\left(\sum_{k=1}^{K} W_{dk} h_k + x_d b_d\right)\right)}
\]

\[
P_\theta(h_k = 1|x) = \frac{1}{1 + \exp\left(-\left(\sum_{d=1}^{D} W_{dk} x_d + h_k c_k\right)\right)}
\]
Introduction: Restricted Boltzmann Machines

- The marginal probability of the visible vector is obtained by summing out over all joint states of the hidden variables.

\[ P_\theta(x) = \frac{1}{\mathcal{Z}} \sum_{h \in \mathcal{H}} \exp(-E_\theta(x, h)) \]

- This sum can be carried out analytically yielding an equivalent model defined in terms of a “free energy”.

\[ P_\theta(x) = \frac{1}{\mathcal{Z}} \exp(-F_\theta(x)) \]

\[ F_\theta(x) = - \left( x^T b + \sum_{k=1}^{K} \log \left( 1 + \exp \left( x^T W_k + c_k \right) \right) \right) \]
Introduction: Restricted Boltzmann Machines

• This construction eliminates the latent, hidden variables, leaving a distribution defined in terms of the visible variables.

• However, computing the normalizing constant (partition function) still has exponential complexity in D.

\[ Z = \sum_{\mathbf{x}' \in \mathcal{X}} \exp \left( -F_\theta(\mathbf{x}') \right) \]

• This work is about inductive principles for RBM learning that circumvent the intractability of the partition function.
Inductive Principles for Restricted Boltzmann Machine Learning
Benjamin Marlin, Kevin Swersky, Bo Chen and Nando de Freitas

Outline:

• Boltzmann Machines and RBMs
• Inductive Principles
  • Maximum Likelihood
  • Contrastive Divergence
  • Pseudo-Likelihood
  • Ratio Matching
  • Generalized Score Matching
• Experiments
• Demo
Stochastic Maximum Likelihood

• Exact maximum likelihood learning is intractable in an RBM. Stochastic ML estimation can instead be applied, usually using a simple block Gibbs sampler.

\[ f^{ML}(\theta) = \sum_{\boldsymbol{x} \in \mathcal{X}} P_e(\boldsymbol{x}) \log P_\theta(\boldsymbol{x}) \]

\[ \nabla f^{ML} \approx -\left( \frac{1}{N} \sum_{n=1}^{N} \nabla F_\theta(\boldsymbol{x}_n) - \frac{1}{S} \sum_{s=1}^{S} \nabla F_\theta(\tilde{\boldsymbol{x}}_s) \right) \]
**Contrastive Divergence**

- The contrastive divergence principle results in a gradient that looks identical to stochastic maximum likelihood. The difference is that CD samples from the T-step Gibbs distribution.

\[
f^{CD}(\theta) = \sum_{\mathbf{x} \in \mathcal{X}} P_e(\mathbf{x}) \log \left( \frac{P_e(\mathbf{x})}{P_\theta(\mathbf{x})} \right) - Q_\theta^t(\mathbf{x}) \log \left( \frac{Q_\theta^t(\mathbf{x})}{P_\theta(\mathbf{x})} \right)
\]

\[
\nabla f^{CD} \approx -\frac{1}{N} \left( \sum_{n=1}^{N} \nabla F_\theta(\mathbf{x}_n) - \nabla F_\theta(\mathbf{\tilde{x}}_n) \right)
\]
Pseudo-Likelihood

- The principle of maximum pseudo-likelihood is based on optimizing a product of one-dimensional conditional densities under a log loss.

\[ f^{PL}(\theta) = \sum_{\mathbf{x} \in \mathcal{X}} \sum_{d=1}^{D} P_e(\mathbf{x}) \log P_\theta(x_d|\mathbf{x}_{-d}) \]

\[ \nabla f^{PL} = -\frac{1}{N} \sum_{n,d} P_\theta(\mathbf{x}^{d}_{dn}|\mathbf{x}_{-dn}) \left( \nabla F_\theta(\mathbf{x}_n) - \nabla F_\theta(\mathbf{x}^{d}_n) \right) \]
Ratio Matching

- The ratio matching principle is very similar to pseudo-likelihood, but is based on minimizing a squared difference between one dimensional conditional distributions.

\[
f^{RM}(\theta) = \sum_{\mathbf{x} \in \mathcal{X}} \sum_{d=1}^{D} \sum_{\xi \in \{0,1\}} P_e(\mathbf{x}) \left( P_{\theta}(X_d = \xi | \mathbf{x}_{-d}) - P_e(X_d = \xi | \mathbf{x}_{-d}) \right)^2
\]

\[
\nabla f^{RM} = \frac{2}{N} \sum_{n=1}^{N} \sum_{d=1}^{D} g(u_{dn})^3 u_{dn} \left( \nabla F_\theta(\mathbf{x}_n) - \nabla F_\theta(\mathbf{x}_n^d) \right)
\]

\[
g(u) = \frac{1}{1 + u}, \quad u_{dn} = \frac{P_\theta(\mathbf{x}_n)}{P_\theta(\mathbf{x}_n^d)}
\]
Generalized Score Matching

• The generalized score matching principle is similar to ratio matching, except that the difference between inverse one dimensional conditional distributions is minimized.

\[
f^{GSM}(\theta) = \sum_{x \in \mathcal{X}} \sum_{d=1}^{D} P_e(x) \left( \frac{1}{P_\theta(x_d|x_{-d})} - \frac{1}{P_e(x_d|x_{-d})} \right)^2
\]

\[
\nabla f^{GSM} = \frac{2}{N} \sum_{n=1}^{N} \sum_{d=1}^{D} (u_{dn}^{-2} - u_{dn}) \left( \nabla F_\theta(x_n) - \nabla F_\theta(x_n^d) \right)
\]

\[
g(u) = u^{-2} - 2u, \quad u_{dn} = \frac{P_\theta(x_n)}{P_\theta(x_n^d)}
\]
Gradient Comparison

\[ \nabla f^{ML} \approx - \left( \frac{1}{N} \sum_{n=1}^{N} \nabla F_{\theta}(x_n) - \frac{1}{S} \sum_{s=1}^{S} \nabla F_{\theta}(\bar{x}_s) \right) \]

\[ \nabla f^{CD} \approx - \frac{1}{N} \left( \sum_{n=1}^{N} \nabla F_{\theta}(x_n) - \nabla F_{\theta}(\bar{x}_n) \right) \]

\[ \nabla f^{PL} = \frac{-1}{N} \sum_{n,d} P_{\theta}(\bar{x}_{dn}^d|x_{-dn}) \left( \nabla F_{\theta}(x_n) - \nabla F_{\theta}(\bar{x}_{dn}^d) \right) \]

\[ \nabla f^{RM} = \frac{2}{N} \sum_{n=1}^{N} \sum_{d=1}^{D} g(u_{dn})^3 u_{dn} \left( \nabla F_{\theta}(x_n) - \nabla F_{\theta}(\bar{x}_{dn}^d) \right) \]

\[ \nabla f^{GSM} = \frac{2}{N} \sum_{n=1}^{N} \sum_{d=1}^{D} (u_{dn}^{-2} - u_{dn}) \left( \nabla F_{\theta}(x_n) - \nabla F_{\theta}(\bar{x}_{dn}^d) \right) \]
Outline:

• Boltzmann Machines and RBMs
• Inductive Principles
  • Maximum Likelihood
  • Contrastive Divergence
  • Pseudo-Likelihood
  • Ratio Matching
  • Generalized Score Matching
• Experiments
• Demo
Experiments:

Data Sets:
- MNIST handwritten digits
- 20 News Groups
- CalTech 101 Silhouettes

Evaluation Criteria:
- Log likelihood (using AIS estimator)
- Classification error
- Reconstruction error
- De-noising
- Novelty detection
Experiments: Log Likelihood

(a) MNIST

(b) 20News

(c) CalTech
Experiments: Classification Error

(a) MNIST
(b) 20News
(c) CalTech
Experiments: De-noising

(a) MNIST  
(b) 20News  
(c) CalTech

---

Department of Computer Science, University of British Columbia
Experiments: Novelty Detection

(a) MNIST  
(b) 20News  
(c) CalTech
Experiments: Learned Weights on MNIST

(a) CD
(b) SML
(c) PL
(d) RM
Discussion:

• As the underlying theory suggests, SML obtains the best test set log likelihoods.

• CD and SML perform very well on MNIST classification, which is consistent with past results.

• Ratio matching obtains the best de-noising results, which is consistent with the observation on MNIST that the filters have more local structure than those produced by SML, CD, and PL.

• PL and RM are actually slower than CD and SML due to the need to consider all one-neighbors for each data case.