Feature Extraction Using Restricted Boltzmann Machine On The MNIST Dataset

Speaker:Ming-Chun Wu

Institute of Statistical Science
Academia Sinica

November 4, 2016

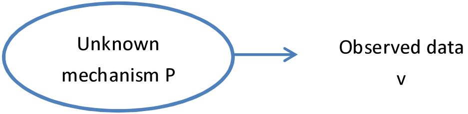
OUTLINE

- Unsupervised learning using generative models
- Restricted Boltzmann Machine (RBM)
- Training RBM: from maximum likelihood to contrastive divergence
- Examples

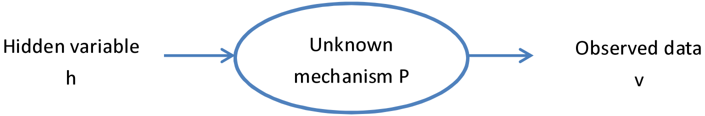
Generative Model

In unsupervised learning

- Q: What are the underlying natures of the data?



- Introducing hidden variable h



- Can we estimate/learn P from the data?

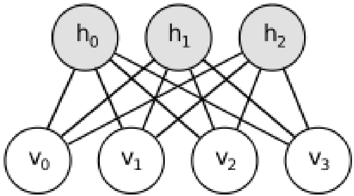
Estimate the Generative Model

Basic framework to estimate the generative model

- Goal: estimate the unknown mechanism P with data v
- Hidden variables h
- Assume joint distribution $P_\theta(v, h)$ with unknown parameters θ
- Marginalization: $P_\theta(v) = \sum_h P_\theta(v, h)$
- Fit model $P_\theta(v)$ to the data v
  - Different "Goodness of fit" with different loss function
  - ML: $\max_\theta \sum_{data} P_\theta(v)$
  - KL: $\min_\theta D(\bar{P}||P_\theta)$ where $\bar{P}$ is the empirical distribution
  - ML $\equiv$ KL
- Next step: How to model $P_\theta(v, h)$?

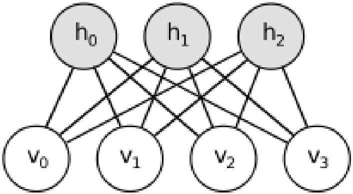
Restricted Boltzmann Machine

- Visible units/data v and hidden variables h
- Energy based model: $P_\theta(v,h) = \frac{e^{-\mathcal{E}(h,v)}}{Z}$
- Normalization: $Z = \sum_{h,v} e^{-\mathcal{E}(h,v)}$
- Boltzmann machine:
  $\mathcal{E}(v,h) = -b'v - c'h - h'Wv - v'Uv - h'Vv$
- Restricted Boltzmann machine:
  $\mathcal{E}(v,h) = -b'v - c'h - h'Wv$
- Goal: estimate the unknown parameters $\theta = \{b, c, W\}$



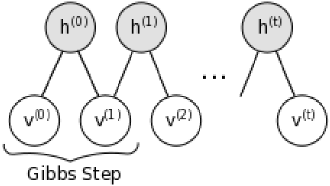
Training RBM

- Recall: ML $\equiv$ minimize KL $\equiv$ fit to the data
- Goal: $\hat{\theta} = \arg\max_{\theta} \log P_\theta(v, h)$
- Gradient descent: $\theta^{(t+1)} \leftarrow \theta^{(t)} + \eta \sum_{v \in Data} [\nabla \log P_{\theta^{(t)}}(v, h)]$
- Q: How to compute $\nabla \log P_\theta(v, h)$?
- $\nabla \log P_\theta(v, h) = -\nabla \mathcal{F}(v) + \sum_v P_\theta(v) \nabla \mathcal{F}(v)$
- Computable free energy: $\mathcal{F}(v) = -b'v - \sum_i \log \left[\sum_{h_i} e^{h_i(c_i + W_i h_i)}\right]$



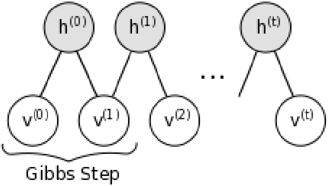
Markov Chain Monte Carlo

- $\nabla \log P_\theta(v, h) = -\nabla \mathcal{F}(v) + E_{P_\theta(v)}[\nabla \mathcal{F}(v)]$
- Law of large number $\frac{1}{N} \sum_{v_n} \nabla \mathcal{F}(v_n) \to E_{P_\theta(v)}[\nabla \mathcal{F}(v)]$
- Monte Carlo estimator: generate $v_n \overset{i.i.d.}{\sim} P_\theta(v)$
- MCMC: Markov chain with stationary distribution $\sim P_\theta(v)$
- Gibbs sampling, why?
  - In RBM, $P(h|v) = \prod_i P(h_i|v), P(v|h) = \prod_i P(v_i|h)$
  - $P(h_i = 1|v) = sigmoid(c_i + W_i v)$



Contrastive Divergence

- When will the Markov Chain be stationary?
- Require many samples $\frac{1}{N}\sum_{v_n} \nabla \mathcal{F}(v_n)$ to estimate the mean
- CD-k
  1. Initial the Markov chain with a data point $v^{(0)} \leftarrow v$
  2. k-step Gibbs chain $v^{(k)}$
  3. One sample approximation
     $\nabla \log P_\theta(v, h) = -\nabla \mathcal{F}(v) + E_{P_\theta(v)}[\nabla \mathcal{F}(v)]$
     $\nabla \log P_\theta(v, h) = \nabla \log P_\theta(v^{(0)}, h) \approx -\nabla \mathcal{F}(v^{(0)}) + \nabla \mathcal{F}(v^{(k)})$



Contrastive Divergence

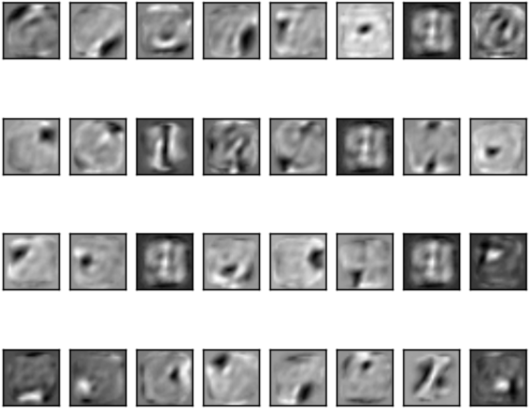
Why CD-k works?

- Gibbs chain: $v^{(0)} \to h^{(0)} \to v^{(1)} ... \to v^{(k)}$
- $\nabla \log P_\theta(v, h) \approx -\nabla \mathcal{F}(v^{(0)}) + \nabla \mathcal{F}(v^{(k)})$
- $\nabla \log P_\theta(v, h) = -\nabla \mathcal{F}(v^{(0)}) + E_{v^k|v^{(0)}}[\nabla \mathcal{F}(v^{(k)})] + E_{v^k|v^{(0)}}[\nabla \log P_\theta(v^{(k)})]$
- Mean of the CD-k estimator $-\nabla \mathcal{F}(v^{(0)}) + E_{v^k|v^{(0)}}[\nabla \mathcal{F}(v^{(k)})]$
- Bias $E_{v^k|v^{(0)}}[\nabla \log P_\theta(v^{(k)})] \to 0$ as $k \to \infty$
- When stationary, bias $= 0$!

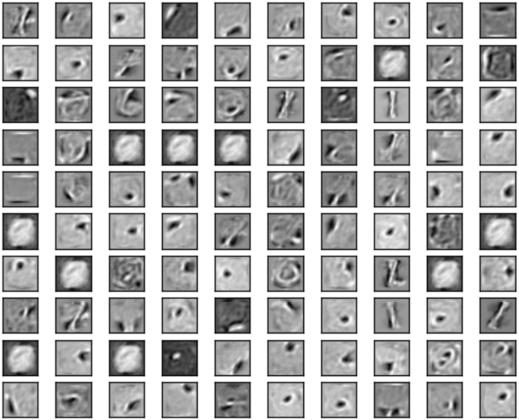
Algorithm of RBM-CD-k

1. For all data point $v \in Data$, update θ as follows
2. CD-k
   - Initial a Markov chain $v^{(0)} \leftarrow v$
   - Target distribution $P_{\theta^{(t)}}(v, h)$
   - Gibbs sampling $v^{(0)} \rightarrow h^{(0)} \rightarrow v^{(1)} \dots \rightarrow v^{(k)}$
   - Approximate gradient $\nabla \log P_{\theta^{(t)}}(v, h) \approx -\nabla \mathcal{F}_{\theta^{(t)}}(v^{(0)}) + \nabla \mathcal{F}_{\theta^{(t)}}(v^{(k)})$
3. Update parameters $\theta^{(t+1)} \leftarrow \theta^{(t)} + \eta \hat{\nabla} \log P_{\theta^{(t)}}$
4. Before convergence, go to step 1

RBM Features of the MNIST Dataset



RBM Features of the MNIST Dataset



Question?

Thank You!