RBM and DBN

赵惜墨

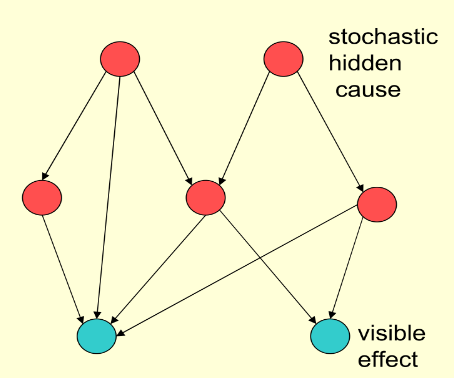
哈尔滨工业大学
计算机学院
智能技术与自然语言处理实验室

November 6, 2013

优点

1. 保留了 bp 利用梯度方法调整权重的有效性、简单性
2. 学习 $P(image)$ 而不是 $P(label|image)$

belief nets



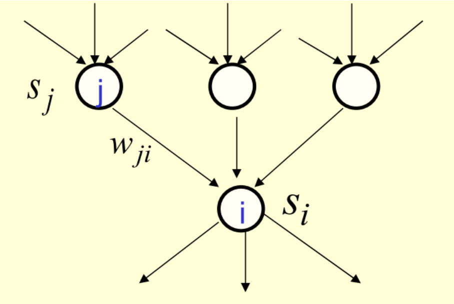
1. 置信网由一个具有随机变量的有向图构成
2. 通过观测可见节点，解决以下两个问题
    1. 推理问题：解决未观察到的节点的状态
    2. 学习问题：通过调整可见节点相互之间的关系使网络产生更正确的观测变量

表示、学习

logistic belief net 由二元随机变量组成。

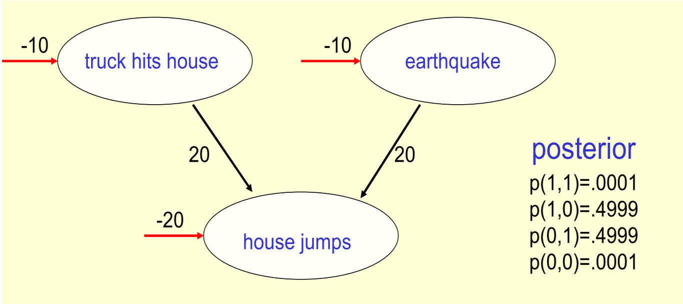
$$p(s_i = 1) = \frac{1}{1 + \exp(-b_i - \sum_j s_j w_{ji})}$$

$$\Delta W_{ji} = \epsilon s_j (s_i - p_i)$$



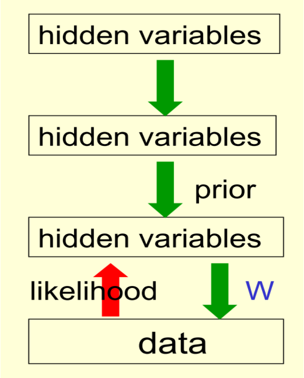
explaining away

即使两个隐含变量独立，在两个变量都能影响到的事件上，它们也能变得相关。
发生地震减弱了发生卡车把房子撞了



explaining away 使有向图推理更困难了。

一次学习一层参数所带来的问题



- 为了学习 W（权重），需要学习第一层隐含层的后验分布

  问题 1 由于 “explaining away”，学习会变得非常复杂

  问题 2 后验既依赖于先验也依赖于似然，所以为了学习（W），需要知道更高层的参数，所有的权重都相关。

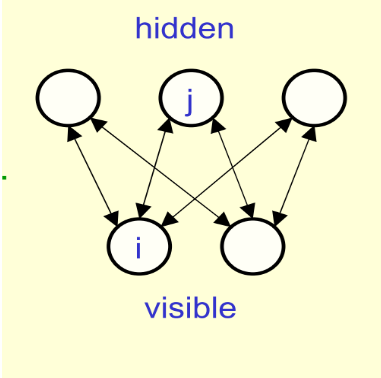
可以采用的学习方法

- MCMC：费时
- 变分：不准确

DBN 所带来的突破

1. 为了高效的学习，需要一次学一层。但是在假设隐含变量独立的情况写学习效果不好。
   - 隐含变量后验分布不独立导致对于非线性模型的推理十分的困难。
   - 在学习的过程中，算法在隐含层中寻找独立的解释，但是实际情况并非如此。
2. 为了解决这些问题，引入了无向图模型。

RBM



- 对于连接进行限制，只有一层
- 隐含层之间没有联系
- 给定可见节点，隐含层间节点相互独立

能量

$$E(v,h) = -\sum_{i,j} v_i h_j w_{ij}$$

$$-\frac{\partial E(v,h)}{\partial w_{ij}} = v_i h_j$$

能量 -> 概率

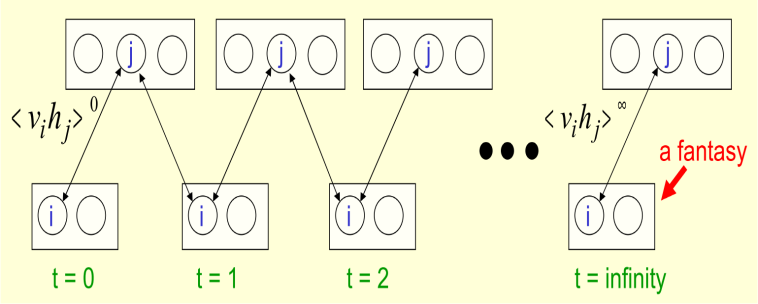
1. 每一个可能的可见节点和隐含节点的组合都有一个能量
2. 能量决定概率

$$p(v, h) = \frac{e^{-E(v,h)}}{\sum_{u,g} e^{-E(u,g)}}$$

3. 可见节点的概率是所有含有该可见节点的组合的加和

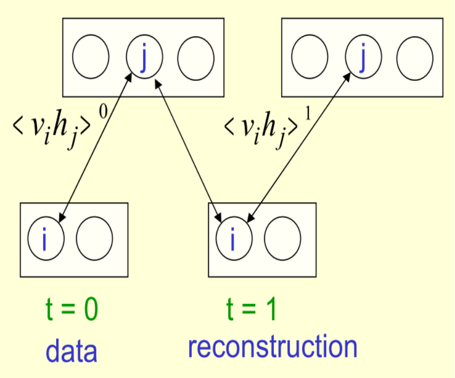
$$p(v) = \frac{\sum_h e^{-E(v,h)}}{\sum_{u,g} e^{-E(u,g)}}$$

MLE learning for RBM



$$\frac{\partial \log p(v)}{\partial w_{ij}} = < v_i h_j >^0 - < v_i h_j >^\infty$$

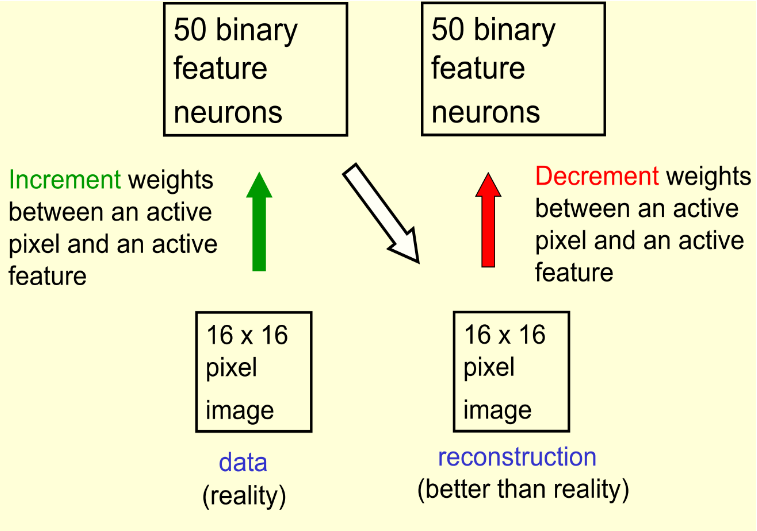
quick way



- 从观测节点开始
- 更新隐含节点
- 重建可见节点
- 再次更新隐含节点

$$\Delta w_{ij} = \epsilon(< v_i h_j >^0 - < v_i h_j >^1)$$

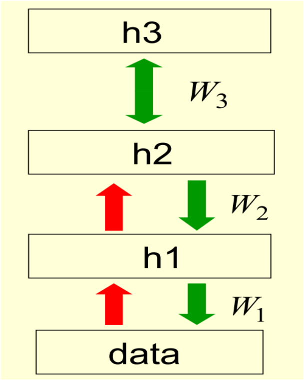
例子


50 binary feature neurons
50 binary feature neurons
Increment weights between an active pixel and an active feature
Decrement weights between an active pixel and an active feature
16 x 16 pixel image
16 x 16 pixel image
data
(reality)
reconstruction
(better than reality)

训练

1. 先训练一个隐含层，从可见层直接接收数据
2. 在将隐含层的输出作为下一层的可见节点，在进行训练

产生数据

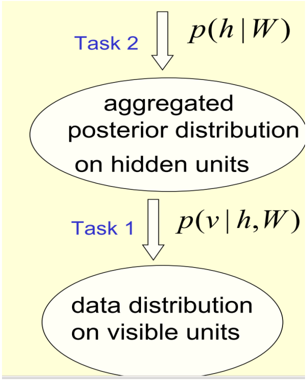


得到数据

- 在最上层进行 Gibbs sampling，得到 h_2
- 对于其他层，根据分布 $P(h^{k-1}|h^k)$ 进行采样
- 最后得到的 $x = h^0$ 即为所得

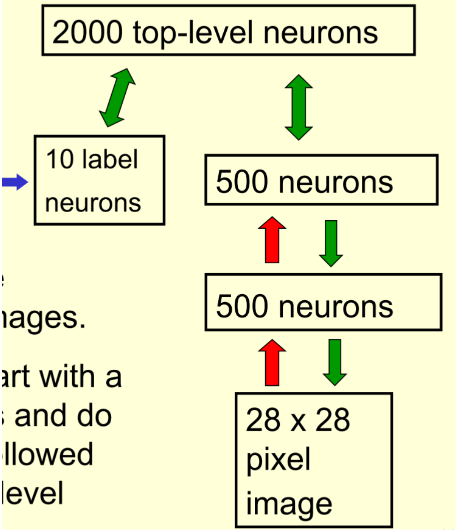
最后几个不是产生式模型的一部分，只是用来进行 inference 的

为什么学习方法有效



- RBM 将数据分布转移到隐含层分布
- 将任务转化成两步：学习 $P(h|W), P(v|h, W)$
- 通过第二步 modeling 数据更容易，因为其更贴近 RBM 所能 model 的
- RBM 能表示为 $P(v) = \sum_h p(h)p(v|h)$，对于 $P(h)$ 提高，自然可以提高 $P(v)$

例子


2000 top-level neurons
10 label neurons
500 neurons
500 neurons
28 x 28 pixel image
ages.
art with a
and do
llowed
level