Restricted Boltzmann Machines

Bo Bernhardsson

Department of Automatic Control LTH,
Lund University

Fun stuff before we get started

A journey trough all the layers of an artificial neural network.

How deep dream works

Fun stuff before we get started

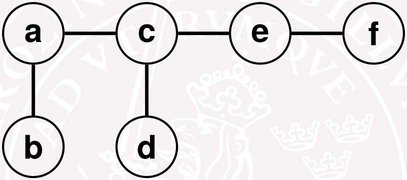
Deep Dream version

Agenda

- Structured Probabilistic Models [Ch. 16]
- RBMs [Ch. 20.1-2]
- Training RBMs [parts of Ch 18]
- Exercise

Watch the linked videos for more information

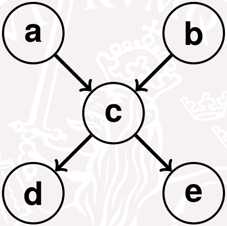
Undirected Graphs



$$p(a,b,c,d,e,f) = \frac{1}{Z}\phi_{ab}(a,b)\phi_{ac}(a,c)\phi_{cd}(c,d)\phi_{ce}(c,e)\phi_{ef}(e,f)$$

The normalisation constant Z ('partition function') often hard to compute

Directed Graphs



$$p(a,b,c,d,e) = \frac{1}{Z} p_a(a)\, p_b(b)\, p_c(c \mid a,b)\, p_d(d \mid c)\, p_e(e \mid c)$$

Conditional Independence

a and b are independent iff

$$p(a, b) = p_a(a)p_b(b)$$

a and b are conditionally independent given s iff

$$p((a, b) \mid s) = p(a \mid s)p(b \mid s)$$

Conditional Independence - undirected graphs

When does graph imply $\mathcal{A}$ is independent of $\mathcal{B}$ given observations in $\mathcal{S}$?

Conditional Independence - undirected graphs

When does graph imply $\mathcal{A}$ is independent of $\mathcal{B}$ given observations in $\mathcal{S}$?

a and b are dependent if s is unobserved (path is ”activate”)

Conditional Independence - undirected graphs

When does graph imply $\mathcal{A}$ is independent of $\mathcal{B}$ given observations in $\mathcal{S}$?

a and b are dependent if s is unobserved (path is ”activate”)

a and b become independent if s is observed (path is ”unactivated”)

Conditional Independence - undirected graphs

When does graph imply $\mathcal{A}$ is independent of $\mathcal{B}$ given observations in $\mathcal{S}$?

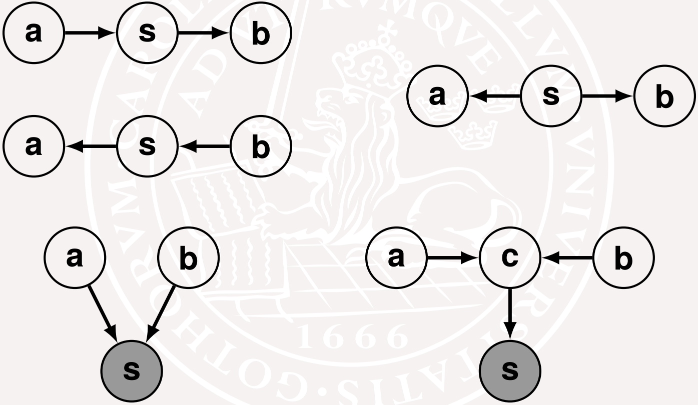
a and b are dependent if s is unobserved (path is ”activate”)

a and b become independent if s is observed (path is ”unactivated”)

a and b are independent if no active path between them exists

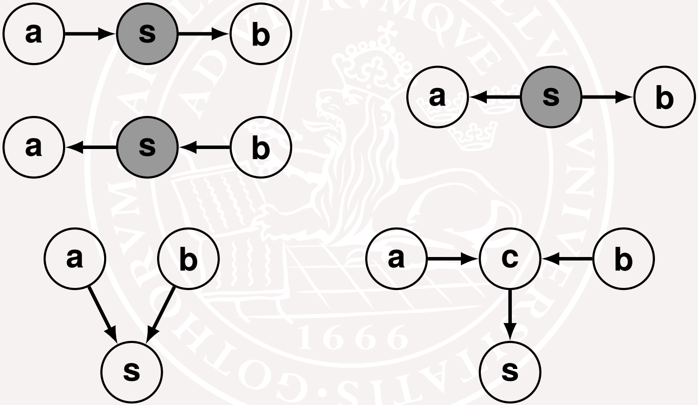
Conditional Independence - Directed Graphs

Examples where a and b are dependent (grey variables observed)



Conditional Independence - Directed Graphs

In these figures a and b are independent (grey variables observed)



Restricted Boltzmann Machine

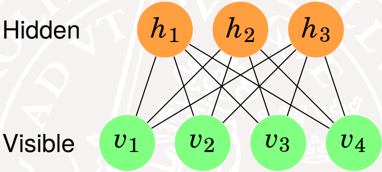
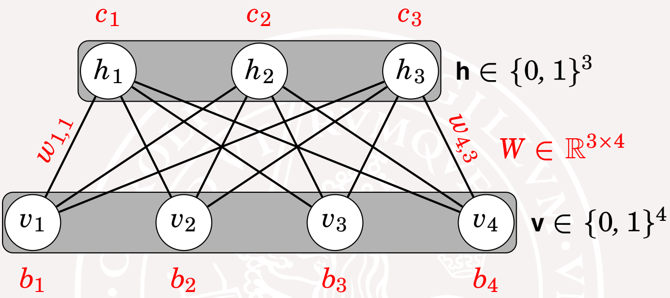


Figure: Restricted Boltzmann topology with 3 hidden units and 4 visible units.

Restricted = No intralayer connections (i.e. graph is bipartite)

h-nodes are independent given the v-nodes (and vice versa)



Popular modeling assumption: The probability distribution has the form

$$p(v, h) = \frac{1}{Z} \exp\left(-E(v, h)\right)$$
$$E(v, h) = -b^T v - c^T h - v^T W h \quad \text{("Energy function")}$$

Structure of p for RBMs

Conditional independence between h variables if v observed

$$p(h \mid v) = \prod_j p(h_j \mid v) \quad \text{and} \quad p(v \mid h) = \prod_i p(v_i \mid h)$$

Probability distributions ”factorize”

Structure of p for RBMs

Conditional independence between h variables if v observed

$$p(h \mid v) = \prod_j p(h_j \mid v) \quad \text{and} \quad p(v \mid h) = \prod_i p(v_i \mid h)$$

Probability distributions ”factorize”

When h and v are binary vectors then

$$\frac{p(h_j = 1 \mid v)}{p(h_j = 0 \mid v)} = \exp(c_j + v^T W_{:,j})$$

Structure of p for RBMs

Conditional independence between h variables if v observed

$$p(h \mid v) = \prod_j p(h_j \mid v) \quad \text{and} \quad p(v \mid h) = \prod_i p(v_i \mid h)$$

Probability distributions ”factorize”

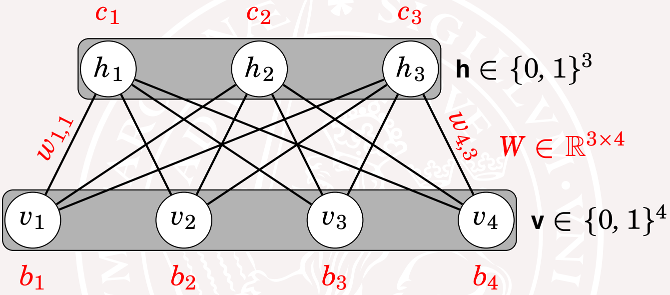
When h and v are binary vectors then

$$\frac{p(h_j = 1 \mid v)}{p(h_j = 0 \mid v)} = \exp(c_j + v^T W_{:,j})$$

hence

$$p(h_j = 1 \mid v) = \sigma(c_j + v^T W_{:,j}), \quad \text{where } \sigma(x) = \frac{\exp(x)}{1 + \exp(x)}$$

Structure of p for RBMs



$$p(h_j = 1 \mid v) = \sigma(c_j + v^T W_{:,j})$$
$$p(v_i = 1 \mid h) = \sigma(b_i + W_{i,:} h)$$

Closed form expressions, no problem with partition function Z!

Structure of $p(v)$ for RBMs

$$
\begin{aligned}
p(v) &= \sum_{h \in \{0,1\}^H} \frac{1}{Z} \exp(b^T h + c^T v + v^T W h) \\
&= \dots \\
&= \frac{1}{Z} \exp \left(c^T v + \sum_{j=1}^{H} \mathrm{softplus}(b_j + v^T W_{:,j}) \right)
\end{aligned}
$$

For details watch videos

Neural networks [5.2] by Hugo Larochelle

Neural networks [5.3] by Hugo Larochelle

Training RBMs

RBMs are popular building blocks for more advanced networks

RBMs admit efficient

- evaluation and differentation of $\tilde{p}(v)$
- MCMC sampling (blocks Gibbs sampling)

Block Gibbs:

`Initialize` h `randomly.`

`Sample` v `from` $p(v|h)$`, sample` h `from` $p(h|v)$`, iterate.`

Might need to iterate many times if the ”mixing” is slow

Training RBMs

RBMs are popular building blocks for more advanced networks

RBMs admit efficient

- evaluation and differentation of $\tilde{p}(v)$
- MCMC sampling (blocks Gibbs sampling)

Block Gibbs:

`Initialize` h `randomly.`

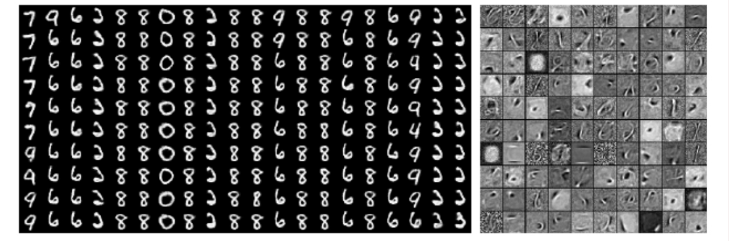
`Sample` v `from` $p(v|h)$`, sample` h `from` $p(h|v)$`, iterate.`

Might need to iterate many times if the ”mixing” is slow

To learn about Gibbs sampling watch this video

Example - MNIST

image $v \in \{0, 1\}^{768}$
features $h \in \{0, 1\}^{500}$



left: Gibbs sampling from $p(v, h)$, showing v as images
right: 100 of the feature column vectors $W_{:,j}$, mapped to grayscale values

Picture from [LISA 2008]

Training RBMs

Goal of learning: Maximize sum of log-probabilities RBM assigns to binary vectors in a training set

$$-\log p(v, h) = E(v, h) + \log Z$$

Note that

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

Must consider how both $E(v, h)$ and Z depend on parameters

Training RBMs - CD and PCD algorithms

Gradient takes surprisingly simple form as difference of two correlations:

$$\frac{\partial \log p(v)}{\partial W_{ij}} = < v_i h_j >_{\text{data } v} - < v_i h_j >_{\text{model}}$$

Watch videos

- Neural networks [5.4] by Hugo Larochelle
- Neural networks [5.5] by Hugo Larochelle
- Neural networks [5.6] by Hugo Larochelle
- Boltzmann Machine Learning by George Hinton
- RBMs by George Hinton

(+related videos by Larochelle and Hinton)

More on Training of RBMs

Recommended reading: [Hinton] A Practical Guide to training RBMs

A cool RBM example

Hinton on RBMs for Collaborative Filtering (Netflix)

Exercise

Train an RBM, for example for the MNIST dataset.

Produce figures as on slide 18 above, trying e.g. with fewer features.

Videos:

Neural networks [5.7] by Hugo Larochelle

Example of RBM Learning by Hinton

Exercise

Code to start from:

Theano implementation and tutorial about RBM

Tensorflow code or some other Tensorflow code

The MNIST data is somewhere in your tensorflow libraries under models/image/mnist/