Graphical Models

Learning with partial observations

Siamak Ravanbakhsh

Winter 2018

Learning objectives

- different types of missing data
- learning with missing data and hidden vars:
  - directed models
  - undirected models
- develop an intuition for expectation maximization
  - variational interpretation

Two settings for partial observations

- missing data
  - each instance in $\mathcal{D}$ is missing some values

Two settings for partial observations

- missing data
  - each instance in $\mathcal{D}$ is missing some values
- hidden variables
  - variables that are never observed

why model hidden variables?

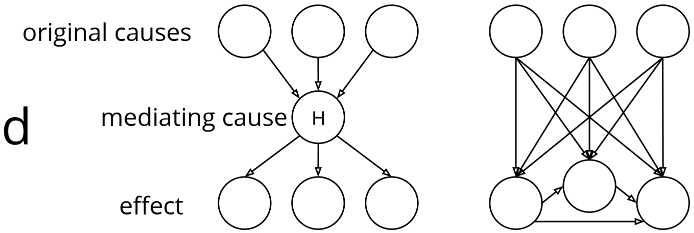


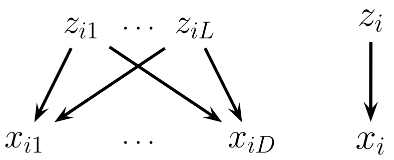
image credit: Murphy's book

Two settings for partial observations

- missing data
  - each instance in $\mathcal{D}$ is missing some values
- hidden variables
  - variables that are never observed

latent variable models

- observations have common cause
- widely used in machine learning



why model hidden variables?

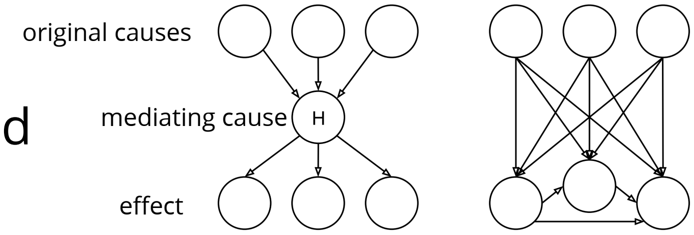


image credit: Murphy's book

Missing data

observation mechanism:

- generate the data point $X = [X_1, \ldots, X_D]$
- decide the values to observe $O_X = [\underset{\text{hide}}{\underset{\uparrow}{1}}, \underset{\text{observe}}{\underset{\uparrow}{0}}, \ldots, 0, 1]$

Missing data

observation mechanism:

- generate the data point $X = [X_1, \ldots, X_D]$
- decide the values to observe $O_X = [1, 0, \ldots, 0, 1]$

hide observe

- observe X_o while X_h is missing $(X = X_h + X_o)$

Missing data

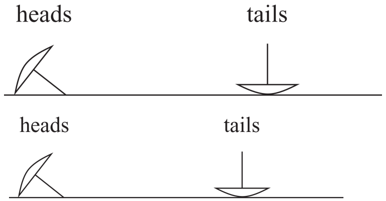
observation mechanism:

- generate the data point $X = [X_1, \ldots, X_D]$
- decide the values to observe $O_X = [1, 0, \ldots, 0, 1]$

hide observe

- observe X_o while X_h is missing $(X = X_h + X_o)$

missing completely at random (MCAR) $P(X, O_X) = P(X)P(O_X)$



$p(x) = \theta^x (1-\theta)^{1-x}$ throw to generate

Missing data

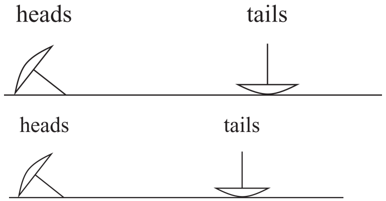
observation mechanism:

- generate the data point $X = [X_1, \ldots, X_D]$
- decide the values to observe $O_X = [1, 0, \ldots, 0, 1]$

hide observe

- observe X_o while X_h is missing $(X = X_h + X_o)$

missing completely at random (MCAR) $P(X, O_X) = P(X)P(O_X)$



$p(x) = \theta^x(1-\theta)^{1-x}$ throw to generate

$p(o) = \psi^o(1-\theta)^{1-o}$ throw to decide show/hide

Missing data

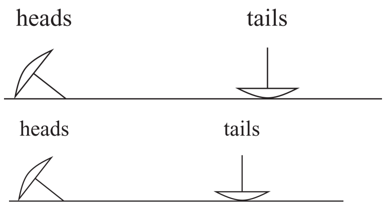
observation mechanism:

- generate the data point $X = [X_1, \dots, X_D]$
- decide the values to observe $O_X = [1, 0, \dots, 0, 1]$

hide observe

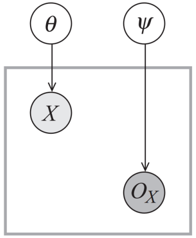
- observe X_o while X_h is missing $(X = X_h + X_o)$

missing completely at random (MCAR) $P(X, O_X) = P(X)P(O_X)$



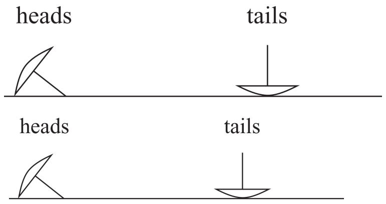
$p(x) = \theta^x (1-\theta)^{1-x}$ throw to generate

$p(o) = \psi^o (1-\theta)^{1-o}$ throw to decide show/hide



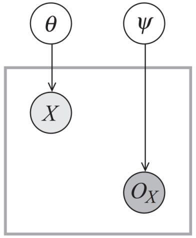
Learning with MCAR

missing completely at random (MCAR) $P(X, O) = P(X)P(O)$



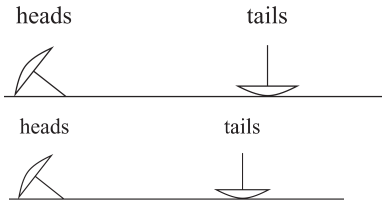
$$p(x) = \theta^x (1-\theta)^{1-x}$$ throw to generate

$$p(o) = \psi^o (1-\theta)^{1-o}$$ throw to decide show/hide



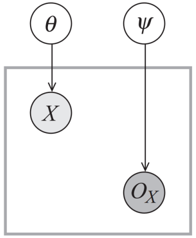
Learning with MCAR

missing completely at random (MCAR) $P(X, O) = P(X)P(O)$



$p(x) = \theta^x (1-\theta)^{1-x}$ throw to generate

$p(o) = \psi^o (1-\theta)^{1-o}$ throw to decide show/hide

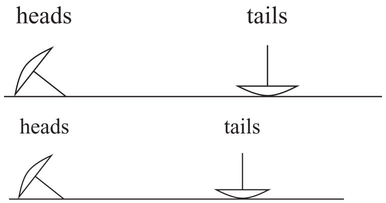


objective: learn a model for X, from the data $\mathcal{D} = \{x_o^{(1)}, \ldots, x_o^{(M)}\}$

each x_o may include values for a different subset of vars.

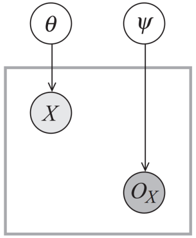
Learning with MCAR

missing completely at random (MCAR) $P(X, O) = P(X)P(O)$



$p(x) = \theta^x(1-\theta)^{1-x}$ throw to generate

$p(o) = \psi^o(1-\theta)^{1-o}$ throw to decide show/hide



objective: learn a model for X, from the data $\mathcal{D} = \{x_o^{(1)}, \ldots, x_o^{(M)}\}$

each x_o may include values for a different subset of vars.

since $P(X, O) = P(X)P(O)$, we can ignore the obs. patterns

optimize: $\ell(\mathcal{D}, \theta) = \sum_{x_o \in \mathcal{D}} \log \sum_{x_h} p(x_o, x_h)$

A more general criteria

missing at random (MAR) $O_X \perp X_h | X_o$

if there is information about the obs. pattern O_X in X_h then it is also in X_o

A more general criteria

missing at random (MAR) $O_X \perp X_h | X_o$

if there is information about the obs. pattern O_X in X_h

then it is also in X_o

example

throw the thumb-tack twice $X = [X_1, X_2]$

if $X_2 = 1$ hide X_1

otherwise show X_1

missing at random ✓

missing completely at random ✗

A more general criteria

missing at random (MAR) $O_X \perp X_h | X_o$

if there is information about the obs. pattern O_X in X_h

then it is also in X_o

example

throw the thumb-tack twice $X = [X_1, X_2]$

if $X_2 = 1$ hide X_1

otherwise show X_1

missing at random ✓

missing completely at random ✗

since there is no "extra" information in the **obs. pattern**, we can ignore it

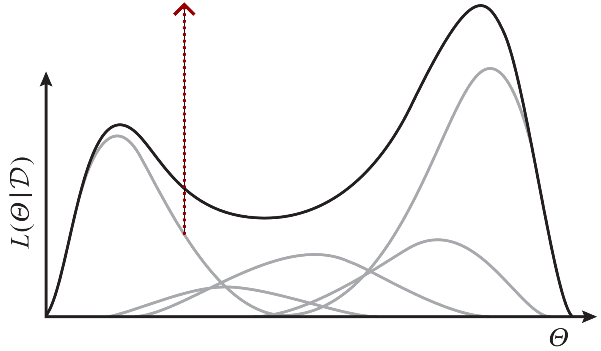
optimize: $\ell(\mathcal{D}, \theta) = \sum_{x_o \in \mathcal{D}} \log \sum_{x_h} p(x_o, x_h)$

marginal **Likelihood function**

for partial observations

- fully observed data:
  - **directed:** likelihood decomposes
  - **undirected:** does not decompose, but it is concave
- partially observed:
  - does not decompose
  - not convex anymore

likelihood for a single assignment to the latent vars.



$$\ell(\mathcal{D}, \theta) = \sum_{x_o \in \mathcal{D}} \log \sum_{x_h} p(x_o, x_h)$$

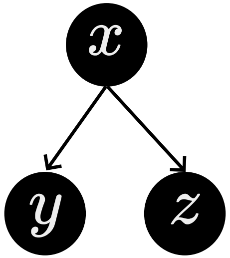
marginal **Likelihood function: example**

for a directed model

fully observed case decomposes:

$$\ell(D,\theta) = \sum_{x,y,z \in \mathcal{D}} \log p(x,y,z)$$

$$= \sum_{x} \log p(x) + \sum_{x,y} \log p(y|x) + \sum_{x,z} \log p(z|x)$$



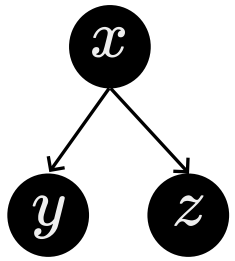
marginal **Likelihood function: example**

for a directed model

fully observed case decomposes:

$$\ell(D,\theta) = \sum_{x,y,z \in \mathcal{D}} \log p(x,y,z)$$

$$= \sum_x \log p(x) + \sum_{x,y} \log p(y|x) + \sum_{x,z} \log p(z|x)$$



x is always missing (e.g., in a **latent variable model**)

$$\ell(D,\theta) = \sum_{y,z \in \mathcal{D}} \log \sum_x p(x)p(y|x)p(z|x)$$

cannot decompose it!

Parameter learning with missing data

Directed models:

option 1: obtain the gradient of marginal likelihood

option 2: expectation maximization (EM)

- variational interpretation (in terms of free energy)
- variational EM
- Bayesian approach: variational Bayes

Parameter learning with missing data

Directed models:

option 1: obtain the gradient of marginal likelihood

option 2: expectation maximization (EM)

- variational interpretation (in terms of free energy)
- variational EM
- Bayesian approach: variational Bayes

undirected models:

obtain the gradient of marginal likelihood

- EM is not a good option here

Parameter learning with missing data

Directed models:

option 1: obtain the gradient of marginal likelihood

option 2: expectation maximization (EM)

- variational interpretation (in terms of free energy)
- variational EM
- Bayesian approach: variational Bayes

undirected models:

obtain the gradient of marginal likelihood

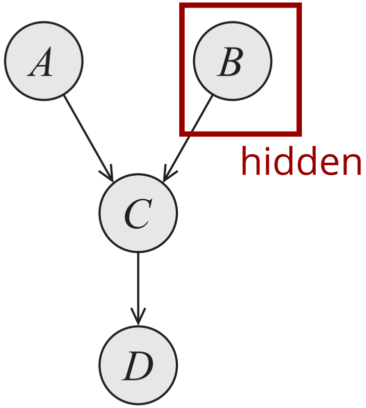
- EM is not a good option here

all of these options need inference for each step of learning

Directed models: gradient of the marginal likelihood

log marginal likelihood:

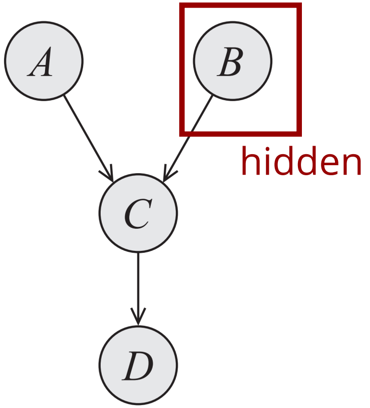
$$\ell(\mathcal{D}) = \sum_{(a,c,d)\in\mathcal{D}} \log \sum_{b} p(a)p(b)p(c|a,b)p(d|c)$$



Directed models: gradient of the marginal likelihood

log marginal likelihood:

$$\ell(\mathcal{D}) = \sum_{(a,c,d)\in\mathcal{D}} \log \sum_b p(a)p(b)p(c|a,b)p(d|c)$$



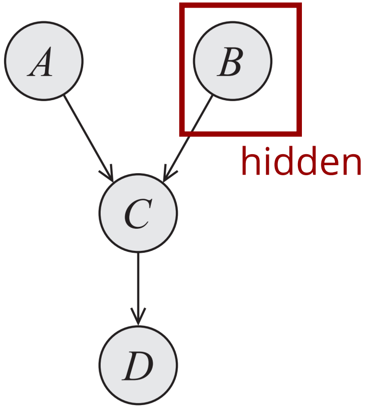
simply take the derivative:

$$\frac{\partial}{\partial p(d'|c')}\ell(\mathcal{D}) = \frac{1}{p(d'|c')} \sum_{(a,c,d)\in\mathcal{D}} p(d',c'|a,c,d)$$

need **inference** for this

what happens to this expression if every variable is observed?

Directed models: gradient of the marginal likelihood



log marginal likelihood:

$$\ell(\mathcal{D}) = \sum_{(a,c,d)\in\mathcal{D}} \log \sum_b p(a)p(b)p(c|a,b)p(d|c)$$

simply take the derivative:

$$\frac{\partial}{\partial p(d'|c')}\ell(\mathcal{D}) = \frac{1}{p(d'|c')}\sum_{(a,c,d)\in\mathcal{D}} p(d',c'|a,c,d)$$

need **inference** for this

what happens to this expression if every variable is observed?

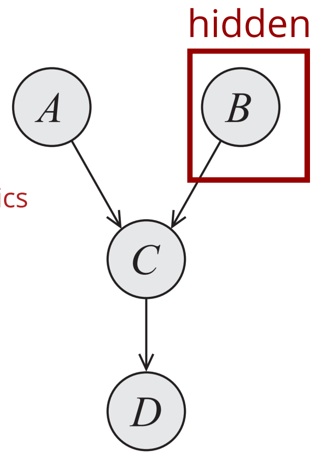
if the cond. prob. is parameterized, use the chain rule:

$$\frac{\partial}{\partial\theta}\ell(\mathcal{D};\theta) = \sum_{(c',d')\in\mathcal{D}} \frac{\partial\ell(\mathcal{D})}{\partial p(d'|c')}\frac{\partial p(d'|c')}{\partial\theta}$$

Directed models: expectation maximization

E-step:
for each $a, c, d \in \mathcal{D}$
use the current parameters θ to get the marginals

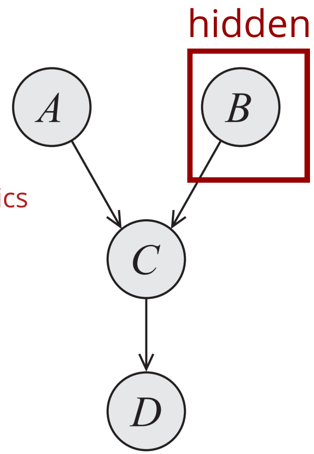
more generally: expected sufficient statistics



Directed models: expectation maximization

E-step:
for each $a, c, d \in \mathcal{D}$
use the current parameters θ to get the marginals

more generally: expected sufficient statistics



$$p(B|\mathcal{D};\theta_B), p(A|\mathcal{D};\theta_A), p(A,B,C|\mathcal{D};\theta_{C|A,B}), p(D,C|\mathcal{D};\theta_{D|C})$$

$$p(B=b'|\mathcal{D};\theta_B) = \frac{1}{N}\sum_{(a,c,d)\in\mathcal{D}} p(b'|a,c,d;\theta_B)$$

need inference here

Directed models: expectation maximization

E-step:

for each $a, c, d \in \mathcal{D}$

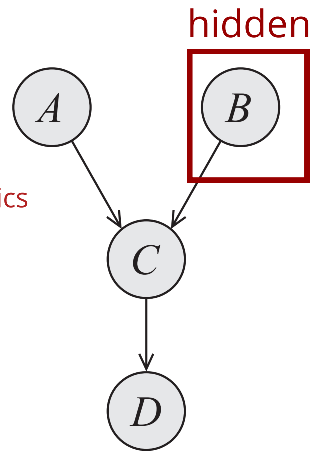
use the current parameters θ to get the marginals

more generally: expected sufficient statistics

$$p(B|\mathcal{D};\theta_B), p(A|\mathcal{D};\theta_A), p(A,B,C|\mathcal{D};\theta_{C|A,B}), p(D,C|\mathcal{D};\theta_{D|C})$$

$$p(B=b'|\mathcal{D};\theta_B) = \frac{1}{N}\sum_{(a,c,d)\in\mathcal{D}} p(b'|a,c,d;\theta_B)$$

need inference here



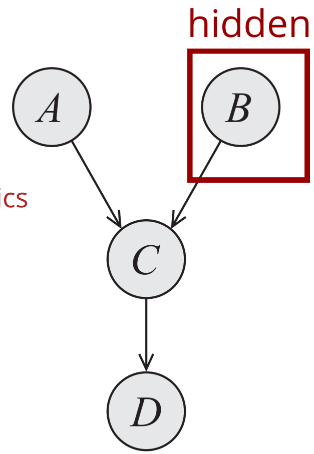
M-step:

use the marginals (similar to completely observed data) to learn θ

more generally: expected sufficient statistics

Directed models: expectation maximization

E-step:
for each $a, c, d \in \mathcal{D}$
use the current parameters θ to get the marginals
more generally: expected sufficient statistics



$$p(B|\mathcal{D};\theta_B), p(A|\mathcal{D};\theta_A), p(A,B,C|\mathcal{D};\theta_{C|A,B}), p(D,C|\mathcal{D};\theta_{D|C})$$

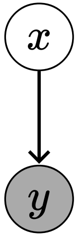
$$p(B=b'|\mathcal{D};\theta_B) = \frac{1}{N}\sum_{(a,c,d)\in\mathcal{D}} p(b'|a,c,d;\theta_B)$$

need inference here

M-step:
use the marginals (similar to completely observed data) to learn θ
more generally: expected sufficient statistics

E.g., update $\theta_{C|A,B}$ using $p(A,B,C|\mathcal{D};\theta_{C|A,B})$ $\rightarrow$ $\theta^{new}_{C|A,B} = \frac{p(A,B,C|\mathcal{D};\theta_{C|A,B})}{p(A,B|\mathcal{D};\theta_{C|A,B})}$

Example: Gaussian mixture model

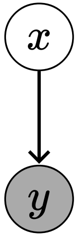


$$p(x; \pi) = \prod_k \pi_k^{\mathbb{I}(x=k)}$$

model parameters

$$p(y|x; \{\mu_k, \Sigma_k\}) = \frac{1}{\sqrt{|2\pi\Sigma_x|}} \exp(-\tfrac{1}{2}(y - \mu_x)^T \Sigma_x^{-1} (y - \mu_x))$$

Example: Gaussian mixture model



$$p(x; \pi) = \prod_k \pi_k^{\mathbb{I}(x=k)}$$

model parameters

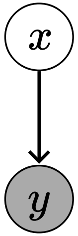
$$p(y|x; \{\mu_k, \Sigma_k\}) = \frac{1}{\sqrt{|2\pi\Sigma_x|}} \exp(-\tfrac{1}{2}(y-\mu_x)^T \Sigma_x^{-1}(y-\mu_x))$$

E-step: calculate $p(x|y)$ for each $y \in \mathcal{D}$

$$p(x|y) \propto p(x;\pi)p(y|x;\mu,\Sigma) \;=\pi_k \mathcal{N}(y;\mu_k,\Sigma_k)$$

- now we have *"probabilistically completed"* instances
- update the parameters (easy in a Bayes-net)

Example: Gaussian mixture model



$$p(x; \pi) = \prod_k \pi_k^{\mathbb{I}(x=k)}$$

model parameters

$$p(y|x; \{\mu_k, \Sigma_k\}) = \frac{1}{\sqrt{|2\pi\Sigma_x|}} \exp(-\tfrac{1}{2}(y - \mu_x)^T \Sigma_x^{-1}(y - \mu_x))$$

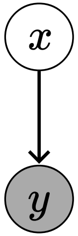
M-step: estimate $\pi, \mu_k, \Sigma_k \forall k$

$$\pi_k = \frac{1}{N} \sum_{y \in \mathcal{D}} \frac{p(x=k|y)}{\sum_{k'} p(x=k'|y)}$$ portion of all particles assigned to this cluster (sum of probs.)

$$\mu_k = \frac{\sum_{y \in \mathcal{D}} p(x=k|y) y}{\sum_{y \in \mathcal{D}} p(x=k|y)}$$ mean of a weighted set of instances

$$\Sigma_k = \frac{\sum_{y \in \mathcal{D}} p(x=k|y)(y-\mu_k)(y-\mu_k)^T}{\sum_{y \in \mathcal{D}} p(x=k|y)}$$ covariance of a weighted set of instances

Example: Gaussian mixture model



$$p(x; \pi) = \prod_k \pi_k^{\mathbb{I}(x=k)}$$

model parameters

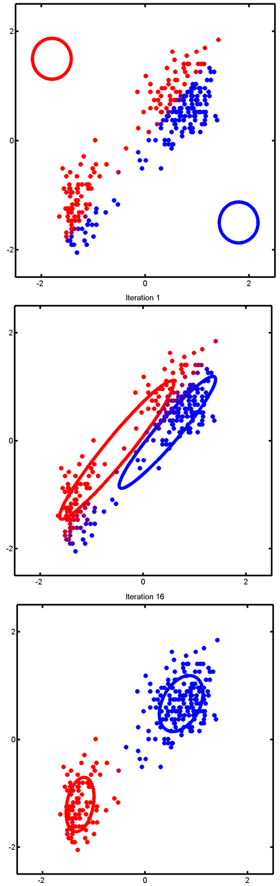
$$p(y|x; \{\mu_k, \Sigma_k\}) = \frac{1}{\sqrt{|2\pi\Sigma_x|}} \exp(-\tfrac{1}{2}(y - \mu_x)^T \Sigma_x^{-1} (y - \mu_x))$$

M-step: estimate $\pi, \mu_k, \Sigma_k \forall k$

$$\pi_k = \frac{1}{N} \sum_{y \in \mathcal{D}} \frac{p(x=k|y)}{\sum_{k'} p(x=k'|y)}$$ portion of all particles assigned to this cluster (sum of probs.)

$$\mu_k = \frac{\sum_{y \in \mathcal{D}} p(x=k|y) y}{\sum_{y \in \mathcal{D}} p(x=k|y)}$$ mean of a weighted set of instances

$$\Sigma_k = \frac{\sum_{y \in \mathcal{D}} p(x=k|y)(y-\mu_k)(y-\mu_k)^T}{\sum_{y \in \mathcal{D}} p(x=k|y)}$$ covariance of a weighted set of instances

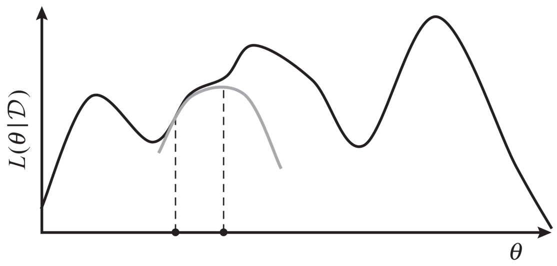


Directed models: expectation maximization

E-step:
for each $a, c, d \in \mathcal{D}$
use the current parameters θ to get the marginals

M-step:
use the marginals (similar to completely observed data) to learn θ

- guaranteed to improve the likelihood at each step
    - first initial steps quickly improve the likelihood, then slows down

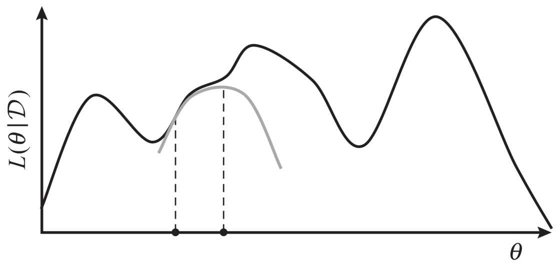


Directed models: expectation maximization

E-step:
for each $a, c, d \in \mathcal{D}$
use the current parameters θ to get the marginals

M-step:
use the marginals (similar to completely observed data) to learn θ

- guaranteed to improve the likelihood at each step
  - first initial steps quickly improve the likelihood, then slows down
- converges to a local optimum:
  - multiple restarts are useful

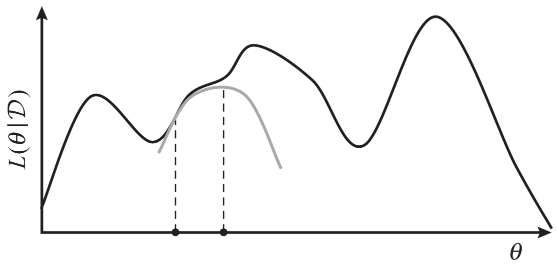


Directed models: expectation maximization

E-step:
for each $a, c, d \in \mathcal{D}$
use the current parameters θ to get the marginals

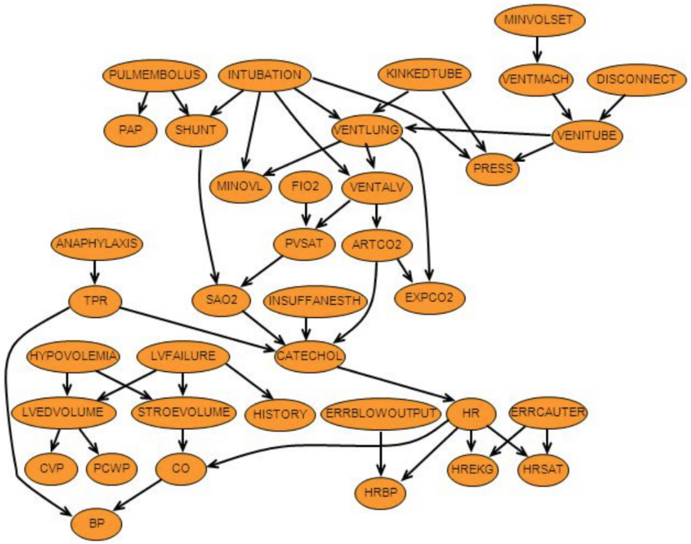
M-step:
use the marginals (similar to completely observed data) to learn θ

- guaranteed to improve the likelihood at each step
  - first initial steps quickly improve the likelihood, then slows down
- converges to a local optimum:
  - multiple restarts are useful
- for undirected models: M-step is the expensive part
  - perform E-step within each iteration of M-step: equivalent to gradient descent



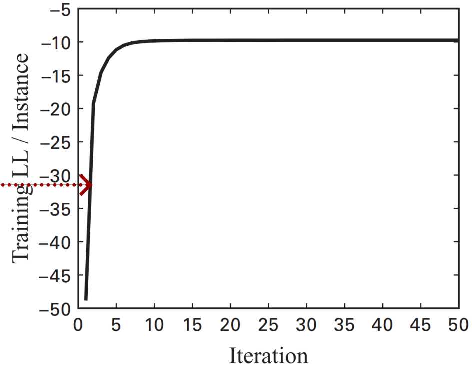
expectation maximization: example

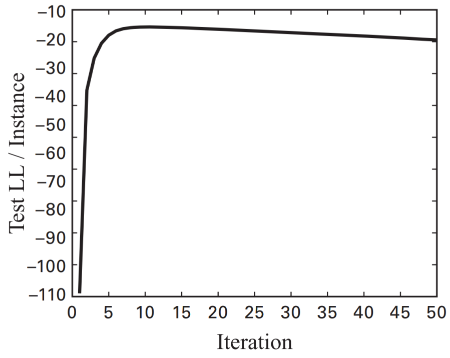
- 1000 training instances
- 50% of variables are observed (in each instance)



alarm network

fast initial improvement



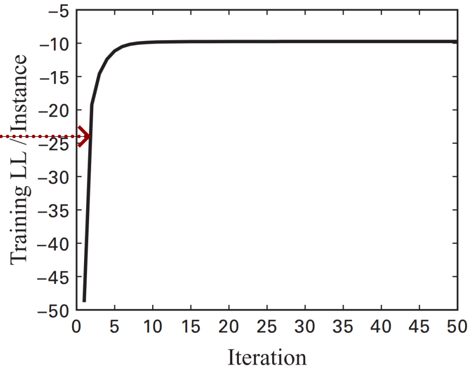


expectation maximization: example

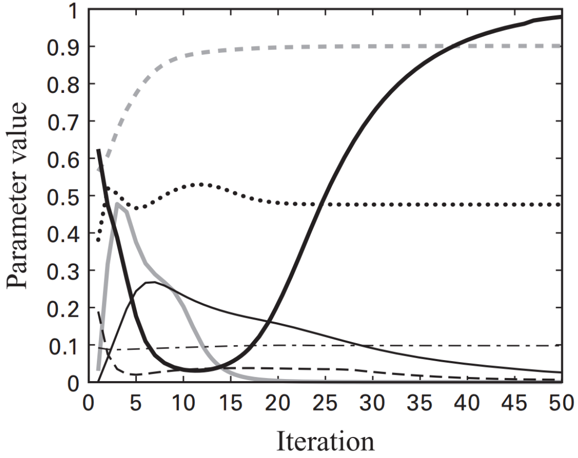
- 1000 training instances
- 50% of variables are observed (in each instance)

train log-likelihood

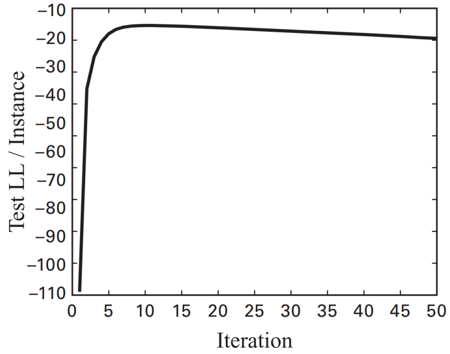
fast initial improvement



change in different parameter values

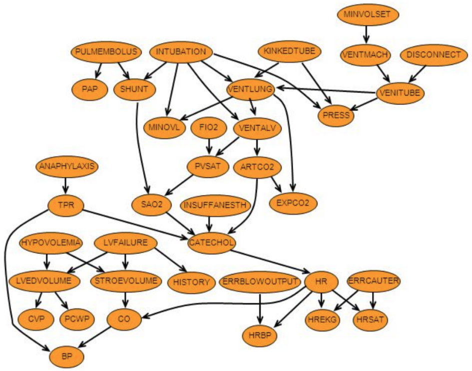


test log-likelihood



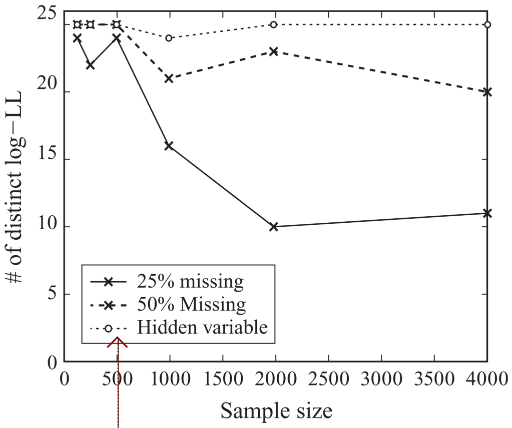
expectation maximization: example

local optima in EM:



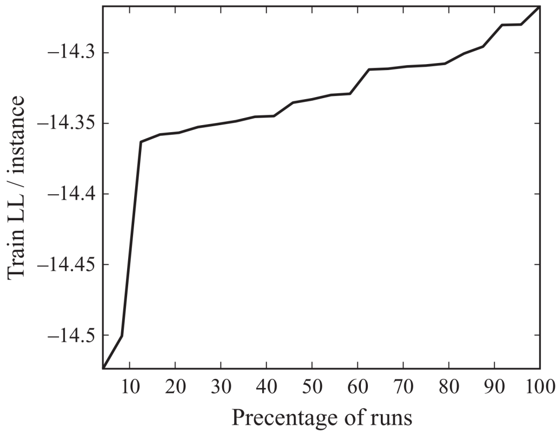
alarm network

number of local maxima



a single hidden variable

effect of multiple restarts



Variational interpretation of EM

posterior $p(h|\mathcal{D};\theta) = \frac{p(h,\mathcal{D};\theta)}{p(\mathcal{D};\theta)}$ a role similar to the partition function $Z(\theta)$

Variational interpretation of EM

posterior $p(h|\mathcal{D};\theta) = \frac{p(h,\mathcal{D};\theta)}{p(\mathcal{D};\theta)}$ a role similar to the partition function $Z(\theta)$

$$D_{KL}(q(h);p(h|\mathcal{D},\theta)) = \underbrace{-H(q) - \mathbb{E}_q[\log p(h,\mathcal{D};\theta)]}_{\text{negative of variational free energy}} + \underbrace{\log p(\mathcal{D};\theta)}_{\text{we want to maximize this!}}$$

Variational interpretation of EM

posterior $p(h|\mathcal{D};\theta) = \frac{p(h,\mathcal{D};\theta)}{p(\mathcal{D};\theta)}$ a role similar to the partition function $Z(\theta)$

$$D_{KL}(q(h);p(h|\mathcal{D},\theta)) = -H(q) - \mathbb{E}_q[\log p(h,\mathcal{D};\theta)] + \log p(\mathcal{D};\theta)$$

negative of variational free energy we want to maximize this!

→ $$\ell(\mathcal{D};\theta) = \underline{H(q) + \mathbb{E}_q[\log p(h,\mathcal{D};\theta)]} + D_{KL}(q(h);p(h|\mathcal{D},\theta))$$

evidence lower bound (ELBO) is a lower-bound on the likelihood

Variational interpretation of EM

posterior $p(h|\mathcal{D};\theta) = \frac{p(h,\mathcal{D};\theta)}{p(\mathcal{D};\theta)}$ a role similar to the partition function $Z(\theta)$

$$D_{KL}(q(h);p(h|\mathcal{D},\theta)) = \underbrace{-H(q) - \mathbb{E}_q[\log p(h,\mathcal{D};\theta)]}_{\text{negative of variational free energy}} + \underbrace{\log p(\mathcal{D};\theta)}_{\text{we want to maximize this!}}$$

$$\rightarrow \quad \ell(\mathcal{D};\theta) = \underbrace{H(q) + \mathbb{E}_q[\log p(h,\mathcal{D};\theta)]}_{\text{evidence lower bound (ELBO) is a lower-bound on the likelihood}} + D_{KL}(q(h);p(h|\mathcal{D},\theta))$$

EM: perform block coordinate ascent

- optimize q to match the posterior (*i.e.*, obtain the posterior)
- optimize θ to increase ELBO

Variational interpretation of EM

$$\ell(\mathcal{D};\theta) = \underbrace{H(q) + \mathbb{E}_q[\log p(h,\mathcal{D};\theta)]}_{} + D_{KL}(q(h);p(h|\mathcal{D},\theta))$$

evidence lower bound (ELBO) is a lower-bound on the likelihood

this interpretation also leads to:

variational EM:

- use a family q and approximate variational inference to obtain q

variational Bayes:

- add a prior $p(\theta)$ and get a posterior over both latent vars (h) and parameters θ

Undirected models with latent variables

recall

linear exponential family $$p(x;\theta) = \frac{1}{Z(\theta)} \exp(\langle \theta, \phi(x) \rangle)$$

gradient in the fully observed setting $$\nabla_\theta \ell(\theta, \mathcal{D}) = |\mathcal{D}|(\mathbb{E}_{\mathcal{D}}[\phi(x)] - \mathbb{E}_{p_\theta}[\phi(x)])$$

expectation wrt the data

expectation wrt the model

Undirected models with latent variables

recall

linear exponential family $p(x;\theta) = \frac{1}{Z(\theta)} \exp(\langle \theta, \phi(x) \rangle)$

gradient in the fully observed setting $\nabla_\theta \ell(\theta, \mathcal{D}) = |\mathcal{D}| (\underbrace{\mathbb{E}_{\mathcal{D}}[\phi(x)]}_{\text{expectation wrt the data}} - \underbrace{\mathbb{E}_{p_\theta}[\phi(x)]}_{\text{expectation wrt the model}})$

partial observation: $x = (x_o, \underbrace{x_h}_{\text{not observed}})$

Undirected models with latent variables

recall

linear exponential family $p(x;\theta) = \frac{1}{Z(\theta)} \exp(\langle \theta, \phi(x) \rangle)$

gradient in the fully observed setting $\nabla_\theta \ell(\theta, \mathcal{D}) = |\mathcal{D}| (\underbrace{\mathbb{E}_{\mathcal{D}}[\phi(x)]}_{\text{expectation wrt the data}} - \underbrace{\mathbb{E}_{p_\theta}[\phi(x)]}_{\text{expectation wrt the model}})$

partial observation: $x = (x_o, \underbrace{x_h}_{\text{not observed}})$

marginal likelihood: $p(x_o;\theta) = \sum_{x_h} \frac{1}{Z(\theta)} \exp(\langle \theta, \phi(x) \rangle)$

Undirected models with latent variables

recall

linear exponential family $p(x;\theta) = \frac{1}{Z(\theta)} \exp(\langle \theta, \phi(x) \rangle)$

gradient in the fully observed setting $\nabla_\theta \ell(\theta, \mathcal{D}) = |\mathcal{D}|(\underbrace{\mathbb{E}_{\mathcal{D}}[\phi(x)]}_{\text{expectation wrt the data}} - \underbrace{\mathbb{E}_{p_\theta}[\phi(x)]}_{\text{expectation wrt the model}})$

partial observation: $x = (x_o, \underbrace{x_h}_{\text{not observed}})$

marginal likelihood: $p(x_o;\theta) = \sum_{x_h} \frac{1}{Z(\theta)} \exp(\langle \theta, \phi(x) \rangle)$

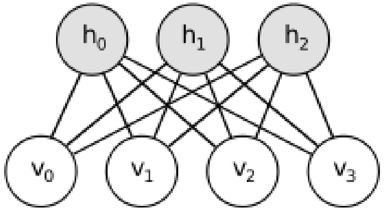
gradient in the partially obs. case $\nabla_\theta \ell(\theta, \mathcal{D}) = |\mathcal{D}|(\mathbb{E}_{\mathcal{D},\theta}[\phi(x)] - \mathbb{E}_{p_\theta}[\phi(x)])$

wrt both data and model: we need to do inference to calculate expected sufficient statistics (similar to EM)

Example: Restricted Boltzmann Machine

recall the binary RBM: $p(h,v) = \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$

for $v_i, h_j \in \{0,1\}$

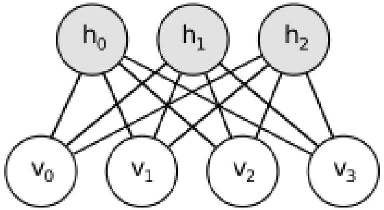


data: $\mathcal{D} = \{v^{(m)}\}_m$

Example: Restricted Boltzmann Machine

recall the binary RBM: $p(h, v) = \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$

for $v_i, h_j \in \{0, 1\}$

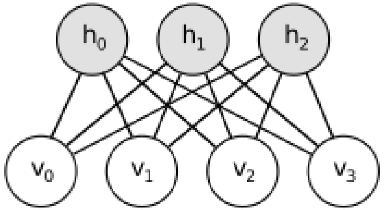


data: $\mathcal{D} = \{v^{(m)}\}_m$

sufficient statistics: $\phi(v_i, h_j) = v_i, h_j$

Example: Restricted Boltzmann Machine

recall the binary RBM: $p(h, v) = \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$



data: $\mathcal{D} = \{v^{(m)}\}_m$

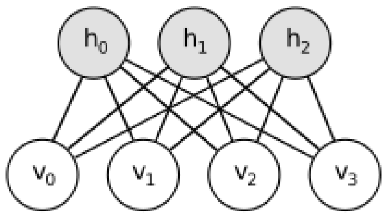
for $v_i, h_j \in \{0, 1\}$

sufficient statistics: $\phi(v_i, h_j) = v_i, h_j$

we want to optimize: $\ell(\mathcal{D}; \theta) = \sum_{v \in \mathcal{D}} \log \sum_h \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$

Example: Restricted Boltzmann Machine

recall the binary RBM: $p(h, v) = \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$



data: $\mathcal{D} = \{v^{(m)}\}_m$

for $v_i, h_j \in \{0, 1\}$

sufficient statistics: $\phi(v_i, h_j) = v_i, h_j$

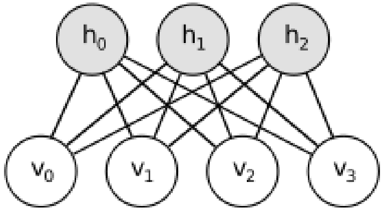
we want to optimize: $\ell(\mathcal{D}; \theta) = \sum_{v \in \mathcal{D}} \log \sum_h \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$

gradient: $\frac{\partial}{\partial \theta_{i,j}} \ell(\mathcal{D}; \theta) \propto \mathbb{E}_{\mathcal{D},\theta}[v_i h_j] - \mathbb{E}_{p_\theta}[v_i h_j]$

$$= (\frac{1}{M} \sum_{v' \in \mathcal{D}} \mathbb{E}_{p_\theta}[h_j | v'_i]) - \mathbb{E}_{p_\theta}[v_i h_j])$$

Example: Restricted Boltzmann Machine

recall the binary RBM: $p(h, v) = \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$



for $v_i, h_j \in \{0, 1\}$

data: $\mathcal{D} = \{v^{(m)}\}_m$

sufficient statistics: $\phi(v_i, h_j) = v_i, h_j$

we want to optimize: $\ell(\mathcal{D}; \theta) = \sum_{v \in \mathcal{D}} \log \sum_h \frac{1}{Z(\theta)} \exp(\sum_{i,j} \theta_{i,j} v_i h_j)$

gradient: $\frac{\partial}{\partial \theta_{i,j}} \ell(\mathcal{D}; \theta) \propto \mathbb{E}_{\mathcal{D},\theta}[v_i h_j] - \mathbb{E}_{p_\theta}[v_i h_j]$

$$= \left(\frac{1}{M} \sum_{v' \in \mathcal{D}} \mathbb{E}_{p_\theta}[h_j | v'_i]\right) - \mathbb{E}_{p_\theta}[v_i h_j])$$

sampling-based inference: sample h | v

use Gibbs sampling:
sample both h,v using current parameters

summary

learning with partial observations:

- missing data
  - optimize the likelihood when missing at random
- latent variables
  - can produce expressive probabilistic models

problem is not convex

how to learn the model?

- directly estimate the gradient
- use EM

both cases require inference within each step