Chapter 11
Learning with Hidden Variables

11.1 Hidden Variables and Missing Data

- Observational variables
  - Visible - those for which we actually know the state
  - Missing – those whose states are missing for a particular datapoint
  - Hidden(latent) - not all variables in the model are observed

11.2 Expectation Maximisation

11.2.1 Variational EM
    individual parameter updates

Consider a single variable pair (v, h)
    v - visible variable
    h - hidden variable
Model $p(v, h|\theta)$
Maximising the marginal likelihood $p(\mathcal{V}|\theta)$ for observed data $\mathcal{V}$

Variational distribution $q(h|v)$
Parametric model $p(h|v,\theta)$:
Kullback-Leibler divergence

$$\begin{aligned}\mathrm{KL}(q(h|v)|p(h|v,\theta)) &\equiv \langle \log q(h|v) - \log p(h|v,\theta)\rangle_{q(h|v)} \\ &= \langle \log q(h|v) - \log p(h,v|\,\theta) + \log p(v|\theta)\rangle_{q(h|v)} \\ &= \langle \log q(h|v)\rangle_{q(h|v)} - \langle \log p(h,v|\,\theta)\rangle_{q(h|v)} + \log p(v|\theta) \\ &\geq 0\end{aligned}$$

Log likelihood

$$\log p(v|\theta) \geq -\langle \log q(h|v)\rangle_{q(h|v)} + \langle \log p(h,v|\,\theta)\rangle_{q(h|v)}$$

$$\log p(\mathcal{V}|\theta) \geq -\sum_{n=1}^{N} \langle \log q(h^n|v^n)\,\rangle_{q(h^n|v^n)} + \sum_{n=1}^{N} \langle \log p(h^n,v^n|\theta)\rangle_{q(h^n|v^n)} = \tilde{L}(q,\theta)$$

EM(expectation maximization)

Find q and θ that maximises $\tilde{L}(q,\theta)$

Repeat

E-step - for fixed θ, find the distributions q that maximise $\tilde{L}(q,\theta)$

M-step – for fixed q, find the distributions θ that maximise $\tilde{L}(q,\theta)$

The maximised $\tilde{L}(q,\theta)$ is equal to $\log p(\mathcal{V}|\theta)$

Algorithm 11.1 Expectation Maximisation. Compute Maximum Likelihood value for data with hidden variables. Input: a distribution $p(x|\theta)$ and dataset $\mathcal{V}$. Returns ML candidate θ.

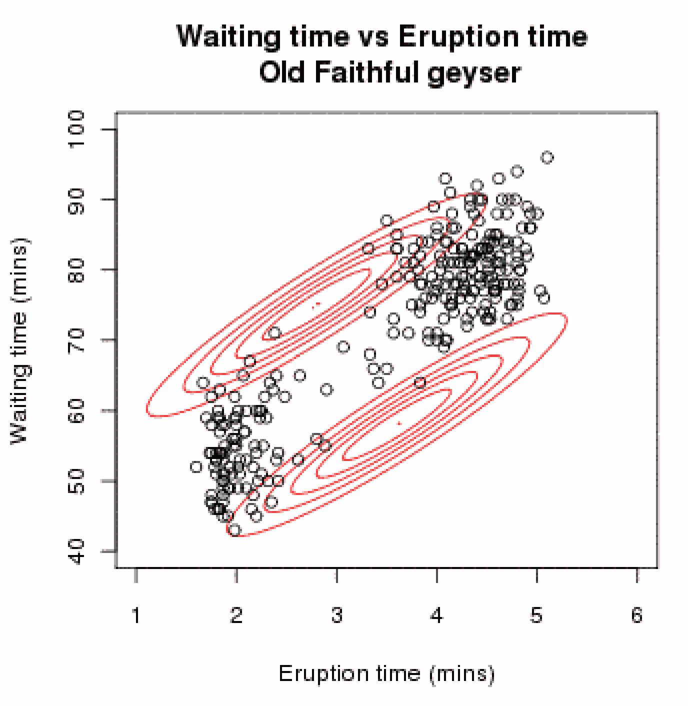
$t = 0$ ▷ Iteration counter
Choose an initial setting for the parameters θ^0. ▷ Initialisation
while θ not converged (or likelihood not converged) **do**
    $t \leftarrow t + 1$
    for $n = 1$ to N **do** ▷ Run over all datapoints
        $q_t^n(h^n|v^n) = p(h^n|v^n, \theta^{t-1})$ ▷ E step
    end for
    $\theta^t = \arg\max_\theta \sum_{n=1}^{N} \langle \log p(h^n, v^n|\theta) \rangle_{q_t^n(h^n|v^n)}$ ▷ M step
end while
return θ^t ▷ The max likelihood parameter estimate.

https://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

Find two gaussian distribution

$\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$

from observations



11.2.2 Classical EM

$$\tilde{L}(q,\theta) = -\sum_{n=1}^{N} \langle \log q(h^n|v^n) \rangle_{q(h^n|v^n)} + \sum_{n=1}^{N} \langle \log p(h^n, v^n|\theta) \rangle_{q(h^n|v^n)}$$

E-step - for fixed θ, find the distributions q that maximise $\tilde{L}(q,\theta)$

optimal solution

$$q(h^n|v^n) = p(h^n|v^n,\theta)$$

M-step – for fixed q, find the distributions θ that maximise $\tilde{L}(q,\theta)$

i.e. maximise

$$\sum_{n=1}^{N} \langle \log p(h^n, v^n|\theta) \rangle_{q(h^n|v^n)}$$

Example 11.2 (EM for a one-parameter model)

visible variable $v \in \mathbb{R}$

hidden variable $h \in \{1,2\}$

model

$$p(v,h|\theta) = p(v|h,\theta)p(h)$$

$$p(v|h,\theta) = \frac{1}{\sqrt{\pi}} e^{-(v-\theta h)^2}$$

$$p(h=1) = p(h=2) = 0.5$$

observation

$$v = 2.75$$

goal – to find θ that maximises $p(v = 2.75|\theta)$

Computational(non-EM) approach

$$
\begin{aligned}
p(v = 2.75|\theta) &= \sum_{h=1,2} p(v = 2.75, h|\theta) \\
&= \sum_{h=1,2} p(v = 2.75|h, \theta)p(h) \\
&= \frac{1}{2\sqrt{\pi}}\left(e^{-(2.75-\theta)^2} + e^{-(2.75-2\theta)^2}\right)
\end{aligned}
$$

answer - $\theta \ = \ 1.325$

EM approach

$$\tilde{L}(q,\theta) = -\langle \log q(h|v) \rangle_{q(h|v)} + \langle \log p(h,v|\theta)\rangle_{q(h|v)}$$

$$\langle \log p(h,v|\theta)\rangle_{q(h|v)} = \langle \log p(v|h,\theta)\rangle_{q(h|v)} + \langle \log p(h)\rangle_{q(h|v)}$$

$$= -\langle (v-\theta h)^2\rangle_{q(h|v)} + \text{const.}$$

$$\tilde{L}(q,\theta) = -\sum_{h=1,2} q(h)\log q(h) - \sum_{h=1,2} q(h)(2.75-\theta h)^2 + \text{const.}$$

E-step

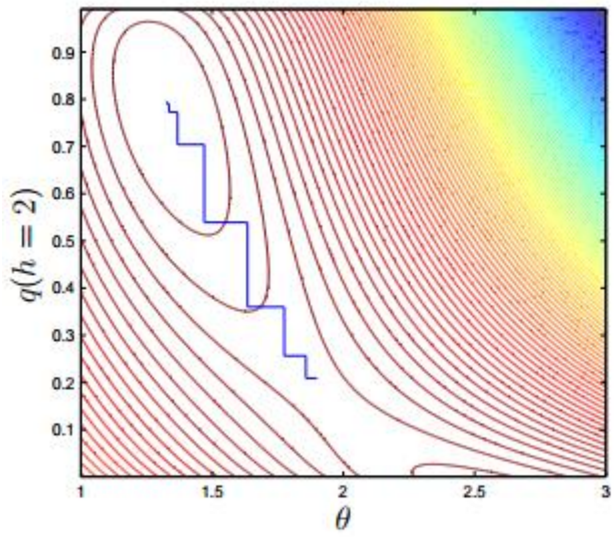
$$q^{new}(h) = p(h|v,\theta)$$

$$q^{new}(h=2) = \frac{p(v=2.75|h=2,\theta)p(h=2)}{p(v=2.75|\theta)} = \frac{e^{-(2.75-2\theta)^2}}{e^{-(2.75-2\theta)^2} + e^{-(2.75-\theta)^2}}$$

$$q^{new}(h=1) = 1 - q^{new}(h=2)$$

M-step

$$\theta^{new} = \frac{2.75\langle h\rangle_{q(h)}}{\langle h^2\rangle_{q(h)}} \Leftarrow \boxed{\frac{d}{d\theta}\sum_{h=1,2} q(h)(2.75-\theta h)^2 = 0}$$


0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
$q(h=2)$
1
1.5
2
2.5
3
θ

Example 11.3

Model

$$p(x1, x2|\theta),\ x_1, x_2 \in \{1,2\}$$

Let

$$p(x_1, x_2|\theta) = \theta_{x_1,x_2},\quad \theta_{1,1} + \theta_{1,2} + \theta_{2,1} + \theta_{2,2} = 1$$

Data

$$\mathbf{x}^1 = (1,1),\ \mathbf{x}^2 = (1,?),\ \mathbf{x}^3 = (?,2)$$

Aim

learn $\theta = (\theta_{1,1}, \theta_{1,2}, \theta_{2,1}, \theta_{2,2})$

EM

$$\tilde{L}(q,\theta) = -\sum_{n=1}^{N} \underbrace{\langle \log q(h^n|v^n) \rangle_{q(h^n|v^n)}}_{\text{entropy}} + \sum_{n=1}^{N} \underbrace{\langle \log p(h^n, v^n|\theta) \rangle_{q(h^n|v^n)}}_{\text{energy}}$$

E-step

$$q^{new}(h) = p(h|x,\theta)$$

M-step

$$\tilde{L}(q,\theta) = \log p(x_1 = 1, x_2 = 1|\theta) + \langle \log p(x_1 = 1, x_2|\theta)\rangle_{p\left(x_2\middle|x_1 = 1, \theta^{old}\right)}$$
$$+ \langle \log p(x_1, x_2 = 2|\theta)\rangle_{p\left(x_1\middle|x_2 = 2, \theta^{old}\right)} + \text{const.}$$

$$= \log \theta_{1,1}$$
$$+ p\left(x_2 = 1\middle|x_1 = 1, \theta^{old}\right)\log\theta_{1,1} + p\left(x_2 = 2\middle|x_1 = 1, \theta^{old}\right)\log\theta_{1,2}$$
$$+ p\left(x_1 = 1\middle|x_2 = 2, \theta^{old}\right)\log\theta_{1,2} + p\left(x_1 = 2\middle|x_2 = 2, \theta^{old}\right)\log\theta_{2,2} + \text{const.}$$

Lagrange multiplier

$$g(\theta) = \theta_{1,1} + \theta_{1,2} + \theta_{2,1} + \theta_{2,2} - 1$$
$$\nabla\tilde{L}(q,\theta) = \lambda\nabla g(\theta)$$

solution

$\theta_{1,1} \propto 1 + p\left(x_2 = 1\middle|x_1 = 1, \theta^{old}\right)$, $\theta_{1,2} \propto p\left(x_2 = 2\middle|x_1 = 1, \theta^{old}\right) + p(x_2 = 1|x_1 = 2, \theta^{old})$

$\theta_{2,1} = 0$, $\theta_{2,2} \propto p\left(x_1 = 2\middle|x_2 = 2, \theta^{old}\right)$