CMPUT463/563
Probabilistic Graphical Models

Partially Observable Learning

Lili Mou

Dept. Computing Science, University of Alberta

lmou@ualberta.ca

Outline

- Partially observable learning is feasible
  - Mixture of Gaussian, HMM
- Expectation—Maximization: Alternate between

  E-step: Compute $q(z|x) = P(z|x)$

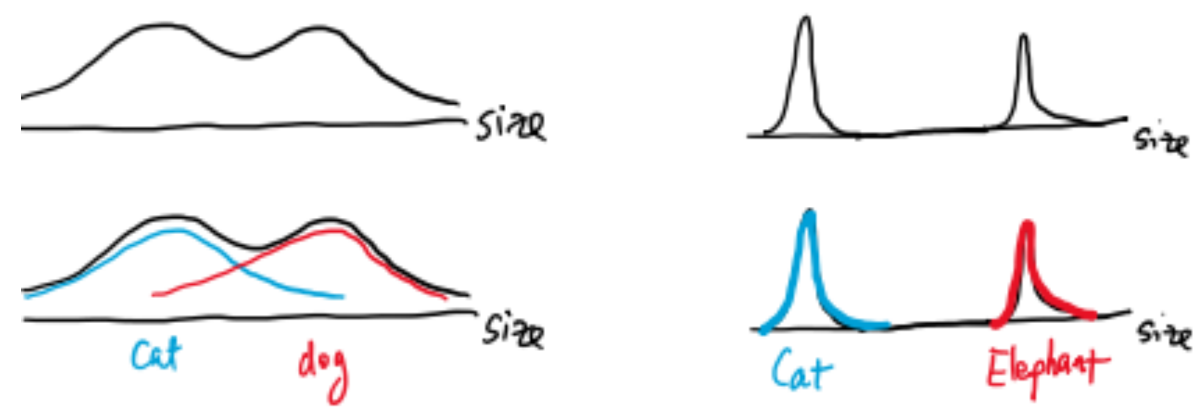
  M-step: Maximize the expected complete likelihood

  $$\theta \leftarrow \text{argmax}\, \mathbb{E}_{z \sim q(z|x)}[P(x, z; \theta)]$$

- EM also works for MN (usually in the log-linear form)
  - Still expectation in data VS expectation in model
  - But the first term takes the expectation over latent variable
- Theoretical justification: EM is MLE, EM is MM.

Partial Observable Learning

- Partially observable learning: some variables are missing in some data samples
- Example: Mixture of Gaussian
  - (Species) $\rightarrow$ (size)
  - Without the supervision of labels, we know they belong to different classes



- Example: Discrete case
  - Clustering of documents based on some latent topic
- Example: HMM
  - Unsupervised learning: $Y_1, \cdots, Y_T$ is not observed in any sample. But don't trust it too much.
  - Weakly supervised learning: Small set of labeled data; massive unlabeled data

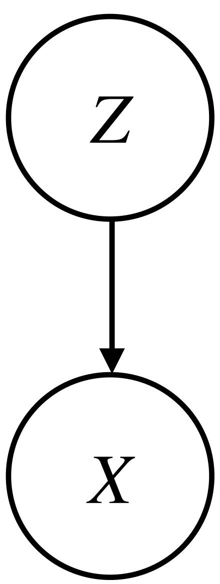
A Heuristic Idea

- If we knew the missing variable
  - Learning would be easy: supervised learning
- Although we do not know the missing variable
  - We can estimate its value and perform supervised-like learning

$$z = k \sim \mathrm{cat}(\pi_1, \pi_2, \cdots, \pi_K)$$
$$\boldsymbol{x} \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

Training data

$$\{\boldsymbol{x}^{(m)}\}_{m=1}^{M}$$



Expectation (E-step)

$$q_k^{(m)} = P(z^{(m)} \mid \boldsymbol{x}^{(m)}; \boldsymbol{\theta})$$

Maximization (M-step)

$$\boldsymbol{\theta} = \mathrm{argmax} \sum_{m=1}^{M} \boxed{\sum_{k=1}^{K} q_k^{(m)}} \boxed{\log P(z^{(m)} = k, \boldsymbol{x}^{(m)}; \boldsymbol{\theta})}$$

Supervised-like training

Soft samples weighted by $q_k^{(m)}$

A variant

- k-means

E-step

$$q_k^{(m)} = \boxed{\text{onehot}[\text{argmax}} \, P(z^{(m)} \mid \boldsymbol{x}^{(m)}; \boldsymbol{\theta})]$$

Hard estimation of the cluster

M-step

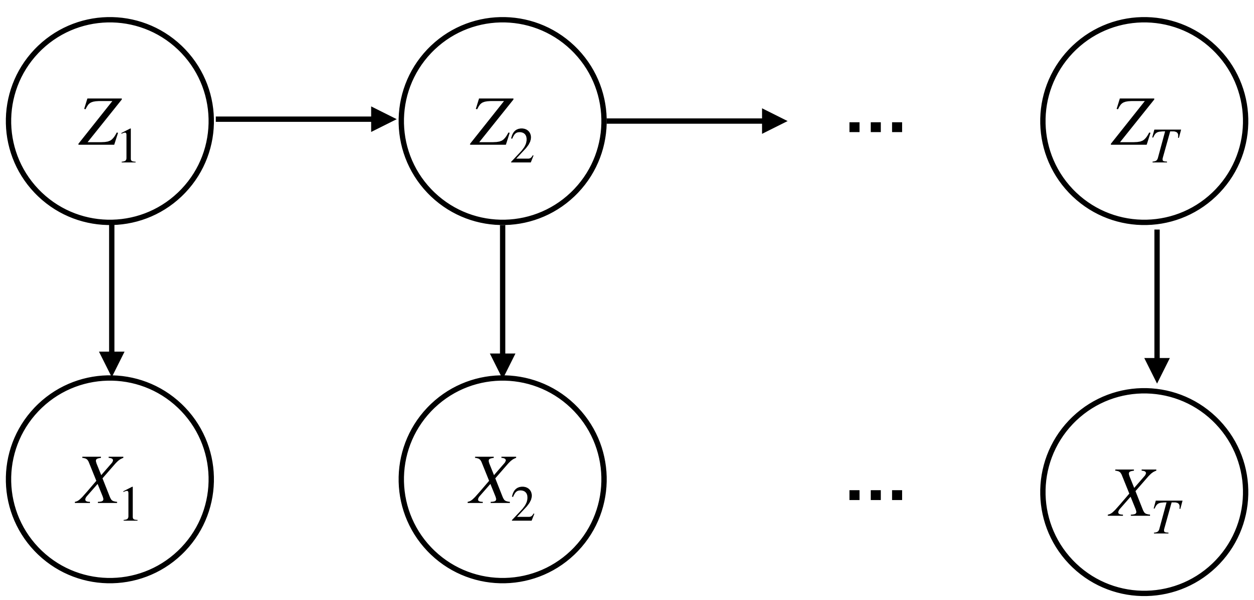
Supervised-like training

$$\boldsymbol{\theta} = \text{argmax} \sum_{m=1}^{M} \boxed{\sum_{k=1}^{K} q_k^{(m)}} \boxed{\log P(z^{(m)} = k, \boldsymbol{x}^{(m)}; \boldsymbol{\theta})}$$

k-means is a hard EM algorithm

The same heuristic for HMM

- Given any observations $x_1^{(m)}, \cdots, x_T^{(m)}$
  - Estimate $z_1^{(m)}, \cdots, z_T^{(m)}$ is not enough
- Learning HMM requires
  - Counting the initial tag
  - Counting the emissions
  - Counting the transitions => requires counting z_t, z_{t+1}



BP on factor/junction trees

Node-wise BP not adequate

EM in General

Loop until convergence

- E-step: Compute the expectation of sufficient statistics

  Compute $q(z \mid x) = P(z \mid x; \theta^{(t)})$

- M-step: Perform supervised learning with soft samples

  Maximize $\theta^{(t+1)} = \text{argmax}_\theta \, \mathbb{E}_{z \sim q(z|x)} P(x, z; \theta)$

EM in daily life

- Midterm mark is observed
- The solution of each question is latent
- Based on the observable mark, estimate the solution of each problem
- Update your belief
- Repeat until convergence

Despite the heuristics, we need a principled approach

Likelihood Estimation

- Fully observable likelihood: $P(x, z)$
- Partially observable likelihood: $P(x) = \sum_z P(x, z)$

$$\ell(\theta) = \sum_{m=1}^{M} \log P(x^{(m)}) = \sum_{m=1}^{M} \log \sum_{z^{(m)}} P(x^{(m)}, z^{(m)})$$

- Gradient descent works
  - You may use autodiff when the marginalization is analytically computed in a tractable way
  - Oftentimes the marginalization is not tractable. We need to other principled methods (e.g., EM) to solve the problem

EM maximizes the (joint) likelihood

$$\log p(\mathcal{D}) = \sum_{m=1}^{M} \log p(x^{(m)})$$

$$= \sum_{m=1}^{M} \left[\underbrace{\sum_z q(z|x^{(m)}) \log \frac{p(z, x^{(m)})}{q(z|x^{(m)})}}_{\mathcal{L}^{(m)}(q,\theta):\ \text{variational lower bound}} + \underbrace{\sum_z q(z|x^{(m)}) \log \frac{q(z|x^{(m)})}{p(z|x^{(m)})}}_{KL(q(z|x^{(m)}) \| p(z|x^{(m)})) \geq 0} \right]$$

$$= \sum_{m=1}^{M} \left[\mathcal{L}^{(m)}(q, \theta) + \underbrace{KL(q(z|x^{(m)}) \| p(z|x^{(m)}))}_{\text{KL between true posterior \& variational posterior}} \right]$$

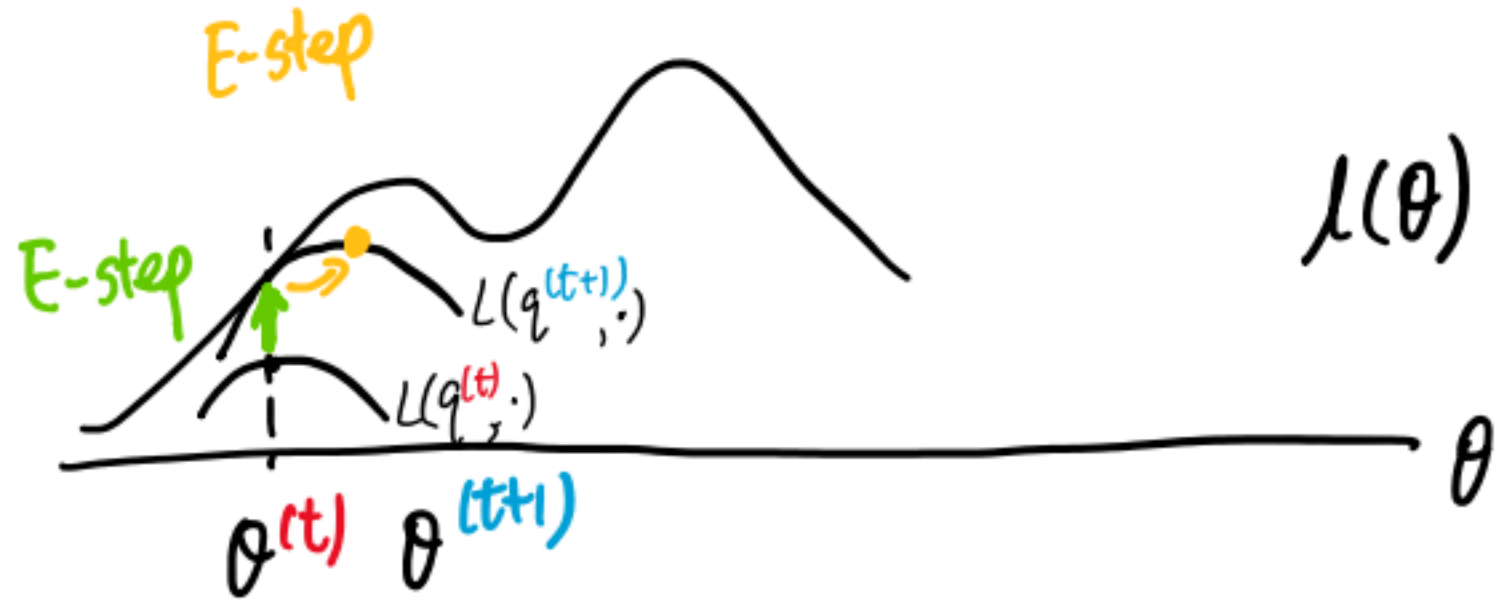
$$\ell(\boldsymbol{\theta}^{(t+1)}) = \sum_m \log p(x^{(m)}; \theta^{(t+1)})$$

$$\geq \sum_m \sum_z q^{(t+1)}(z|x^{(m)}) \log \frac{p(z, x^{(m)}; \boldsymbol{\theta}^{(t+1)})}{q^{(t+1)}(z|x^{(m)})} \quad \text{[lower bounds hold for any } q\text{]}$$

$$\geq \sum_m \sum_z q^{(t+1)}(z|x^{(m)}) \log \frac{p(z, x^{(m)}; \boldsymbol{\theta}^{(t)})}{q^{(t+1)}(z|x^{(m)})} \quad \text{[due to M-step]}$$

$$= \ell(\boldsymbol{\theta}^{(t)}) \quad \text{[due to E-step making lower-bound tight]}$$

EM is MM (minorize-maximization)



A function g minorizes function f at y if

- $g(x) \leq f(x)$ for every x in the domain
- $g(y) = f(y)$

EM also Works for MN

E-step: Compute $q(z \mid x) = P(z \mid x; \theta^{(t)})$

Inference algorithms are near identical in BNs and MNs

M-step: $\theta^{(t+1)} = \mathrm{argmax}_\theta \, \mathbb{E}_{z \sim q(z|x)} \log P(x, z; \theta)$

The log-linear representation assumes

$$P(x, z; \theta) = \frac{1}{Z} \exp\left\{ \sum_i \theta_i f_i(x, z) \right\}$$

Then,

$$\mathbb{E}_{z \sim q(z|x)} P(x, z; \theta) = \mathbb{E}_{z \sim q(z|x;\theta)}[\theta_i f_i] - A(\theta)$$

Observation 1: MNs do not have closed form solution; require gradient-based optimization

Observation 2: We won't do M-step from random initialization. We'll do keep the last $\theta^{(t)}$

Gradient Descent for Partially Observable MNs

Randomly initialize $\theta^{(0)}$

For iteration $t = 0,1,2,\cdots$ with a sample x:

$$\theta^{(t+1)} = \theta^{(t)} - \alpha \cdot g(\theta^{(t)})$$

where

$$g(\theta^{(t)}) = -\,\mathbb{E}_{z\sim P(z|x;\theta^{(t)})}[f_i(x,z)] + \mathbb{E}_{x,z\sim P(x,z;\theta^{(t)})}[f_i(x,z)]$$

Expectation in data **Expectation in model**

Example: Restricted Boltzmann Machine