Probabilistic Graphical Models

Lecture 16: Learning: Partial Observability

Matthew Walter

TTI-Chicago

June 4, 2020

Maximum Likelihood Estimation

- We've seen (e.g., Lecture 13) that choosing parameters that maximize the empirical log-likelihood of data is an effective approach to learning

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \frac{1}{|\mathcal{D}|} \sum_{m=1}^{|\mathcal{D}|} \log \hat{P}(\boldsymbol{\xi}; \boldsymbol{\theta})$$

- Suppose instead that the joint distribution was

$$P(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta})$$

where $\mathcal{D}$ provides samples of $\boldsymbol{X}$ but $\boldsymbol{Z}$ is never observed, i.e.,

$$\mathcal{D} = \{(0,1,0,?,?,?), (1,1,1,?,?,?), (0,1,1,?,?,?), \ldots\}$$

- Assume also that the hidden variables are *missing completely at random* (otherwise, we should model *why* these values are missing)

Identifiability

- In the fully observed, IID case, as $|\mathcal{D}| \to \infty$, the empirical log-likelihood approaches the true expected log-likelihood
- In the partially observed setting, what happens if we have infinite data?
- Is it possible to uniquely identify the true parameters?

Maximum Likelihood

- We can still use the same maximum likelihood approach
- The objective that we are maximizing becomes

$$\ell(\boldsymbol{\theta}) = \frac{1}{|\mathcal{D}|} \sum_{m=1}^{|\mathcal{D}|} \log \sum_{\boldsymbol{Z}} P(\boldsymbol{X}^{(m)}, \boldsymbol{Z}; \boldsymbol{\theta})$$

- For Bayesian networks, as a result of the marginalization over $\boldsymbol{Z}$:
  - the objective is no longer locally or globally decomposable
  - there is no longer a closed-form solution for $\boldsymbol{\theta}^*$
- Furthermore, the objective is no longer convex, and may have a different mode for every possible assignment $\boldsymbol{Z}$
- One approach is to employ gradient ascent as a general purpose optimization method to reach a local maxima of $\ell(\boldsymbol{\theta})$

Expectation Maximization

- The expectation maximization (EM) algorithm provides an alternative approach to finding the local maximum of $\ell(\boldsymbol{\theta})$
- EM is particularly useful when a closed-form solution for $\boldsymbol{\theta}^{\mathsf{ML}}$ exists in the fully observed setting
- For example, in Bayesian networks, we have the following

$$\hat{\theta}^{\mathsf{ML}}_{x \mid \boldsymbol{u}} = \frac{\#[x, \boldsymbol{u}]}{\#[\boldsymbol{u}]}$$

where $\boldsymbol{U}$ are the parents of X

Expectation Maximization

The EM algorithm follows as

1. Write down the *complete log-likelihood* $\log P(\boldsymbol{X}, \boldsymbol{Z}; \boldsymbol{\theta})$ in a way that is linear in $\boldsymbol{Z}$
2. Initialize $\boldsymbol{\theta}_0$ at random or using a heuristic
3. Repeat until convergence

$$\theta_{t+1} = \arg\max_{\boldsymbol{\theta}} \sum_{m=1}^{|\mathcal{D}|} \mathbb{E}_{P(\boldsymbol{Z}^{(m)} \mid \boldsymbol{X}^{(m)}; \boldsymbol{\theta}_t)} \left[\log P(\boldsymbol{X}^{(m)}), \boldsymbol{Z}; \boldsymbol{\theta} \right]$$

- Notice that $\log P(\boldsymbol{X}^{(m)}), \boldsymbol{Z}; \boldsymbol{\theta}$ is a random function because $\boldsymbol{Z}$ is unknown
- By linearity of expectation, the objective decomposes into expectation terms and data terms
- "E" stem corresponds to computing the objective (i.e., the *expectations*)
- "M" step corresponds to *maximizing* the objective