Machine Learning for Data Science (CS4786)
Lecture 25

Graphical Models: Approximate Inference

Course Webpage :
http://www.cs.cornell.edu/Courses/cs4786/2016sp/

Announcement

- Competition I was hard, $\ldots$, but it was real world data
- We don't care about your kaggle rank but only care about what you tried and how. This is what matter for your grades
- Competition II is synthetic data designed to be **easy**.
- Will be released tomorrow and due on May 20th
- Data is sequence data generated from HMM
- Your goal is to fill in missing values
- Very small percentage sequences are reversed!
- Report ≈ 5 pages, think of it as a proxy for final exam.
- Don't spend more than 6-7 hours on this.

Inference and Learning in Graphical Models

- Model data as a graphical model (use hidden or latent varibles)
- Inference:
  - What is the probability of some unobserved variable(s) given/conditioned on observation
  - What are the marginal probability of variables in the model
- Learning: based on observation pick the best parameters that explain the data
  - MLE:

$$\theta^* = \text{argmax}_{\theta \in \Theta} P(\text{Observations}|\theta)$$

  - MAP:

$$\begin{aligned}\theta^* &= \text{argmax}_{\theta \in \Theta} P(\theta|\text{Observations})\\ &= P(\theta|\text{Observations}) \times P(\theta)\end{aligned}$$

Learning in Graphical Models: EM

- Power of wishful thinking: start with a wild guess
- E-step: perform inference to infer distributions over latent variables given observation (under current guess of parameters)

$$Q^t(\text{Latent}) = P_{\theta^{t-1}}(\text{Latent}|\text{Observation})$$

- Under the inferred distribution over latent variables, find parameters that optimize joint likelihood of variables

$$\begin{aligned}\theta^t &= \text{argmax}_{\theta\in\Theta} \sum_{\text{Latent}} Q^t(\text{Latent}) \log P_\theta(\text{Observed}, \text{Latent}) \\ &= \text{argmax}_{\theta\in\Theta} \mathbb{E}_{\text{Latent}\sim Q^t}\left[\log P_\theta(\text{Observed}, \text{Latent})\right]\end{aligned}$$

Inference required for EM (learning in general)

Exact Inference

Calculate the marginals/conditionals given parameter exactly

- Variable Elimination:
  - Always guaranteed to work
  - Can be computationally prohibitive
- Belief Propagation/Message Passing
  - Guaranteed to work only on tree structures and few other structure
  - Highly parallelizable, for many problems works well in practice

Exact inference in worst case is computationally hard!

Approximate Inference

Two approaches:

- Inference via sampling:
  generate instances from the model, compute marginals
- Use exact inference but move to a close enough simplified model

Inference Via Sampling

- Law of large numbers: empirical distribution using large samples approximates the true distribution
- Some approaches:
  - Rejection sampling: sample all the variables, retain only ones that match evidence
  - Importance sampling: Sample from a different distribution but then apply correction while computing empirical marginals
  - Gibbs sampling: iteratively sample from distributions closer and closer to the true one

Gibbs Sampling

- Fix values of observed variables v to the observations ($x_v^1 = x_v$)
- Randomly initialize all other variables u by randomly sampling x_u^1
- For $t = 2$ to n
- For $i = 1$ to N

  If X_i is observed set

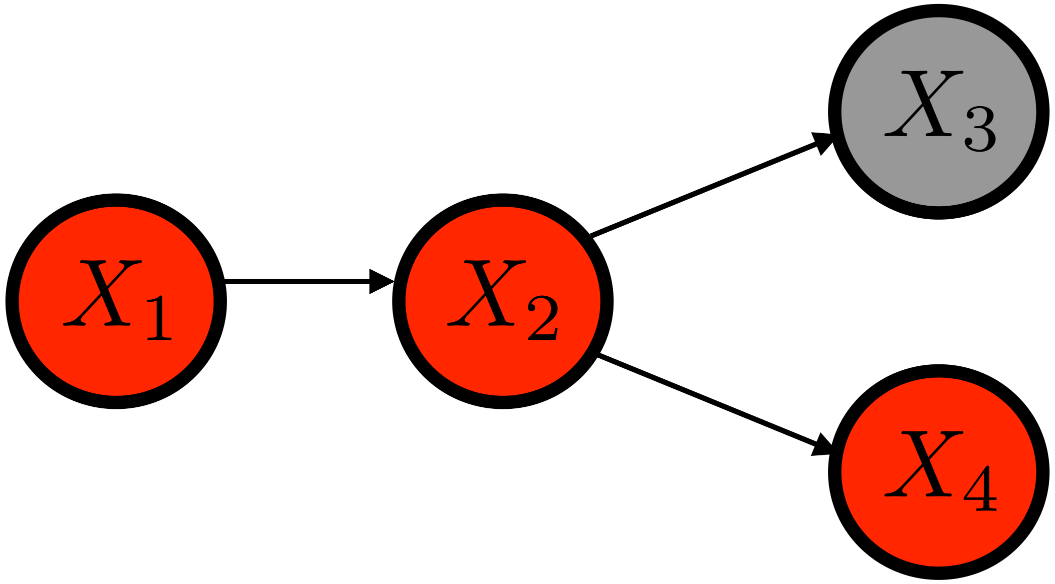
$$x_i^t = x_i^{t-1}$$

  Else sample x_i^{t+1} from

$$x_i^{t+1} \sim P(X_i | X_1 = x_1^{t+1}, \ldots, X_{i-1} = x_{i-1}^{t+1}, X_{i+1} = x_{i+1}^t, \ldots, X_N = x_N^t)$$

  - End For
- End For
- Take $(x_1^n, \ldots, x_N^n)$ as one sample and repeat

Gibbs Sampling for Bayesian Networks



Notice that:

$$
\begin{aligned}
&P(X_i = x_i | X_1 = x_1^{t+1}, \ldots, X_{i-1} = x_{i-1}^{t+1}, X_{i+1} = x_{i+1}^{t}, \ldots, X_N = x_N^t) \\
&\quad \propto P(X_i = x_i, X_1 = x_1^{t+1}, \ldots, X_{i-1} = x_{i-1}^{t+1}, X_{i+1} = x_{i+1}^{t}, \ldots, X_N = x_N^t) \\
&\quad \propto P(X_i = x_i | X_{\text{Parents}(i)} = x_{\text{Parents}(i)}^{t+1}) \times \prod_{j \in \text{Child}(X_i)} P(X_j = x_j^t | X_{\text{Parents}(j)}, X_i = x_i)
\end{aligned}
$$

MCMC Sampling in General

- Gibbs sampling belongs o a class of methods called Markov Chain Monte Carlo methods
- We start by sampling from some simple distribution
- Set up a markov chain whose stationary distribution is the target distribution
- That is, based on previous sample (state) we transit to the next state, and then to the next state and so on
- If the transition probabilities are set up right, after multiple transitions, our sample looks like one from target distribution

Approximate Inference

- Variational inference:
  - Instead of true posterior, calculate posterior in a restricted family of distributions close to true one
  - Latent variables get their own set of parameters which we pick on the fly to make then close to true posterior
- Approximate message passing, expectation propagation, …

Variational Inference

- Basic idea: we want to infer $P(\text{Unobserved}|\text{Observed})$
  We create a new parametric distribution $Q_\theta(\text{Unobserved})$ where
  θ is picked based on Obervations
- We pick θ such that, Q_θ is close to $P(\text{Unobserved}|\text{Observed})$
- Closeness measured using KL divergence
- Mean-field approximation,

$$Q_\theta(X_1, \ldots, X_m) = \prod_{j=1}^{m} Q_{\theta_j}(X_j)$$