COMPSCI 688: Probabilistic Graphical Models

Lecture 24: Final Review

Dan Sheldon

Manning College of Information and Computer Sciences
University of Massachusetts Amherst

Partially based on materials by Benjamin M. Marlin (marlin@cs.umass.edu) and Justin Domke (domke@cs.umass.edu)

Logistics

Exam Logistics

- Monday, 5/13/2023
- 1:00pm - 3:00PM
- Goessmann Lab Room 20 (here)

Exam Procedures

- Open notes
- No electronic devices
- Paper exam: bring pencils, erasers, etc.
- Double seating (spread out, leave every other seat/row empty)

Review

Foundations

- sample space, events, random variables, probability mass function
- independence, conditional independence
- joint, marginal, and conditional distributions; chain rule
- density function, expectation, variance, covariance

Directed Models

$$p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i | x_1, \ldots, x_{i-1}) \quad \text{chain rule (always true)}$$

- Assume $p(x_i | x_1, \ldots, x_{i-1}) = p(x_i | x_{\text{pa}(i)})$, then

$$p(x_1, \ldots, x_D) = \prod_{i=1}^{D} p(x_i | x_{\text{pa}(i)}).$$

- Starting CIs (conditional independencies) imply other CIs.
- Use D-separation to test if any CI is implied by the graph.
- ML (Maximum likelihood) = KL divergence minimization
- Can do ML learning by "counting"

Undirected Models

- Global Markov property: $X_i \perp X_{-i} | X_{\text{nb}(i)}$.
- MRF (Markov Random Field): $p(x) = \frac{1}{Z} \prod_{c \in C} \phi(x_c)$.
- Hammersley–Clifford theorem: The global Markov property implies that p can be written as an MRF. (And vice-versa.) (all assuming $p(x) > 0$)
- CRF (conditional random field): $p_\theta(y|x) = \frac{1}{Z(x,\theta)} \prod_{c \in C} \phi_c(x, y_c; \theta)$
- Why a CRF might be better (or worse) than an MRF.

Inference

- If Bayes net, first convert to MRF
- Conditioning = factor reduction
- Variable elimination = general purpose inference algorithm:
  - marginalize variables one-by-one in some order
  - replace set of factors containing each variable by one new factor

Message Passing

- $p(x) = \frac{1}{Z} \prod_i \phi_i(x_i) \prod_{(i,j)} \phi(x_i, x_j)$. (assume tree structures.)
- Derivation principle: leaf-first variable elimination
- $m_{i \to j}(x_i) = \sum_{x_i} \phi(x_i)\phi(x_i, x_j) \prod_{l \in \text{nb}(i) \setminus j} m_{l \to i}(x_i)$.
- Once we have passed messages, we can compute $Z, p(x_i), p(x_i, x_j)$.

Exponential Family

- $p_\theta(x) = h(x) \exp(\theta^\top T(x) - A(\theta))$.
- $\frac{\partial A}{\partial \theta} = \mathbb{E}\,[T(X)] \quad \frac{\partial^2 A}{\partial \theta \partial \theta^\top} = \text{Var}[T(X)]$.
- Moment-matching conditions for ML.
- A MRF $p(x) = \frac{1}{Z} \prod_{c \in C} \phi_c(x_c)$ can be written as an exponential family with

$$T(x) = [\ \mathbb{I}[x_c = a], \quad \forall c \in C, \quad \forall a \in \text{Val}(X_c)]$$

Directed vs Undirected

	Directed	Undirected
assume	$X_i \perp X_{\mathsf{nd}(i)} \mid X_{\mathsf{pa}(i)}$	$X_i \perp X_{-i} \mid X_{\mathsf{nb}(i)}$
$p(x)$	$\prod_i p(x_i \mid x_{\mathsf{pa}(i)})$	$\frac{1}{Z} \prod_{c \in C} \phi_c(x_c)$
ML condition	$p(x_i \mid x_{\mathsf{pa}(i)}) = \frac{\#(X_i = x_i, X_{\mathsf{pa}(i)} = x_{\mathsf{pa}(i)})}{\#(X_{\mathsf{pa}(i)} = x_{\mathsf{pa}(i)})}$	$p(x_c) = \frac{\#(X_c = x_c)}{N}$

Markov chain Monte Carlo (MCMC)

- Want to sample $x \sim p(x)$.
- Create a random walk that will approximately do this.
- Regularity: $\exists n$ st $(T^n)_{ij} > 0$ for all i, j. $\Longrightarrow$ Markov chain has only one stationary distribution.
- Detailed balance: $p(x)T(x'|x) = p(x')T(x|x')$. $\Longrightarrow$ p stationary distribution of T.
- Gibbs / Metropolis-Hastings sampling
- Hamiltonian Monte Carlo: M-H that uses $\nabla \log p$ to make big steps

Bayesian Inference

- You create $p(\theta, \mathrm{Data})$.
- You observe Data.
- You make predictions using $p(\theta|\mathrm{Data})$.

- Prior, likelihood, posterior
- Conjugate inference: prior and posterior in same family
- "Easy way": drop terms that don't involve θ, combine prior/likelihood to get posterior

Variational Inference (VI)

- ELBO decomposition: $\log p(x) = \mathrm{ELBO} + \mathrm{KL}$. Gives both:
  - bound on $\log p(x)$
  - approximation of $p(z|x)$
- Black-box VI: Optimize ELBO using stochastic gradients
- Reparameterization estimation

GPs and Flows

Not on the exam

Question Topics

Expect ~10 questions drawn from ones like these:

- Bayes nets
  - given a graph, write factorization
  - given factorization, draw graph
  - answer questions about CPTs
  - given graph, which CI properties hold?
  - answer queries using distribution properties
- Markov networks
  - given a graph, write factorization
  - given factorization, draw graph
  - given graph, which CI properties hold?
  - answer queries
  - conceptual questions about CRFs vs MRFs

- Inference
  - Conceptual questions about variable elimination and its efficiency
  - Conceptual questions about message passing
  - Execute message-passing on toy model and/or use messages to get marginals, Z
- Learning
  - Conceptual questions about log-likelihood, KL divergence, conditional log-likelihood
  - Use Bayes net learning rule on toy problem
  - Derive log-likelihood and gradient for some simple graphical model
- Exponential families
  - Conceptual questions about exponential families and their properties
  - Given a sample distribution, is θ an optima of likelihood? (use moment matching)

- MCMC
  - Conceptual questions about Markov chains, stationary distributions
  - Is a Markov chain regular?
  - Does a Markov chain satisfy detailed balance with respect to a distribution?
  - Conceptual questions about Gibbs sampler, Metropolis-Hastings
  - Apply Gibbs sampler of Metropolis-Hastings to a toy problem
- Bayesian inference
  - Conceptual questions about latent variable models, prior distributions, likelihood functions, posterior distributions
  - Derive formula for posterior distribution in conjugate Bayesian model
- Variational inference
  - Conceptual questions about ELBO, mean-field VI, stochastic gradient VI, reparameterization trick
  - Given simple p and q, derive ELBO, or gradient estimator