Inference and Representation: Recap

Rahul G. Krishnan

New York University

Lab 8, October 26, 2016

Outline

Random Variables

- We began by representing objects of interest in our world as random variables
- Univariate or Multivariate
- Discrete or Categorical
- Having them in isolation does not let us represent our entire environment

Bayesian networks

- Directed, generative (potentially causal) process for our data
- Joint distribution factorizes as $p(X_1, \ldots, X_N) = \prod_i p(X_i | X_{pa(i)})$
- Conditional independance statements in the random variables are encoded in the graph structure
- Three different structures for inference : chain, common child (v-structure), common parent
- Markov Blanket: Parents, Children and Co-parents
- Parameterize the CPDs (tables, logistic function, neural network)

Markov Random Fields

- Undirected graphical model
- Joint distribution factorizes over cliques
  $p(X_1, \ldots, X_N) = \prod_{c \in \mathcal{C}(G)} \phi_c(X_c)$
- Conditional independance as graph separation
- Markov Blanket: Neighbors
- Different ways to parameterize this distribution: Ising Model etc.

BN $\rightarrow$ MRF

- Bayesian networks may be moralized to form Markov Random Fields
- Lose v-structures (not an invertible process)
- Different graphical models allow us to admit different conditional independance statements

Inference

- We want to estimate the result of some probablistic query $p(X_i|X_1, X_2)$
- Often this query is conditioned on some evidence
- Bayes Ball: Algorithm to check d-separation via rolling a ball on a Bayesian network. Gives us a way to measure influence of a random variable on another.
- If the random variables are d-separated under the query, they are (conditionally) independant

Algorithm 1: Variable Elimination for Exact Inference

- Variable Elimination: Moralize Bayesian network if necessary
- Elimination Ordering: Ordering in which variables will be eliminated
- Elimination Process: Collect factors which contain a variable, sum out the influence of the variable on their product and create an intermediate factor
- Fill edges: Additional edges added between nodes to accomodate intermediate factors
- Treewidth of the graph: Measure of complexity for exact inference. Number of nodes in the largest intermediate factor during exact inference
- Chordal graphs: Graphs for which there exists an elimination ordering that yields no fill edges

Algorithm 2: Belief Propagation

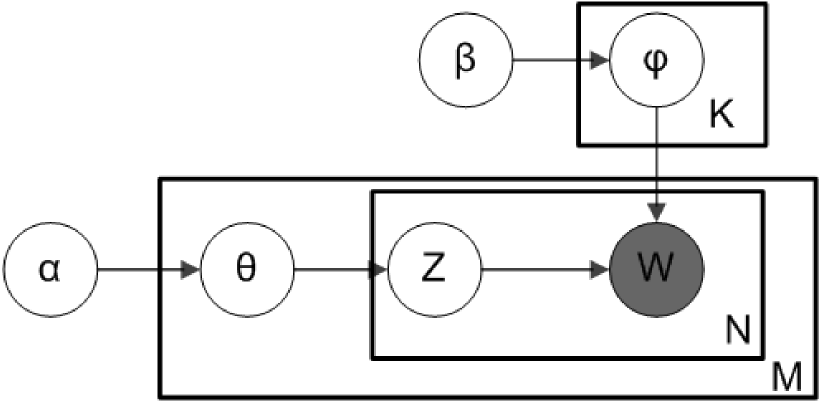
- *Exact* on tree structured MRFs and *approximate* on loopy graphs
- Runs by passing messages on graphs. Start from leaf send to root and back
- Marginal probabilities estimated by taking the product of incoming messages from all neighbors along each node
- Can be viewed as a way to cache computation taking place in variable elimination

Algorithm 3: Gibbs Sampling

- An algorithm for inference in Bayesian networks.
- Inference via Gibbs Sampling : Sample from $p(X_i|X_{\neg i})$ to estimate marginal probabilities
- Is an example of a Markov Chain Monte Carlo algorithm

Topic Models: Latent Dirichlet Allocation

- Unsupervised learning, interpretable generative model
- Bag-of-Words assumption on documents



Jensen's Inequality

- Jensen's Inequality: For concave f, we have

$$f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$$

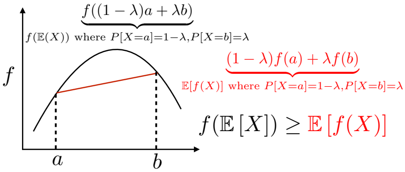


Figure: Jensen's Inequality

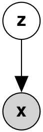
Maximum Likelihood

- We assume that for $\mathcal{D} = \{x_1, \ldots, x_N\}$, $x_i \sim p(x)$ i.i.d
- We hypothesize a model (with parameters θ) for how the data is generated
- The Maximum Likelihood Principle: $\max_\theta p(\mathcal{D}; \theta) = \prod_{i=1}^{N} p(x_i; \theta)$
- Typically work with the log probability: i.e $\max_\theta \sum_{i=1}^{N} \log p(x_i; \theta)$

ML for Learning in Bayesian networks

- **Fully Observed Model:** The factorization of the joint distribution implies the maximizations can be distributed. For a Bayesian network with parameters θ
  $\arg\max_\theta \log p(x_1, \ldots, x_N; \theta) = \sum_i \arg\max_{\theta_i} \log p(x_i | x_{pa(i)}; \theta_i)$
- **Latent Variable Model:** Challenging Integral.
  $\arg\max_\theta p(x; \theta) = \arg\max_\theta \int_z p(x, z; \theta)$

A simple Bayesian Network



- We assume that the data is generated i.i.d as:

$$z \sim p(z) \qquad x \sim p(x|z)$$

- z is latent/hidden and x is observed. Corresponds to FA, PPCA etc.

Bounding the Marginal Likelihood

- Log-Likelihood of a single datapoint $x \in \mathcal{D}$ under the model: $\log p(x; \theta)$
- Important: Assume $\exists q(z)$

$$\log p(x) = \log \int_z p(x, z) \text{ (Multiply and divide by } q(z))$$

$$= \log \int_z \frac{q(z)p(x,z)}{q(z)} = \log \mathbb{E}_{z \sim q(z)} \left[\frac{p(x,z)}{q(z)} \right] \text{ (By Jensen's Inequality)}$$

$$\geq \int_z q(z) \log \frac{p(x,z)}{q(z)} = \mathcal{L}(x; \theta)$$

$$= \underbrace{\mathbb{E}_{q(z)}[\log p(x,z)]}_{\text{Expectation of Joint distribution}} + \underbrace{\mathrm{H}(q(z))}_{\text{Entropy}}$$

Evidence Lower BOund (ELBO)/Variational Bound

- When is the lower bound tight?
- Look at: function - lower bound

$$\log p(x;\theta) - \mathcal{L}(x;\theta)$$

$$\begin{aligned}
&\log p(x) - \int_z q(z) \log \frac{p(x,z)}{q(z)} \\
&= \int_z q(z) \log p(x) - \int_z q(z) \log \frac{p(x,z)}{q(z)} \\
&= \int_z q(z) \log \frac{q(z)p(x)}{p(x,z)} \\
&= \mathrm{KL}(q(z)||p(z|x))
\end{aligned}$$

Overview

Key Point

The optimal $q(z)$ corresponds to the one that realizes $\text{KL}(q(z)||p(z|x)) = 0 \iff q(z) = p(z|x)$

- In order to estimate the liklihood of the entire dataset $\mathcal{D}$, we need $\sum_{i=1}^{N} \log p(x_i;\theta)$
- Summing up over datapoints we get:

$$\max_\theta \sum_{i=1}^{N} \log p(x_i;\theta) \geq \max_\theta \underbrace{\sum_{i=1}^{N} \mathcal{L}(x_i, q(z_i), \theta)}_{ELBO}$$

Expectation Maximization

- Is an iterative procedure for the maximization of ELBO.
- Consider learning in the context of a single data point x and let k index time
- **E step:** $q^{(k)}(z) = \arg\max_{q(z)} \mathcal{L}(x, q(z), \theta^{(k-1)})$
- **M step:** $\theta^{(k)} = \arg\max_{\theta} \mathcal{L}(x, q^{(k)}(z), \theta)$
- Repeat till convergence

Summary

- **Fully Observed Bayesian Networks**
- **Latent Variable Models**
  - **Analytic Posterior Distribution:** Solve E step exactly to find the optimal q and then perform co-ordinate maximization over θ
  - **Intractable Posterior Distribution:** Choose one of the following ways to sample
    - **Variational EM:** Approximate the complex posterior distribution with a simpler family of distributions (variational distributions) to approximately maximize E-step
    - **MCMC:** Form a Monte-Carlo approximation to the expectation in ELBO using MCMC to sample from the posterior distribution

Graphical Model

- $z \sim \mathcal{N}(0, \mathbb{I})$
- $x \sim \mathcal{N}(Wz + \mu, \sigma^2 \mathbb{I})$
- Define $\Psi = \sigma^2 \mathbb{I}$
- For a data point x_i, the posterior distribution under the model may be obtained analytically

E & M step

- **E Step:** The posterior distribution is analytic.
- $q(z_i) = \mathcal{N}(\mu_{z_i|x_i}, \Sigma_{z_i|x_i})$ where
  $\mu_{z_i|x_i} = W^T(WW^T + \Psi)^{-1}(x_i - \mu)$,
  $\Sigma_{z_i|x_i} = \mathbb{I} - W^T(WW^T + \Psi)^{-1}W$
- **M Step:** Maximize ELBO over all the data points with respect to W.
- $W^* = \arg\max_W \sum_{n=1}^{N} \mathbb{E}_{q(z_i)}[\frac{\log p(x_i,z_i)}{q(z_i)}] + \underbrace{H(q(z_i))}_{\text{const. wrt. } \theta}$
- Take gradients, set to 0, solve for W to yield:
  $W^* = (\sum_{n=1}^{N}(x_i - \mu)\mu_{z_i|x_i}^T)(\sum_{n=1}^{N} \mu_{z_i|x_i}\mu_{z_i|x_i}^T + \Sigma_{z_i|x_i})^{-1}$

References

- Andrew Ng's Coursera Handouts: https://see.stanford.edu/materials/aimlcs229/cs229-notes9.pdf