Notes on Inference and Representation

Aodong Li

June 9, 2018

This document only serves as informal personal thoughts or notes.

Bayesian Network

- Key idea:
  - Represent the whole world as a collection of observed and unobserved variables with a joint distribution $P(X_1, .., X_n, Z_1, ..., Z_m)$.
  - Learn this distribution from data.
  - Perform inference to compute posterior distribution $P(Z|X_1 = x_1, ..., X_n = x_n)$.
- The introduction of the latent variables can greatly reinforce the power of the model and allows us to represent the complex distribution using very simple distributions at each node. For example, the marginal distribution of Gaussian mixtures is able to represent complicated multi-modal distributions but the joint distribution is tractable.
- If we model the variables independent, it is might not be useful because the connection between variables are cut. However, if we introduce conditional independence by introducing an additional latent variable, connections are recovered and the model is simple but powerful.
- V-structure accounts for the phenomena of explaining away, which makes the Bayesian Network so powerful.
- d-separation reduces statistical independencies (hard) to connectivity in graphs (easy).
- d-separation is important because it allows us to quickly prune the Bayesian network, finding just the relevant variables for answering a query.
- Different Bayesian Networks can be equivalent in that they encode precisely the same conditional independence assertions. The notations of the nodes are arbitrary.
- 1) If two networks have the same skeleton and v-structure, then the two networks are equivalent. But the reverse is not necessarily true.

  2) If two networks have the same skeleton and immoralities if and only if the two networks are equivalent.

Markov Random Field

- Bayesian Network has limitations: Not every distribution has a perfect map as a DAG. Prove this by counterexample. BN cannot satisfy such a relationship at the same time $A \perp C|\{B, D\}$ and $B \perp D|\{A, C\}$.
- Markov Random Field/Markov Networks (MRF/MN) has potential functions over cliques and partition functions as normalization. The normalization makes the potential scale invariant.
- MRF does not show the causal relationship but the correlation of variables **within the clique**. For example, the potential function encourages the co-occurance of some values.
- The conditional independence is specified by graph separation.
- The probability of a variable conditioned on its Markov blanket depends only on the potentials involving that node.
- Pairwise MRF only involves node potentials and pairwise potentials, and every discrete MRF can be transformed into pairwise MRF by introducing an additional node for each clique. (See hw2)
- MRF does not reveal the structure of the distribution. So factor graph comes into play by defining a potential on each factor node.
- Moralization: Marry the parents and introduce one potential for each CPD.
- Factorization and Conditional independence (Separation in graph)
    1. Soundness of separation: If $p(x)$ is a Gibbs distribution for G, then G is an I-map for $p(x)$.
    2. HC: If $p(x)$ is a positive distribution and G is an I-map for $p(x)$, then $p(x)$ is a Gibbs distribution that factorizes over G.
    3. 1) Specify a graph structure G; 2) Search the Gibbs distribution over G – parameterize the distribution.
- If each random variable conforms to a distribution in exponential family, the joint distribution of such iid random variables is in exponential family.
- We can learn model parameters for a fixed structure, or both the structure and model parameters.
- Density estimation: minimizing KL-divergence $D(p^*||p_\theta)$ equals maximizing expected log-likelihood $\mathbb{E}_{p^*} \log p_\theta(x)$. Because of log, samples x where $p_\theta(x)0$ weigh heavily in objective.

- Since we have iid samples, we can compute empirical log-likelihood:

$$\mathbb{E}_{p^*} \log p_\theta(x) = \frac{1}{|D|} \sum_{x \in D} \log p_\theta(x).$$

- Because BN has well-defined distribution for each node, that is, nodes are clear from a whole normalization, ML optimization decomposes into an independent optimization over each node. However, the occurance of normalization constant impedes the ML estimation for MRF.

Exact Inference

- Naively marginalizing over other unrelated/unobserved variables requires an exponential number of computations.
- 3-SAT indicates there exist NP-hard inference problem.
- In practice, tree-structured graph can be performed inference in linear time.
- HMM

1. filter problem by DFS. Given $Y_1 = y_1, ..., Y_n = y_n$,

$$\begin{aligned} p(X_n, y_1, ..., y_n) &= \sum_{X_{n-1}} p(X_{n-1}, y_1, ..., y_{n-1}) p(X_n, y_n | X_{n-1}, y_1, ..., y_{n-1}) \\ &= \sum_{X_{n-1}} p(X_{n-1}, y_1, ..., y_{n-1}) p(X_n, y_n | X_{n-1}) \\ &= \sum_{X_{n-1}} p(X_{n-1}, y_1, ..., y_{n-1}) p(X_n | X_{n-1}) p(y_n | X_n) \end{aligned}$$

   The initiliztion is $p(X_1, y_1) = p(X_1) p(y_1 | X_1)$.

2. MAP inference by Viterbi algorithm (dynamic programming). Given $Y_1 = y_1, ..., Y_n = y_n$,

$$\begin{aligned} \arg\max_X p(X_1, ..., X_n | y_1, ..., y_n) &= \arg\max_X p(X_1, ..., X_n, y_1, ..., y_n) \\ &= \arg\max_X \log p(X_1, ..., X_n, y_1, ..., y_n) \\ &= \arg\max_X \log p(X_1, ..., X_{n-1}, y_1, ..., y_{n-1}) p(X_n | X_{n-1}) p(y_n | X_n) \\ &= \arg\max_X \log p(X_1) p(y_1 | X_1) + \sum_{i=2}^{n} p(X_i | X_{i-1}) p(y_i | X_i) \end{aligned}$$

- Running time of VE (Variable Elimination) depends on the graph structure.
- Uses dynamic programming to circumvent enumerating all assignments.

- Key idea: push the summation inside the product. This procedure is dynamic programming: computation is inside out instead of outside in – cache the computations that are otherwise computed exponentially many times.
- Factor marginalization gives a new factor that do not contain marginalized factors.
- How to eliminate a factor: 1) multiply all the factors that involve the corresponding factor Z; 2) marginalize the Z, generating a smaller factor; 3) replace the old factors with the new factor.
- The computation complexity $O(mk^N)$ depends on (is exponential of) the size N of the largest clique of the induced graph, thus depends on the elimination order.
- The treewidth provides a bound on the best running time achievable by VE on a distribution that factorizes over G.
- Doing exact inference is NP-hard, instead, we do approximate inference: 1) Monte-carlo methods; 2) Variational inference.
- Monte-carlo algorithm:

$$\begin{aligned} p(X_1 = x_1) &= \sum_X p(X) f(X) \\ &= \mathbb{E}_X f(X), \end{aligned}$$

  where $f(X) = 1[X_1 = x_1]$. The estimate is $\frac{1}{M} \sum_{m=1}^{M} f(x^m)$ for M samples.
- MRF has the property that the maximum likelihood estimator for the empirical marginal distribution has moment matching property – $\hat{p}(x_i, x_j) = \text{count}(x_i, x_j)/N$.

Topic modeling and Gibbs sampling

- Monte-Carlo estimate is an unbiased estimate, i.e., the expectation is the same as the true value.
- Depending on whether we care about additive error or multiplicative error, we can use Hoeffding bound or Chernoff bound to express the convergence speed.
- Computing the conditional queries is very hard if the probability of the evidence is very small, by Chernoff bound. Same reasoning applies for undirected graphical model – the normalizer is difficult to compute because some assignments have very low probability.
- Gibbs sampling: when we update a variable, we use its new value for sampling other variables.
- The full conditional of variable x is just the Markov blanket of x. We can use this knowledge to simplify the computation.
- We can monitor the convergence by plotting samples from multiple MH runs.

- Every distribution in exponential family has conjugate prior. For example, Dirichlet distribution is the conjugate prior of Multinomial distribution.
- By Rao-Blackwell theorem, analytically integration of some unknown quantities decreases the variance of the MC estimate. This is where collapsed Gibbs sampling comes from. For topic model, integrating out θ and β gives an easier sampler.

Factor analysis

- Monte-Carlo estimate is an unbiased estimate but it may need way too many samples when some low probability evidence is involved, in order to reach a satisfactory error bar.
- Factor analysis for survey data is to extract interpretable, summary information out of a series of correlated survey responses. The latent factors are uncorrelated variables.
- We can think of each column of design/data matrix is a delegate of a random variable, or a feature.
- Principal component analysis
  - Simplest setting $X_l = \sum_{j=1}^{J} \alpha_{jl} Y_j, l = 1, ..., L$.
  - If X_l is Gaussian then Y_j is also Gaussian.
  - For Gaussian random variables, independence if and only if uncorrelation.
  - Let $Y = AX + b$, we have

    $$\mu_Y = A\mu_X + b$$
    $$\Sigma_Y = A\Sigma_X A^\top.$$

    Uncorrelated random vector Y implies diagonal covariance matrix Σ_Y. So our goal is to find such A to make Σ_Y diagonal.
  - Since $\Sigma_X \in \mathbf{S}_+$, eigendecomposition exists. A corresponds to the eigenvectors of the covariance matrix. But the decomposition is not unique, any orthogonal transformation on A and on Y will generate the same solution.
  - The empirical covariance matrix satisfies that for $||\hat{\Sigma}_N - \Sigma|| \leq \epsilon$, we need $O((\log\log L)^\alpha L) \approx O(L)$ samples. *It turns out that PCA does not suffer from the curse of dimensionality.*
  - Naively computing the eigen-decomposition requires $O(NL^2)+O(L^3)$. Computing the first p principal components by SVD gives $O(pNL)$.
  - We have the prior knowledge that the data matrix has a low rank. We can randomly sample points (approximate basis) to capture the range of the data matrix.
  - Sample Gaussian matrix $\Omega \in \mathbf{R}^{L\times(k+p)}$. The additional p points make sure $k+p$ to capture the whole range due to some noise or variation in the data matrix. The computation improves from $O(pNL)$ to $O(\log(p)NL)$.

- Factor analysis
  - PCA can be seen as a linear latent model

$$X_j = \sum_{l=1}^{L} \alpha_{jl} Y_l + \mu_j, j = 1, ..., J$$

    where Y is uncorrealted and unit variance.
  - The solution is lack of unicity. The underlying assumption is that data has low-rank, i.e., covariance directly reveals dependencies in data.
  - Factor analysis:

$$X_j = \sum_{l=1}^{L} \alpha_{jl} Y_l + \mu_j + \epsilon_j, j = 1, ..., J$$

    where Y is uncorrealted and unit variance and ϵ are uncorrelated and zero-mean.
  - The noise for each random variable captures the variability associated with that variable. Y act as common factors of variability.
  - We usually assume X and Y are Gaussian, then $\Sigma_X = AA^\top + \text{diag}(\beta)$ is a sufficient statistic. But it still suffers from lack of unicity.
- Only decorrelation between latent factors does not suffice to lead to unicity, because any orthogonal transformation is also uncorrelated.
- So we hope to gain independence among latent factors. But Gaussian is an exception because uncorrelation $\Leftrightarrow$ independence for Gaussian.
- Instead, we hope Y to be independent and non-Gaussian, which is a form of inverse central limit theorem method.
- The mutual information measures independence

$$I(Y) = \sum_{n=1}^{N} H(Y_n) - H(Y) \geq 0$$

  and if A is unitary and $Y = AX$, then $H(Y) = H(X)$.
- The challenge is computing entropy requires estimating the density, which is exposed to the curse of dimensionality.
- PCA captures the covariance structure while Factor analysis captures the correlation structure; PCA is sensitive to the scale of the data while factor analysis does not.

EM and MCMC

- Floyd algorithm for K-means is a specific instance of EM algorithm. In particular, we seek to minimize the cost function

$$C(r, c) = \sum_i \sum_k r_i(k)||x_i - c_k||^2$$

for which the optimization takes an alternative manner.

- Gaussian Mixture Model: Define the generation process as first generate an indicator for responsible cluster and then generate the observed data. The marginal density of the data looks like

$$\begin{aligned} p(z) &= \prod_{k=1}^{K} \pi_k^{z_k} \\ p(x|z) &= \prod_{k=1}^{K} \mathcal{N}(x; \mu_k, \Sigma_k)^{z_k} \\ p(x) &= \sum_z p(x|z)p(z) \\ &= \sum_z \prod_{k=1}^{K} (\pi_k \mathcal{N}(x; \mu_k, \Sigma_k))^{z_k} \\ &= \sum_{k=1}^{K} \pi_k \mathcal{N}(x; \mu_k, \Sigma_k). \end{aligned}$$

- Solve the maximum log-likelihood of GMM gives weighted average over points or empirical covariance. The wights in turn depend on the parameters.

- The EM algorithm tries to solve the log-likelihood problem with discrete latent variables. Specifically, we try to solve

$$\log p(X|\theta) = \log \sum_z p(X, Z|\theta).$$

The E-step computes the expectation of total likelihood of the model

$$\begin{aligned} Q(\theta, \theta^{old}) &= \mathbb{E}[\log p(X, Z|\theta)|X, \theta^{old}] \\ &= \sum_i \sum_k z_{ik} \log \pi_k + z_{ik} \log \mathcal{N}(x_i; \mu_k, \Sigma_k) \end{aligned}$$

where z_{ik} conforms to the posterior distribution.

The M-step maximizes the likelihood given by E-step.

- Variational Bound:

$$\begin{aligned}
\log p(X|\theta) &= \log \sum_Z p(X, Z|\theta) \\
&= \log \sum_Z \frac{p(X, Z|\theta) q(Z)}{q(Z)} \\
&\geq \sum_Z q(Z) \log \frac{p(X, Z|\theta)}{q(Z)} \\
&= L(q, \theta) \\
\log p(X|\theta) &= L(q, \theta) + KL(q(z)||p(z|x, \theta))
\end{aligned}$$

- To see the correctness of EM, we can prove that it is improved after each iteration. Take $q(z) = p(z|x, \theta^{(n)})$

$$\begin{aligned}
\log p(X|\theta) - \log p(X|\theta^{(n)}) &\geq L(p(z|x, \theta^{(n)})) - \log p(X|\theta^{(n)}) \\
&= \sum_z p(z|x, \theta^{(n)}) \log \frac{p(x, z|\theta)}{p(z|x, \theta^{(n)}) p(x|\theta^{(n)})}
\end{aligned}$$

  with $\theta = \theta^{(n)}$ the RHS equals zero.

  Thus,

$$\begin{aligned}
\theta^{(n+1)} &= \arg\max_\theta \sum_z p(z|x, \theta^{(n)}) \log \frac{p(x, z|\theta)}{p(z|x, \theta^{(n)}) p(x|\theta^{(n)})} \\
&= \arg\max_\theta \sum_z p(z|x, \theta^{(n)}) \log p(x, z|\theta) \\
&= \arg\max_\theta \mathbb{E}_{z \sim p(z|x, \theta^{(n)})} p(x, z|\theta)
\end{aligned}$$

- Importance sampling is an improved variance reduction trick over rejection sampling, i.e., accept every sample by weighting.

- Importance sampling to calculate the normalizer.

$$\begin{aligned}
Z &= \int p(y|\theta) p(\theta) \mathrm{d}\ \theta \\
&= \int \frac{p(y|\theta) p(\theta)}{q(\theta)} q(\theta) \mathrm{d}\ \theta \\
&= \int w(\theta) q(\theta) \mathrm{d}\ \theta \\
&\approx \frac{1}{N} \sum_{i=1}^N w(\theta^{(i)})
\end{aligned}$$

- Represent the histogram in math – approximating the probability by samples,

$$\begin{aligned} p(\mathrm{d}\ \theta|data) &= \frac{1}{N}\sum_{i=1}^{N} w(\theta^{(i)})\delta_{\theta^{(i)}}(\mathrm{d}\ \theta) \\ &= p(\theta|data)\mathrm{d}\ \theta, \end{aligned}$$

  which is the measure. The probability is proportional to the samples falling into the region.

- Unnormalized importance sampling,

$$\begin{aligned} p(y_{t+1}|x_{t+1}, D) &= \frac{\int p(y_{t+1}|x_{t+1},\theta)p(D|\theta)p(\theta)\mathrm{d}\ \theta}{\int p(D|\theta)p(\theta)\mathrm{d}\ \theta} \\ &= \frac{\int p(y_{t+1}|x_{t+1},\theta)p(D|\theta)p(\theta)\frac{q(\theta)}{q(\theta)}\mathrm{d}\ \theta}{\int p(D|\theta)p(\theta)\frac{q(\theta)}{q(\theta)}\mathrm{d}\ \theta} \\ &= \frac{\int p(y_{t+1}|x_{t+1},\theta)w(\theta)q(\theta)\mathrm{d}\ \theta}{\int w(\theta)q(\theta)\mathrm{d}\ \theta} \\ &\approx \frac{\sum_{i=1}^{N} p(y_{t+1}|x_{t+1},\theta^{(i)})w(\theta^{(i)})}{\sum_{j=1}^{N} w(\theta^{(j)})} \end{aligned}$$

- Importance sampling suffers the curse of the dimension. It is hard to capture the typical set of the distribution. In other words, in 2D, we need N^2 points and in 3D we need N^3 points.

- As long as the graph (state space) is aperiodic and irreducible, we have a stationary distribution for the stochastic matrix corresponding to the eigenvalue of 1.

  Irreducibility: The graph is connected and every cluster can be accessed.

  Aperiodicity: There is no cycles in the graph, or there is oscillation occurring.

- In order to change a periodic graph into an aperiodic graph, we simply add transition connection to other nodes.

- Stein's method suggests: can we characterize the distribution by finding functional relationships between moments?

- Stein method is invariant to the normalizer, which can be used to find the mixing time of MCMC.

Variational Inference

- What does variational mean?

  –In general, it refers to the idea of expressing a quantity of interest θ^* (e.g. a posterior probability) as the solution of an optimization problem

$$\theta^* = \inf_{\theta \in M} f(\theta).$$

- Approximating the solution can now be accomplished by 1) simplifying the domain and 2) simplifying the function.
- Such approximations are particularly powerful in presence of convex structures.
- The maximum entropy subject to the empirical moments matching

$$p^* = \arg\max_p H(p), \;\; s.t. \mathbb{E}_p\{\phi(X)\} = \hat{\mu}$$

gives the solution in the form

$$p(x) \propto \exp\left\{\sum_k \lambda_k \phi_k(x)\right\},$$

which lies in the realm of exponential family, i.e.,

$$p_\theta(x) = \exp\{\langle\theta, \phi(x)\rangle - A(\theta)\}.$$

- The log-partition function $A(\theta)$ is convex in its domain because its second-order derivative is positive semi-definite.
- Conjugate duality of convex function gives

$$A^*(\mu) = \sup_\theta \langle\theta, \mu\rangle - A(\theta)$$
$$A(\theta) = \sup_\mu \langle\theta, \mu\rangle - A^*(\mu)$$

- $A^*(\mu)$ is the negative entropy of $p_{\theta(\mu)}$ where $p_{\theta(\mu)}$ is the exponential family distribution such that

$$\mathbb{E}_{\theta(\mu)}\phi(X) = \mu.$$

It comes from $\bigtriangledown_\theta \langle\theta, \mu\rangle - A(\theta) = 0$.

- Variational inference and duality:
  -

$$p_\theta(x, z) = \exp\{\langle\theta, \phi(x, z)\rangle - A(\theta)\}$$
$$p_\theta(z|x) = \exp\{\langle\theta, \phi(x, z)\rangle - A_x(\theta)\}$$

    where $A_x(\theta) = \log \int \exp\{\langle\theta, \phi(x, z)\rangle\}\mathrm{d}\ z$
  - $\mathcal{L}(\theta, x) = \log \int_z \exp\{\langle\theta, \phi(x, z)\rangle - A(\theta)\}\mathrm{d}\ z = A_x(\theta) - A(\theta)$.
    Because $A_x(\theta) = \sup_{\mu_x} \langle\theta, \mu_x\rangle - A_x^*(\mu_x)$, we have

$$\mathcal{L}(\theta, x) \geq \langle\theta, \mu_x\rangle - A_x^*(\mu_x) - A(\theta) = \tilde{\mathcal{L}}(\theta, x).$$

  - The E step gives the maximizer of $\tilde{\mathcal{L}}(\theta, x)$, $\mu_x^{(t+1)} = \mathbb{E}_{\theta^{(t)}}\phi(x, z)$. After the E step, the inequality becomes an equality, thus M step increases log-likelihood.

- Mean-field approximation: we model hidden variables as being independent.

- The term of $\bigtriangledown_\beta \mathbb{E}_{q(z|\beta)} f(Z)$ may be problematic, so we transform

$$\begin{aligned}
\bigtriangledown_\beta \mathbb{E}_{q(z|\beta)} f(Z) &= \int f(z) \bigtriangledown_\beta q(z|\beta) \mathrm{d}\ z \\
&= \int f(z) q(z|\beta) \bigtriangledown_\beta \log q(z|\beta) \mathrm{d}\ z \\
&= \mathbb{E}_{q(z|\beta)} f(z) \bigtriangledown_\beta \log q(z|\beta)
\end{aligned}$$