Approximate Inference

PAI Tutorial

Johannes Kirschner

October 25, 2019

Statistical Inference

What is statistical inference?

Statistical inference is the process of using data analysis to deduce properties of an underlying probability distribution. (Wikipedia)

- Use a model to relate data to a quantity of interest.

Examples:

- Maximum likelihood $x_{\text{ML}} = \arg\max_x P(Y = y|X = x)$
- **Bayesian inference** $P(X|Y) \propto P(Y|X)P(Y)$.
- Computing conditional/marginals in a Bayes Net

Bayesian Inference

The **Bayesian model** specifies:

- $P(Y|X)$, likelihood of Y given hypothesis/latent factors X.
- $P(X)$ prior probability of X ("belief").

→ Joint distribution $P(X, Y) = P(X)P(Y|X)$.

Bayesian inference asks to compute, e.g.

- *posterior distribution* $P(X|Y = y) = \frac{P(Y=y|X)P(X)}{P(Y=y)}$
- *maximum a posteriori* estimate $x_{\text{MAP}} = \arg\max_x P(X = x|Y = y)$
- *posterior sample* $x \sim P(X|Y = y)$

But: Evidence $P(Y = y) = \sum_x P(y, x)$ is often difficult to compute.

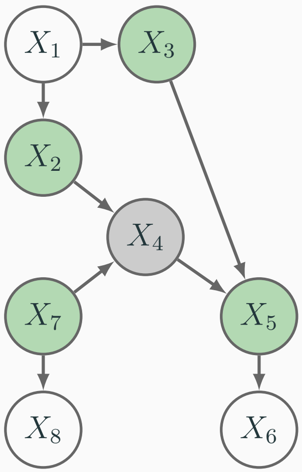
Bayesian Networks

Bayes Net:

- Defines a joint distribution $P(X_1, \ldots, X_m)$
- Inference: e.g. $P(X_1 | X_4 = T)$

Queries are difficult to compute in general:

- Need to marginalize.



Approximate Inference

Exact inference can be difficult:

- Bayesian inference requires to compute the marginal $P(Y)$.
- Computing conditional/marginals in Bayes Nets not always efficient.
- Factor distribution $P(X) = \frac{1}{Z} \prod_{i=1}^{m} \Phi_i(X_{A_i})$.

Goal: Approximate $P(X) = \frac{1}{Z}\Phi(X)$, where $\Phi(X) \geq 0$ and $Z = \sum_x \Phi(x)$.

Two popular approaches: Variational Inference and MCMC

Goal: Approximate $P(X) = \frac{1}{Z}\Phi(X)$, where $\Phi(X) \geq 0$, $Z = \sum_x \Phi(x)$.

Variational Inference

$$Q = \arg\min_{Q \in \mathcal{Q}} \mathrm{KL}(Q \| P)$$

- Choose a class of distributions $\mathcal{Q}$
- Solve optimization problem.
- Jordan et al. (1999)

Markov Chain Monte Carlo

- *Idea:* Set up a Markov chain
- Define transition kernel $T(X, X')$
- With stationary distribution $P(X)$
- Metropolis et al. (1953)

$\rightarrow$ *Sample P(X) from the Markov chain.*

Variational Inference:

$$Q = \arg\min_{Q \in \mathcal{Q}} \mathrm{KL}(Q \| P)$$

Variational Inference: Evidence Lower Bound

Goal: Approximate $P(X) = \frac{1}{Z}\Phi(X)$, where $\Phi(X) \geq 0$, $Z = \sum_x \Phi(x)$.

$$\begin{aligned}
\mathrm{KL}(Q\|P) &= \sum_x Q(x) \log \frac{Q(x)}{P(x)} \\
&= \sum_x Q(x) \log Q(x) - \sum_x Q(x) \log \Phi(x) + \log(Z)
\end{aligned}$$

$$\arg\min_{Q \in \mathcal{Q}} KL(Q\|P) = \arg\max_{Q \in \mathcal{Q}} \underbrace{\sum_x Q(x) \log \Phi(x) - \sum_x Q(x) \log Q(x)}_{:= \mathsf{ELBO}(Q)}$$

Evidence Lower Bound for Bayesian Inference

Goal: Approximate $P(X|Y=y) \propto \Phi(X) = P(Y=y|X)P(X)$

$$
\begin{aligned}
\text{ELBO}(Q) &= \sum_x Q(x)\log(P(Y|x)P(x)) - \sum_x Q(x)\log Q(x) \\
&= \sum_x Q(x)\log(P(Y|x)) - \text{KL}(Q(X)||P(X))
\end{aligned}
$$

$$
\begin{aligned}
&= \sum_x Q(x)\log\left(\frac{P(Y,x)}{P(x)}\right) - \sum_x Q(x)\log\frac{Q(x)}{P(x)}. \\
&= \sum_x Q(x)\log(P(x|Y)P(Y)) - \sum_x Q(x)\log Q(x) \\
&= \sum_x Q(x)\log P(Y) - \sum_x Q(x)\log\frac{Q(x)}{P(x|Y)} \\
&= \log P(Y) - KL(Q(X)\,||\,P(X|Y)).
\end{aligned}
$$

→ *Trade-off between likelihood and prior.*

$$
\begin{aligned}
\text{ELBO}(Q) &= \log P(Y) - \text{KL}(Q(X)||P(X|Y)) \\
&\leq \log P(Y)
\end{aligned}
$$

→ *Maximize lower-bound on evidence.*

Mean Field Variational Inference

Need to choose class of distributions $\mathcal{Q}$.

Mean field approximation: $Q(x) = \prod_{i=1}^{m} Q_i(x_i)$.

$$\mathrm{ELBO}(Q) = \sum_x \prod_{i=1}^{m} Q_i(x_i) \log \Phi(x) - \sum_{i=1}^{m} \sum_{x_i} Q_i(x_i) \log Q_i(x_i)$$

Objective for Q_j:

$$\mathrm{ELBO}_j(Q) = \sum_{x_j} Q_j(x_j) \underbrace{\sum_{x_{j-}} \prod_{i \neq j} Q_i(x_i) \log \Phi(x)}_{\log \exp(\sum \prod \cdots)} - \sum_{x_j} Q_j(x_j) \log Q_j(x_j) + const$$

Exact update for Q_j:

$$Q_j^* = \arg\max_{Q_j} \mathrm{ELBO}_j(Q) \propto \exp\left(\sum_{x_{j-}} \prod_{i \neq j} Q_i(x_i) \log \Phi(x)\right)$$

Mean Field Coordinate Ascent

CAVI (Coordinate Ascent Variational Inference):

Choose ordering of $X_1, \ldots, X_m$
Initialize $Q_1, \ldots, Q_m$
repeat
  for $j = 1{:}m$ **do**
    $Q_j \leftarrow \arg\max_{Q_j} \mathsf{ELBO}_j(Q) \propto \exp\left(\sum_{x_{j-}} \prod_{i \neq j} Q_i(x_i) \log \Phi(x)\right)$
until converged

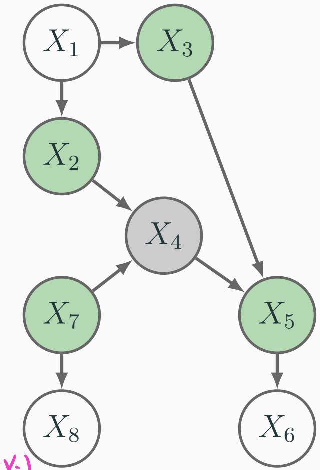
$$\mathsf{ELBO}_j(Q) = \sum_{x_j} Q_j(x_j) \sum_{x_{j-}} \prod_{i \neq j} Q_i(x_i) \log \Phi(x) - \sum_{x} Q_j(x_j) \log Q_j(x_j)$$

Mean Field Variational Inference: Bayes Nets

Markov Blanket:

$\text{mb}(X_j) = \text{parents}(X_j) \cup \text{children}(X_j) \cup \text{co-parents}(X_j)$

- ▷ $\text{co-parents}(X_j) = \text{parents}(\text{children}(X_j)) \setminus X_j$
- ▷ $\forall i \neq j \, , X_j \perp X_i | X_{\text{mb}(X_j)}$



CAVI update only depends on Markov Blanket:

$$Q_j \propto \exp\left(\sum_{x_{j-}} \prod_{i \neq j} Q_i(x_i) \log \Phi(x)\right)$$

$$\propto \exp\left(\sum_{x_{\text{mb}(X_j)}} \prod_{i \in \text{mb}(X_j)} Q_i(x_i) \log P(X_j | X_{\text{mb}(X_j)})\right)$$

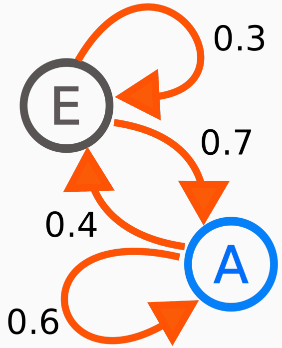
只需更新 M.B. 里面的元素. mb(Xj)

Markov Chain Monte Carlo

Markov Chains

Markov Chain:

- Sequence of random variables $X^{(1)}, X^{(2)}, X^{(3)}, \ldots$ on $\mathcal{S}$
- States space $\mathcal{S} = \{S_1, \ldots, S_k\}$
- Initial distribution: $P(X^{(1)} = S)$
- Transition kernel: $T(S, S') = P(X^{(t)} = S' | X^{(t-1)} = S)$
- Can be written as $k \times k$ matrix T.



$T = \begin{pmatrix} 0.3 & 0.7 \\ 0.4 & 0.6 \end{pmatrix}$

Stationary distribution: $\pi(S) = \lim_{t\to\infty} P(X^{(t)} = S)$

- Satisfies $\pi(S)T(S, S') = \pi(S)$
- Matrix notation: $\pi^\top T = \pi^\top$ (left eigenvector)
- Independent of initial distribution (mild conditions)

MCMC: General Idea

Goal: Approximate $P(X) = \frac{1}{Z}\Phi(X)$, where $\Phi(X) \geq 0$, $Z = \sum_x \Phi(x)$.

Markov Chain Monte Carlo (MCMC)

Idea: Construct Markov chain with stationary distribution $P(X)$.

- Simulating the Markov chain to draw samples $\sim P(X)$

Detailed balance condition:

$$\pi(S)T(S, S') = \pi(S')T(S', S)$$

- Implies that $\pi(S)$ is the stationary distribution.
- Define $T(X, X')$ that satisfies detailed balance for Φ:

$$\Phi(X)T(X, X') = \Phi(X')T(X', X)$$

MCMC: Gibbs Sampling

Gibbs Sampling:

Initialize $x_1, \ldots, x_m$

repeat

  for $j = 1{:}m$ **do**

    $x_j \leftarrow X_j \sim \frac{1}{Z_j}\Phi(X_1 = x_1, \ldots, X_j, \ldots, X_m = x_m)$

until converged

- ▷ Draw samples conditioned on all other variables.
- ▷ We'll show detailed balance for $j \sim \text{Uniform}(\{1, \ldots, m\})$

Gibbs Sampling: Detailed Balance

Gibbs sampling generates Markov chain $X^{(1)}, X^{(2)}, X^{(3)}, \ldots$

- $X^{(t)} = (X_1^{(t)}, \ldots, X_m^{(t)})$

$$T(X, X') = \begin{cases} 0 & \text{if } X_i \neq X_i' \text{ and } X_j \neq X_j' \text{ for } i \neq j \\ \propto P(X_1, \ldots, X_m) & X = X' \\ \propto \frac{1}{m} P(X_{1:j-1}, X_j', X_{j+1:m}) & \exists j,\ X_j \neq X_j' \text{ and } X_i = X_i' \,\forall\, i \neq j \end{cases}$$

Detailed Balance:

- Case 1) & 2) are easy
- Case 3) $j,\ X_j \neq X_j'$ and $X_i = X_i' \,\forall\, i \neq j$

$$\underline{P(X_{1:m})} \cdot \frac{1}{m} \underline{P(X_{1:j-1}, X_j', X_{j+1:m})} = \underline{P(X_{1:m}')} \cdot \frac{1}{m} \underline{P(X_{1:j-1}', X_j, X_{j+1:m}')}$$

Gibbs Sampling on Bayes Nets

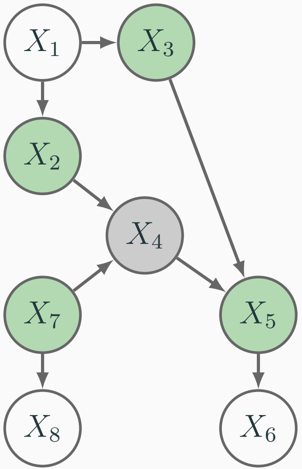
Markov Blanket:

$\text{mb}(X_j) = \text{parents}(X_j) \cup \text{children}(X_j) \cup \text{co-parents}(X_j)$

- co-parents(X_j) = parents(children(X_j)) $\setminus X_j$
- $\forall i \neq j \;, X_j \perp X_i | X_{\text{mb}(X_j)}$

Gibbs sample only depends on Markov Blanket:

$$X_j \sim P(X_{1:j-1} = x_{1:j-1}, X_j, X_{j+1:m} = x_{j+1:m})$$
$$\sim P(X_j, X_{\text{mb}(X_j)} = x_{\text{mb}(X_j)})$$



MCMC and Variational Inference

MCMC: 渐近.

- ▷ Asymptotically correct
- ▷ Computationally expensive, if it takes long to converge
- ▷ Difficult to detect convergence (burn-in)
- ▷ Does not generate iid samples

Variational Inference:

- ▷ Faster
- ▷ Easy to monitor optimization progress
- ▷ Class of distributions $\mathcal{Q}$ often misspecifed, i.e. $P \notin \mathcal{Q}$.
- ▷ Optimization can be difficult

Further Reading

- Bayesian Inference, MCMC and Variational Inference:

https://towardsdatascience.com/25a8aa9bce29

- Technical Review on Variational Inference:

https://arxiv.org/abs/1601.00670

- Forward & Reverse KL:

https://blog.evjang.com/2016/08/variational-bayes.html