Brief Intro to Variational Inference

WHY, WHAT and HOW

Only basic statistics background needed

(Nov 02, 2016)

YANG Jiancheng

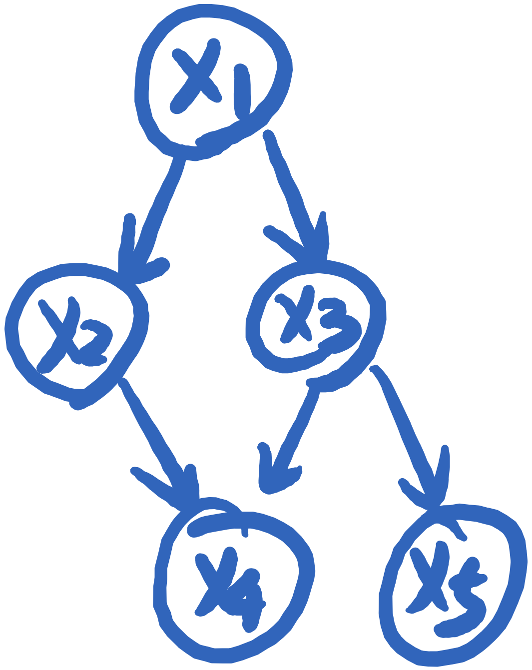
Outline

- I. WHY
  - Probability Review
  - Graphic Models
  - The Point of Variational Inference
- II. WHAT
  - Ways to infer
  - Short Words
- III. HOW
  - Information and Entropy
  - KL divergence
  - Derivation
  - Example

• I. WHY

- **Probability Review**

pdf / pmf = probability density / mass function



$\boxed{\text{Bayes Rules}}$

joint pdf

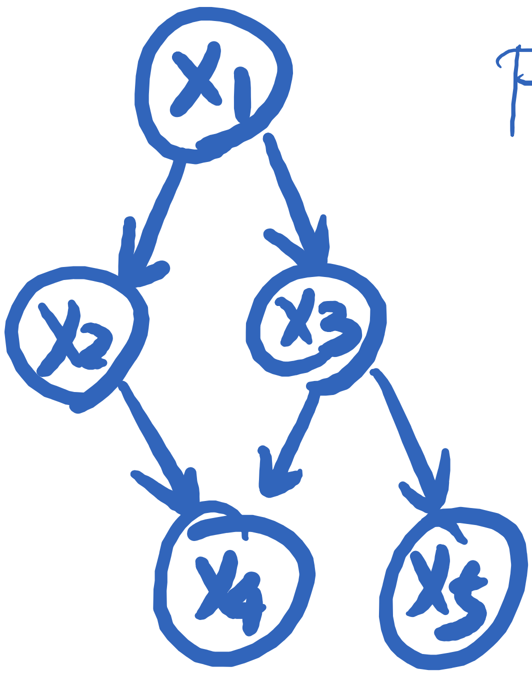
$$\boxed{p(X|Y)} = \frac{P(X,Y)}{\boxed{P(Y)}} \quad \frac{\boxed{P(X,Y)}}{\int_x P(x,y)\,dx}$$

Conditional pdf

prior

• I. WHY

- **Graphic Models**



EASY

Joint pdf

$$P(x_1,x_2,x_3,x_4,x_5)=P(x_5|x_3)\cdot P(x_4|x_2,x_3)\cdot P(x_3|x_1)\cdot P(x_2|x_1)\cdot P(x_1)$$

Conditional pdf

$$P(x_3,x_4|x_1,x_2,x_5)=\frac{P(x_1,x_2,x_3,x_4,x_5)}{P(x_1,x_2,x_5)}$$

unknown ←

$$=\frac{P(x_1,x_2,x_3,x_4,x_5)}{\iint_{x_3\,x_4} P(x_1,x_2,x_3,x_4,x_5)\,dx_3\,dx_4}$$

hard ←

HARD

•I. WHY

- **The Point of Variation Inference**

JOINT $\xrightarrow{?}$ CONDITIONAL

easy hard.

Answer is YES.

• II. WHAT

- Ways to infer

a) **MCMC**
- *Gibbs Sampling*
- *Metropolis Hastings*

Exact Easy
Sampling Slow

b) **Variational Inference**

Good Approximate Hard
Deterministic Fast

c) **Laplace Inference**

Poor Approximate Easy
Deterministic Fast

• II. WHAT

- **Short Words**

We try to use a simple distribution

$$q(z)$$

to approximate a complex conditional

$$p(z|x)$$

When we don't/hardly know $P(x)$

• III. HOW

- **Information and Entropy**

Information $I = -\log P(x)$, x is event.

$\boxed{P\downarrow,\ I\uparrow}$

Entropy (Average Information) → The measure of uncertainty.

discrete: $H = \sum_i P_i \log P_i$

continuous: $H = -\int_x P(x) \log P(x)\, dx$

• III. HOW

- **KL divergence**

Kullback-Leibler divergence

$$KL(\underline{p \| q}) = -\sum_i p_i \log \frac{q_i}{p_i}$$

p, q is distribution

$$= -\sum_i p_i \log q_i - \left(-\sum_i p_i \log p_i\right)$$

$$\underline{H(p,q)} - H(p)$$

$\downarrow$

Cross-entropy

• III. HOW

- **KL divergence**

but not symmetrical!

Measure the "distance" of two distributions

Information inequality

$$\boxed{\begin{array}{c} KL(p,q) = H(p,q) - H(p) \geq 0 \\ KL = 0 \text{ iff } p = q \end{array}}$$

use Jensen's inequality to prove

Cross-entropy as loss function

$$H(p,q) = KL(p,q) - H(p)$$

$H(p,q) \downarrow$

$KL \downarrow$

p, q similarity $\uparrow$

• III. HOW

- **KL divergence**

Maximum Entropy Principle

(the principle of insufficent reasoning)

The discrete distribution with max entropy is uniform distribution $u(x)$

$$0 \le KL(p\|u) = \sum_i p_i \log \frac{p_i}{u_i} = \sum_i p_i \log p_i - \sum_i p_i \log u_i$$

$$\Rightarrow \quad H(p) \le H(u) = \log |X|$$

$\hookrightarrow |X|$ = the number of state

$H(p)$ get max iff $p = u$.

• III. HOW

• **Derivation**

We want to use $q(z)$ to estimate $P(z|x)$

known $P(z,x)$

$$\boxed{\min_{q} KL(q(z)||P(z|x))}$$

• III. HOW

- **Derivation**

$$KL = -\sum_z q(z)\log\frac{P(z|x)}{q(z)}$$

$$\frac{P(x,z)}{P(z)} = P(z|x)$$

$$\Rightarrow KL = -\sum q(z)\log\frac{P(x,z)}{q(z)} + \underbrace{\sum q(z)}_{\text{sum to } 1}\cdot\log P(x)$$

given x, fixed.

$$\therefore \underbrace{KL}_{\geq 0} + \boxed{\sum q(z)\cdot\log\frac{P(x,z)}{q(z)}} = \log P(x) \leq 0$$

Lower Bound $(\mathcal{L}) \leq 0$

• III. HOW

- **Derivation**

$$KL + \mathcal{L} = \log P(x) \quad (\text{fixed})$$

known

$$\therefore \min_q KL \Longleftrightarrow \max_q \mathcal{L} = \sum_z q(z) \log \frac{\boxed{P(x,z)}}{q(z)}$$

$P(x|z)$ disappear!

Though, estimate $q(z)$ is still not easy.

- # III. HOW
  - **Derivation**

q is a "constructed" distribution, so we ASSUME:

$\boxed{z = (z_1, z_2, z_3),\ z_1, z_2, z_3 \text{ is independent.}}$

thus,

$$q(z) = q(z_1, z_2, z_3) = q(z_1) \cdot q(z_2) \cdot q(z_3)$$

and,

$$\mathcal{L} = \sum_{z_1} \sum_{z_2} \sum_{z_3} q(z_1) \cdot q(z_2) \cdot q(z_3) \left(\log P(x, z) - \log q(z_1) - \log q(z_2) - \log q(z_3) \right)$$

• III. HOW

- **Derivation**

Still difficult, how about $\boxed{\text{imagining}}$:

We know z_2, z_3, want to solve z_1.

$$\sum_{z_1}\sum_{z_2}\sum_{z_3} q(z_1)\, q(z_2)\, q(z_3) \cdot \begin{cases} ① \ \log P(X,Z) \\ ② \ -\log q(z_1) \\ ③ \ -\log q(z_2) - \log q(z_3) \end{cases}$$

$$① = \sum_{z_1} q(z_1) \cdot \boxed{\sum_{z_2}\sum_{z_3} q(z_2)\cdot q(z_3) \log P(X,Z)} = \sum_{z_1} q(z_1)\, E_{z_2 z_3} \log P(X,Z)$$

$$② = -\sum_{z_1} q(z_1) \cdot \log q(z_1) \boxed{\sum_{z_2}\sum_{z_3} q(z_2)\, q(z_3)} = 1$$

$$③ = -\sum_{z_1} q(z_1) \cdot \sum_{z_2}\sum_{z_3} q(z_2)\, q(z_3) \left(\log q(z_2) + \log (z_3)\right) = -\sum_{z_1} q(z_1) \cdot K = K$$

• III. HOW

- **Derivation**

$$\mathcal{L} = \sum_{z_1} q(z_1)\left[\mathop{E}_{z_2 z_3} \log P(x,z) - K\right] - \sum_{z_1} q(z_1) \cdot \log q(z_1)$$

Let $f(x,z) = C_1 \, e^{\mathop{E}_{z_2 z_3} \log P(x,z)}$ is a distribution.

$\therefore$

$$\mathcal{L} = \sum_{z_1} q(z_1) \log f(x,z) - \sum_{z_1} q(z_1) \log q(z_1) + \text{const}.$$

$$= \sum_{z_1} q(z_1) \cdot \log \frac{f(x,z)}{q(z_1)} + \text{const}$$

$$= -KL\left(q(z_1) \,\|\, f(x,z)\right) + \text{Const}.$$

• III. HOW

- **Derivation**

Got $\mathcal{L} + KL(q(z_1) \| f(x, z)) = \text{const}$.

$\max \mathcal{L} \Longleftrightarrow \min KL$

$$\therefore \boxed{q(z_1) = f(x, z) = C_1 e^{\underset{z_2 z_3}{E} \log P(x, z)}}$$

So as $q(z_2)$ and $q(z_3)$

Then:

$$q(z) = q(z_1) \cdot q(z_2) \cdot q(z_3)$$

• III. HOW

- **Derivation**

In many cases, given q is distribution, we can infer q directly (use $\int q_\theta \, dz = 1$)

But, if we cannot, just $\boxed{\text{iterate}}$

$$\begin{tikzcd}
 & P(z_2) \arrow[d] \\
P(z_1) \arrow[ur] & P(z_3) \arrow[l]
\end{tikzcd}$$

- **III. HOW**
  - **Example**

Given

$$p(x, y, z) = \lambda_1 \lambda_2 \lambda_3 e^{-\lambda_1 x - \lambda_2 y - \lambda_3 z},$$

calculate $P(x, y | z)$.

Simple Bayes Rule Solution:

$$P(x, y | z) = \frac{P(x, y, z)}{P(z)} = \frac{P(x, y, z)}{\int_{x=0}^{+\infty} \int_{y=0}^{+\infty} p(x, y, z)\, dx\, dy} = \lambda_1 \lambda_2 e^{-\lambda_1 x_1 - \lambda_2 y}$$

• III. HOW

- Example

Variational Inference Solution:

Use $q(x,y) = q(x)\cdot q(y)$ to estimate $P(x,y|z)$

Use the equation:

$$\ln q(x) = \underset{y}{E} \ln p(x,y,z) + k$$

$$= \underset{y}{E}\left(\ln \lambda_1\lambda_2\lambda_3 - \lambda_1 x - \lambda_2 y - \lambda_3 z + k\right)$$

$$= -\lambda_1 x - \underline{\lambda_2 E y} - \underline{\lambda_3 z} + \underline{\ln\lambda_1\lambda_2\lambda_3} + \underline{k}$$

Const

• III. HOW

- **Example**

$$\because \ln q(x) = -\lambda_1 x + K_1$$

$$\Rightarrow q(x) = C_1 e^{-\lambda_1 x}$$

$$\int q(x) = 1 \Rightarrow C_1 = \lambda_1$$

So as $q(y)$.

$$\therefore q(x,y) = \lambda_1 \lambda_2 e^{-\lambda_1 x - \lambda_2 y}$$

The same as Bayes Rule Solution!

A "perfect" approximate!

Bibliography

- Variational Inference Tutorial Series by Chieu from NEU
- Machine Learning: A Probabilistic Perspective (Kevin P. Murphy) Chapter 2

Thanks for listening!