Inferencia Aproximada

→ Modelos con múltiples capas generan posteriores intratables.

→ El concepto de inferencia exacta se puede describir como un problema de optimización.

→ Se pretende estimar $\log(p(v;\theta))$

v: datos observados

h: variables latentes

θ: parámetros

→ Se busca calcular una cota (Lower Bound) $\mathcal{L}(v,\theta,q)$ de $\log(p(v;\theta))$; denominada Evidence Lower Bound (ELBO)

→ ELBO también se conoce como la negative variational free energy.

$$\mathcal{L}(v,\theta,q) = \log p(v;\theta) - D_{KL}(q(h|v) \,||\, p(h|v;\theta))$$

Supanga que se quiere marginalizar:

$$p(x) = \int_z p(x,z)\, \partial z$$

$$\log p(x) = \log \int_z p(x,z)\, \partial z$$

Si se arega $q(z)$:

$$\log p(x) = \log \int_z p(x,z) \frac{q(z)}{q(z)} \partial z = \log \mathcal{E}_q \left\{ \frac{p(x,z)}{q(z)} \right\}$$

Por la desigualdad de Jensen:

$$\log \mathcal{E}_q \left\{ \frac{p(x,z)}{q(z)} \right\} \geq \mathcal{E}_q \left\{ \log \frac{p(x,z)}{q(z)} \right\}$$

$$\log p(x) = \log \mathcal{E}_q \left\{ \frac{p(x,z)}{q(z)} \right\} \geq \mathcal{E}_q \left\{ \log \frac{p(x,z)}{q(z)} \right\}$$

$$\log p(x) \geq \mathcal{E}_q \left\{ \log \frac{p(x,z)}{q(z)} \right\} = \mathcal{L}$$

$\mathcal{L}$ es cota sobre la evidencia $p(x)$.

Si se analiza desde el posterior. $p(z|x)$ aprox. por $q(z)$.

$$D_{KL}(q(z) \| p(z|x)) = -\mathcal{E}_q \left\{ \log \frac{p(z|x)}{q(z)} \right\}$$

$$= -\mathcal{E}_q \left\{ \log p(z|x) - \log q(z) \right\}$$

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

$$D_{KL}(q(z) \| p(z|x)) = -\mathbb{E}_q\left\{\log \frac{p(z,x)}{p(x)} - \log q(z)\right\} =$$

$$-\mathbb{E}_q\{\log p(z,x) - \log q(z) - \log p(x)\} =$$

$$\underbrace{-\mathbb{E}_q\{\log(p(z,x)/q(z))\}}_{-\mathcal{L}} + \mathbb{E}_q\{\log p(x)\}$$

$$D_{KL}(q(z) \| p(z|x)) = -\mathcal{L} + \log p(x)$$

$$\mathcal{L} = \log p(x) - D_{KL}(q(z) \| p(z|x))$$

→ Para $q(z)$ apropiado, $\mathcal{L}$ es tratable.

→ Se busca $q(z)$ que maximice $\mathcal{L}$.

→ Incluso, en forma canónica, y respecto a las variables h, v y θ:

$$\mathcal{L}(v,\theta,q) = \mathbb{E}_{h\sim q}\{\log p(h,v)\} + H(q)$$

$$q(z) = q(h|v)$$

$\mathcal{L}$ es tratable.

Expectation maximization

→ Algoritmo clásico basado en ELBO

E-step: $\theta^{(0)}$: ptos. iniciales.

$$q(h^{(i)} \mid v) = p(h^{(i)} \mid v^{(i)}; \theta^{(0)})$$

M-step:

$$\max_{\theta} \sum_i \mathcal{L}(v^{(i)}, \theta, q)$$

Para PL M-step es difícil.

Maximum a Posteriori - MAP.

→ Generalmente estamos interesados en calcular las variables latentes $p(h \mid v)$.

→ Como alternativa al cálculo de la distribución, se busca calcular los valores más probables:

$$h^* = \arg\max_h p(h \mid v)$$

→ MAP indirectamente provee un valor de q en inferencia aproximada desde $\mathcal{L}$.

Variational Inference.

→ El proceso de inferencia busca q maximizando $\mathcal{L}$.

→ El proceso de aprendizaje (learning), busca θ maximizando $\mathcal{L}$

→ Variational Learning: max. $\mathcal{L}$ sobre una familia restringida de distribuciones q.

→ Se busca facilitar el cálculo de
$\mathbb{E}_q\{\log p(h,v)\}$

→ En general, se restringe la factorización de q:

$$q(h|v) = \prod_i q(h_i|v)$$

(mean Field).

→ Se impone un modelo gráfico estructurado sobre q para fijar las interacciones que se quieren capturar con la aprox.

→ Para variables discretas se optimiza un # finito de variables discretas

→ Para variables continuas se debe considerar calculus of variations.
↳ Resolver para una función $f(x)$ no para $\theta \in \mathbb{R}^d$