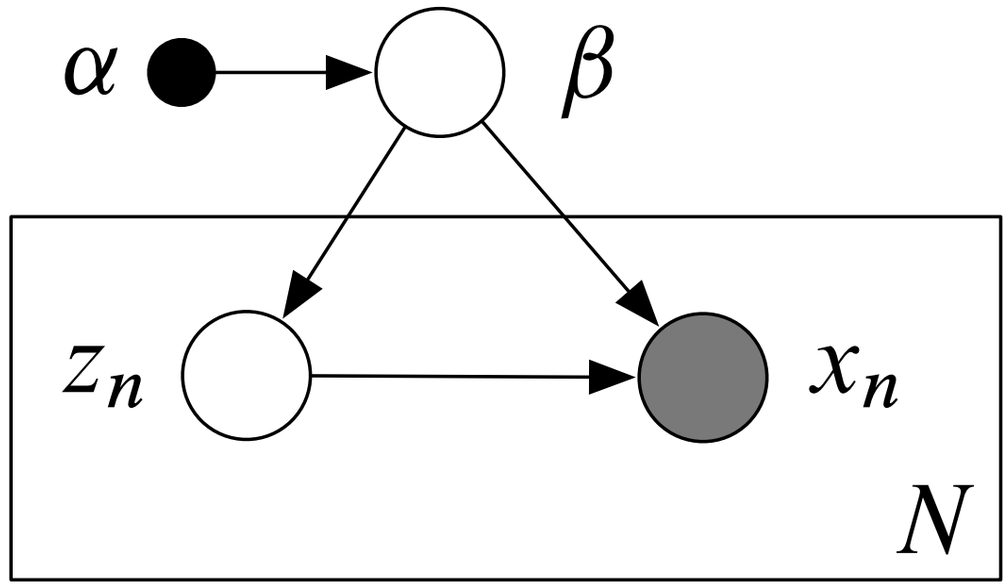
Стохастический вариационный вывод

Сергей Бартунов, ВЦ РАН

Рассматриваемый класс моделей:



$$p(\mathbf{x}, \mathbf{z}, \beta|\alpha) = p(\beta|\alpha) \prod_{i=1}^{N} p(z_i|\beta) p(x_i|z_i, \beta)$$

Вариационный вывод

Апостериорное распределение

$$p(\mathbf{z}, \beta|\mathbf{x}, \alpha) = \frac{p(\mathbf{x}, \mathbf{z}, \beta|\alpha)}{\int p(\mathbf{x}, \mathbf{z}, \beta|\alpha) d\mathbf{z} d\beta}$$

Будем искать наилучшее приближение

$$q^*(\mathbf{z}, \beta) = \arg\min_q D_{KL}(q||p(\cdot|\mathbf{x}, \alpha)) \approx p(\mathbf{z}, \beta|\mathbf{x}, \alpha)$$

Дивергенция Кульбака-Лейблера:

$$D_{KL}(q||p(\cdot|\mathbf{x}, \alpha)) = \mathbb{E}_q \left[\log q(\mathbf{z}, \beta) - \log p(\mathbf{z}, \beta|\mathbf{x}, \alpha)\right]$$

Вариационная нижняя оценка

Обоснованность модели:

$$\log p(\mathbf{x}|\alpha) = \log \int p(\mathbf{x}, \mathbf{z}, \beta|\alpha) d\mathbf{z} d\beta$$

Введем некоторое распределение q:

$$\log p(\mathbf{x}|\alpha) = \log \int \frac{p(\mathbf{x}, \mathbf{z}, \beta|\alpha) q(\mathbf{z}, \beta)}{q(\mathbf{z}, \beta)} d\mathbf{z} d\beta$$

Применим неравенство Йенсена:

$$\log p(\mathbf{x}|\alpha) \geq \int q(\mathbf{z}, \beta) \log \frac{p(\mathbf{x}, \mathbf{z}, \beta|\alpha)}{q(\mathbf{z}, \beta)} d\mathbf{z} d\beta$$

Вариационная нижняя оценка

$$\log p(\mathbf{x}|\alpha) \geq \mathcal{L}(q) = \mathbb{E}_q \left[\log p(\mathbf{x}, \mathbf{z}, \beta|\alpha) - \log q(\mathbf{z}, \beta)\right]$$

Можно показать, что

$$\log p(\mathbf{x}|\alpha) = \mathcal{L}(q) + D_{KL}(q||p(\cdot|\mathbf{x}, \alpha))$$

Mean-field variational inference

Будем искать приближение в семействе полностью факторизованных распределений

$$q(\mathbf{z}, \beta) = q(\beta) \prod_{i=1}^{N} q(z_i)$$

Формулы пересчета:

$$\log q(z_i) \propto \mathbb{E}_{q(\beta)} \left[\log p(x_i | z_i, \beta) + \log p(z_i | \beta) \right]$$

$$\log q(\beta) \propto \mathbb{E}_{q(\mathbf{z})} \left[\sum_{i=1}^{N} \log p(x_i, z_i | \beta) \right] + \log p(\beta | \alpha)$$

Снова надо считать нормированную константу!

$$q(\beta) \propto p(\beta|\alpha) \exp(\sum_{i=1}^{N} \mathbb{E}_{q(z_i)} \log p(x_i, z_i|\beta))$$

Все распределения лежат в экспоненциально семействе

$$p(\beta|\alpha) = h(\beta)\exp\{\alpha^T t(\beta) - a_g(\alpha)\}$$

$$p(x_i, z_i|\beta) = h(x_i, z_i)\exp\{\beta^T t(x_i, z_i) - a_l(\beta)\}$$

Все распределения лежат в экспоненциально семействе

$$p(\beta|\alpha) = h(\beta)\exp\{\alpha^T t(\beta) - a_g(\alpha)\}$$
$$p(x_i, z_i|\beta) = h(x_i, z_i)\exp\{\beta^T t(x_i, z_i) - a_l(\beta)\}$$

и образуют сопряженную пару

$$p(\beta|\mathbf{x}, \mathbf{z}, \alpha) = p(\beta|\eta_g(\mathbf{x}, \mathbf{z}, \alpha))$$
$$= h(\beta)\exp\{\eta_g(\mathbf{x}, \mathbf{z}, \alpha)^T t(\beta) - a_g(\eta_g(\mathbf{x}, \mathbf{z}, \alpha))\}$$

Все распределения лежат в экспоненциально семействе

$$p(\beta|\alpha) = h(\beta)\exp\{\alpha^T t(\beta) - a_g(\alpha)\}$$
$$p(x_i, z_i|\beta) = h(x_i, z_i)\exp\{\beta^T t(x_i, z_i) - a_l(\beta)\}$$

и образуют сопряженную пару

$$p(\beta|\mathbf{x}, \mathbf{z}, \alpha) = p(\beta|\eta_g(\mathbf{x}, \mathbf{z}, \alpha))$$
$$= h(\beta)\exp\{\eta_g(\mathbf{x}, \mathbf{z}, \alpha)^T t(\beta) - a_g(\eta_g(\mathbf{x}, \mathbf{z}, \alpha))\}$$

Отсюда следует:

$$\alpha = (\alpha_1, \alpha_2)$$
$$t(\beta) = (\beta, -a_l(\beta))$$
$$\eta_g(\mathbf{x}, \mathbf{z}, \alpha) = (\alpha_1 + \sum_{i=1}^{N} t(x_i, z_i), \alpha_2 + N)$$

Вариационные распределения будут иметь ту же экспоненциальную форму

$$q(\beta|\lambda) = h(\beta)\exp\{\lambda^T t(\beta) - a_g(\lambda)\}$$

$$q(z_i|\phi_i) = h(z_i)\exp\{\phi_i^T t(z_i) - a_l(\phi_i)\}$$

Вариационный пересчет

Вариационная нижняя оценка как функция от λ

$$\begin{aligned}\mathcal{L}(\lambda) &= \mathbb{E}_q\left[\log p(\beta|\mathbf{x},\mathbf{z},\alpha) - \log q(\beta|\lambda)\right] + \text{const} \\ &= \mathbb{E}_q\left[\eta_g(\mathbf{x},\mathbf{z},\alpha)^T t(\beta) - \lambda^T t(\beta) + a_g(\lambda)\right] + \text{const} \\ &= \mathbb{E}_q\left[\eta_g(\mathbf{x},\mathbf{z},\alpha)\right]^T \nabla_\lambda a_g(\lambda) - \lambda^T \nabla_\lambda a_g(\lambda) + a_g(\lambda) + \text{const}\end{aligned}$$

Вариационный пересчет

Вариационная нижняя оценка как функция от λ

$$\begin{aligned}\mathcal{L}(\lambda) &= \mathbb{E}_q\left[\log p(\beta|\mathbf{x},\mathbf{z},\alpha) - \log q(\beta|\lambda)\right] + \text{const} \\ &= \mathbb{E}_q\left[\eta_g(\mathbf{x},\mathbf{z},\alpha)^T t(\beta) - \lambda^T t(\beta) + a_g(\lambda)\right] + \text{const} \\ &= \mathbb{E}_q\left[\eta_g(\mathbf{x},\mathbf{z},\alpha)\right]^T \nabla_\lambda a_g(\lambda) - \lambda^T \nabla_\lambda a_g(\lambda) + a_g(\lambda) + \text{const}\end{aligned}$$

Градиент вариационной нижней оценки

$$\nabla_\lambda \mathcal{L}(\lambda) = \nabla^2_\lambda a_g(\lambda)(\mathbb{E}_{q(\mathbf{z})}\left[\eta_g(\mathbf{x},\mathbf{z},\alpha)\right] - \lambda)$$

Вариационный пересчет

Вариационная нижняя оценка как функция от λ

$$\begin{aligned}\mathcal{L}(\lambda) &= \mathbb{E}_q \left[\log p(\beta|\mathbf{x}, \mathbf{z}, \alpha) - \log q(\beta|\lambda)\right] + \text{const} \\ &= \mathbb{E}_q \left[\eta_g(\mathbf{x}, \mathbf{z}, \alpha)^T t(\beta) - \lambda^T t(\beta) + a_g(\lambda)\right] + \text{const} \\ &= \mathbb{E}_q \left[\eta_g(\mathbf{x}, \mathbf{z}, \alpha)\right]^T \nabla_\lambda a_g(\lambda) - \lambda^T \nabla_\lambda a_g(\lambda) + a_g(\lambda) + \text{const}\end{aligned}$$

Градиент вариационной нижней оценки

$$\nabla_\lambda \mathcal{L}(\lambda) = \nabla^2_\lambda a_g(\lambda)(\mathbb{E}_{q(\mathbf{z})} \left[\eta_g(\mathbf{x}, \mathbf{z}, \alpha)\right] - \lambda)$$

Формула пересчета

$$\lambda = \mathbb{E}_{q(\mathbf{z})} \eta_g(\mathbf{x}, \mathbf{z}, \alpha) \Leftrightarrow q(\beta) \propto p(\beta|\alpha) \exp(\sum_{i=1}^{N} \mathbb{E}_{q(z_i)} \log p(x_i, z_i|\beta))$$

Локальные вариационные распределения детерминировано зависят от глобального

$$\mathcal{L}(q(\beta)) = \mathcal{L}(q(\beta), q^*(\mathbf{z}))$$

$$q^*(\mathbf{z}) \propto \exp(\mathbb{E}_{q(\beta)} \log p(x_i, z_i|\beta))$$

Стохастическая оптимизация

Типичная задача машинного обучения

$$L(\theta) = \sum_{i=1}^{N} l(y_i, f(x_i, \theta)) + Cg(\theta) \to \min_{\theta}$$

Стохастический градиентный спуск

$$\theta^{t+1} = \theta - \rho_{t+1} \nabla L(\theta^t)$$

$$\nabla L(\theta) \approx N \nabla l(y_i, f(x_i, \theta)) + C \nabla g(\theta)$$

Стохастическая оптимизация

Типичная задача вариационного вывода

$$\mathcal{L}(\lambda) = \sum_{i=1}^{N} \underbrace{\mathbb{E}_{q^*(z_i)} \left[\log p(x_i, z_i|\beta) - \log q(z_i)\right]}_{l(x_i, q(\beta))} + \underbrace{\mathbb{E}_{q(\beta|\lambda)} \left[\log p(\beta|\alpha) - \log q(\beta|\lambda)\right]}_{g(q(\beta))}$$

Стохастический вариационный вывод (?)

$$\lambda^{t+1} = \lambda^t + \rho_t \nabla \mathcal{L}(\lambda^t)$$

$$\nabla \mathcal{L}(\lambda^t) \approx \nabla^2_\lambda a_g(\lambda^t)(\mathbb{E}_{q(z_i)} \eta^i_g(\mathbf{x}, \mathbf{z}, \alpha) - \lambda^t)$$

$$\eta^i_g(\mathbf{x}, \mathbf{z}, \alpha) = (\alpha_1 + N t(x_i, z_i), \alpha_2 + N)$$

Обычный градиент

$$\nabla f(\lambda) = \arg\max_{d\lambda} f(\lambda + d\lambda) \quad \text{s.t. } \|d\lambda\|_2 \to 0$$

Обычный градиент

$$\nabla f(\lambda) = \arg\max_{d\lambda} f(\lambda + d\lambda) \quad \text{s.t. } \|d\lambda\|_2 \to 0$$

$$\begin{aligned}
\arg\max_{\|d\lambda\|\leq\epsilon} f(\lambda + d\lambda) &\approx \arg\max_{\|d\lambda\|\leq\epsilon} f(\lambda) + \nabla f(\lambda)^T d\lambda \\
&= \arg\max_{\|d\lambda\|\leq\epsilon} \nabla f(\lambda)^T d\lambda \\
&= \frac{\nabla f(\lambda)}{\|\nabla f(\lambda)\|}\epsilon
\end{aligned}$$

Пространство параметров имеет значение

В эвклидовой метрике

$$\|d\lambda\|_2^2 = \sum_{j=1}^{d} (d\lambda_j)^2$$

В римановской метрике

$$\|d\lambda\|_G^2 = (d\lambda)^T G(\lambda)(d\lambda)$$

Разумно выбрать метрику такую, чтобы

$$\|d\lambda\|_G^2 = \underbrace{D_{KL}(q(\cdot|\lambda)\|q(\cdot|\lambda+d\lambda)) + D_{KL}(q(\cdot|\lambda+d\lambda)\|q(\cdot|\lambda))}_{D_{KL}^{sym}(\lambda,\lambda+d\lambda)}$$

Направление наискорейшего возрастания
в римановской метрике

$$\begin{aligned}
\arg \max_{\|d\lambda\|_G \leq \epsilon} f(\lambda + d\lambda) &\approx \arg \max_{\|d\lambda\|_G \leq \epsilon} f(\lambda) + (\nabla f(\lambda))^T d\lambda \\
&= \arg \max_{\|d\lambda\|_G \leq \epsilon} (\nabla f(\lambda))^T d\lambda \\
&= \frac{G^{-1}(\lambda)\nabla f(\lambda)}{\gamma}
\end{aligned}$$

Натуральный градиент

$$\tilde{\nabla} f(\lambda) = G^{-1}(\lambda)\nabla f(\lambda)$$

Известно, что

$$\|d\lambda\|_G^2 = D_{KL}^{sym}(\lambda, \lambda + d\lambda) = d\lambda^T G(\lambda) d\lambda$$

Что это за метрика G?

$$\begin{aligned} G(\lambda) &= \mathbb{E}_\lambda \left[(\nabla_\lambda \log q(\beta|\lambda))(\nabla_\lambda \log q(\beta|\lambda))^T \right] \\ &= \nabla_\lambda^2 a_g(\lambda) \end{aligned}$$

Теперь вспомним про градиент нижней оценки

$$\begin{aligned} \nabla_\lambda \mathcal{L}(\lambda) &= \nabla_\lambda^2 a_g(\lambda)(\mathbb{E}_{q(\mathbf{z})} \left[\eta_g(\mathbf{x}, \mathbf{z}, \alpha) \right] - \lambda) \\ &= G(\lambda)(\mathbb{E}_{q(\mathbf{z})} \left[\eta_g(\mathbf{x}, \mathbf{z}, \alpha) \right] - \lambda) \end{aligned}$$

Натуральный градиент нижней оценки

$$\tilde{\nabla}_\lambda \mathcal{L}(\lambda) = G^{-1}(\lambda) \underbrace{G(\lambda)(\mathbb{E}_{q^*(\mathbf{z})}\left[\eta_g(\mathbf{x}, \mathbf{z}, \alpha)\right] - \lambda)}_{\nabla_\lambda \mathcal{L}(\lambda)}$$

Можно ничего не делать! (кроме вар. пересчета)

Стохастическая оптимизация

Типичная задача вариационного вывода

$$\mathcal{L}(\lambda) = \sum_{i=1}^{N} \underbrace{\mathbb{E}_{q^*(z_i)} \left[\log p(x_i, z_i|\beta) - \log q(z_i)\right]}_{l(x_i, q(\beta))} + \underbrace{\mathbb{E}_{q(\beta|\lambda)} \left[\log p(\beta|\alpha) - \log q(\beta|\lambda)\right]}_{g(q(\beta))}$$

Стохастический вариационный вывод

$$\lambda^{t+1} = \lambda + \rho_t \tilde{\nabla} \mathcal{L}(\lambda^t)$$

$$\tilde{\nabla} \mathcal{L}(\lambda^t) \approx \mathbb{E}_{q(z_i)} \eta_g^i(\mathbf{x}, \mathbf{z}, \alpha) - \lambda^t$$