Вариационный вывод

Бахтеев Олег

МФТИ

07.09.2017

Вариационный вывод

Вариационный метод — метод решения математических задач с помощью минимизации определенного функционала, используя пробную функцию, которая зависит от небольшого количества параметров.

Глобальный вариационный вывод

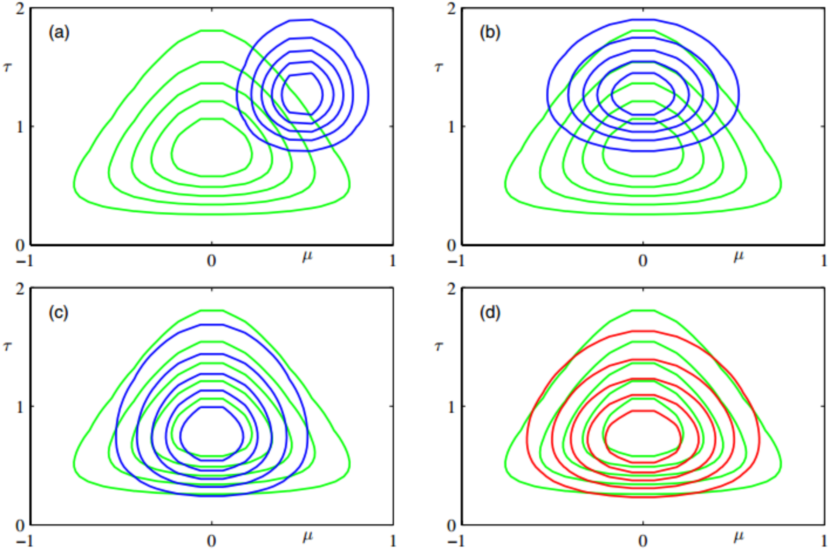
Пусть правдоподобие (модели или выборки) $p(\mathbf{X})$ зависит от скрытой переменной $\mathbf{w}$. Введем аппроксимирующее распределение $q(\mathbf{w})$. Тогда $\log p(\mathbf{X})$ представимо в следующем виде:

$$\log p(\mathbf{X}) = \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{X}, \mathbf{w})}{q(\mathbf{w})} - \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathbf{X})}{q(\mathbf{w})} =$$

$$= \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{X}, \mathbf{w})}{q(\mathbf{w})} + D_{\mathrm{KL}}(q || p(\mathbf{w}|\mathbf{X})).$$

Выбор подходящего распределения позволяет свести задачу поиска правдоподобия к EM-алгоритму.

Пример: аппроксимация нормальным распределением



ELBO

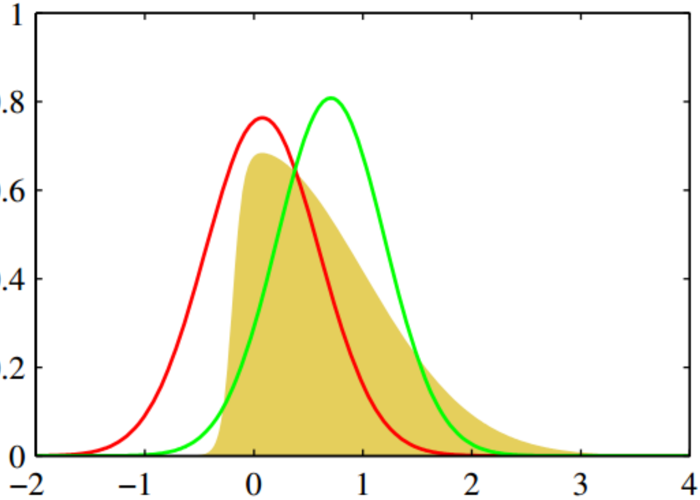
$$\log p(\mathbf{X}) = \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{X}, \mathbf{w})}{q(\mathbf{w})} + D_{\mathrm{KL}}(q||p(\mathbf{w}|\mathbf{X})).$$

Т.к. $D_{\mathrm{KL}}(q||p) \geq 0$, часто проводят оптимизацию только первого слагаемого *(Evidence Lower Bound)*:

$$\log p(\mathbf{X}) \geq \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{X}, \mathbf{w})}{q(\mathbf{w})} =$$

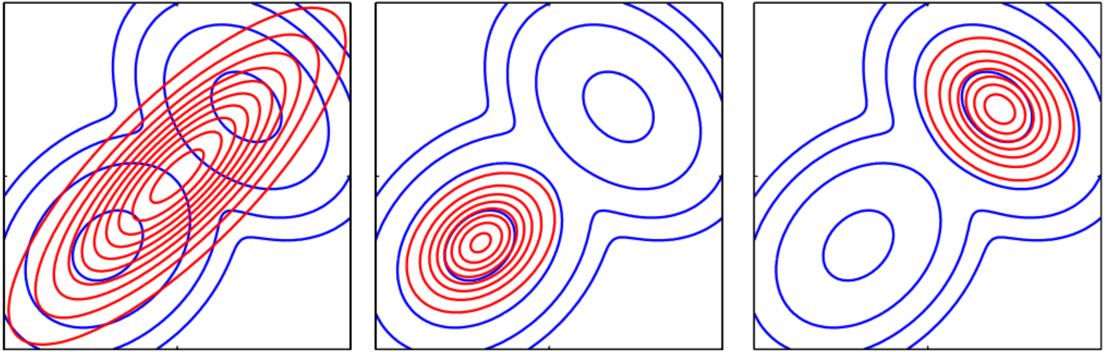
$$= -\mathrm{D}_{\mathrm{KL}}(q(\mathbf{w})||p(\mathbf{w})) + \int_{\mathbf{w}} q(\mathbf{w}) \log\, p(\mathbf{X}|\mathbf{w}).$$

Следствие: Максимизация ELBO эквивалентна минимизации $D_{\mathrm{KL}}(q||p(\mathbf{w}|\mathbf{X}))$.



Апроксимация Лапласа и вариационная оценка

Expectation propagation

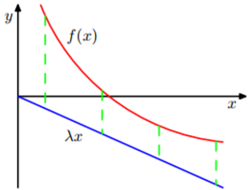


Другие вариационные методы

Локальная аппроксимация

Пусть $p(x) = \exp(-x) = f(x)$.

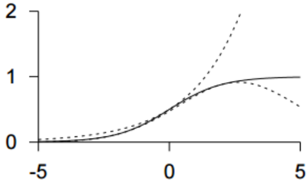
Тогда $\hat{f}(x) \leq f(x_0) + f'(x_0)(x - x0)$.



Локальная аппроксимация: пример

Другие вариационные методы

Ограничения распределения, Jaakkola and Jordan



Upper bound

$$\frac{1}{1+e^{-a}} \leq \exp(\mu a - H_2^e(\mu)) \qquad \mu \in [0,1]$$

Lower bound

$$\frac{1}{1+e^{-a}} \geq g(\nu)\exp\left[(a-\nu)/2 - \lambda(\nu)(a^2-\nu^2)\right]$$

where $\lambda(\nu) = [g(\nu) - 1/2]/2\nu$.

Использование вариационной нижней оценки

Для чего используют variational inference?

- получение оценок Evidence;
- получение оценок распределений моделей со скрытыми переменными (тематическое моделирование, снижение размерности).

Зачем используют variational inference?

- сводит задачу нахождения апостериорной вероятности к методам оптимизации;
- проще масштабируется, чем аппроксимация Лапласа;
- проще в использовании, чем сэмплирующие методы.

Variational Inference может давать сильно заниженную оценку.

Правдоподобие модели (“Evidence”)

$$p(\mathbf{X}|\mathbf{f}) = \int_{\mathbf{w}} p(\mathbf{X}|\mathbf{w})p(\mathbf{w}|\mathbf{f})d\mathbf{w}.$$

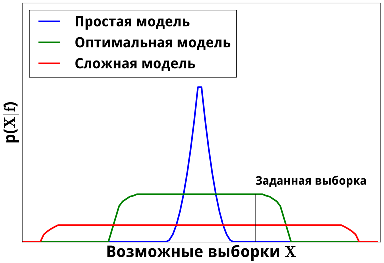
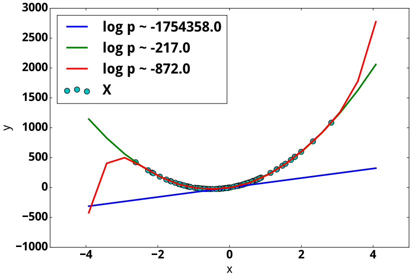


Схема выбора модели по правдоподобию



Пример: полиномы

Разделяющие модели: правдоподобие

Пусть $q \sim \mathcal{N}(\boldsymbol{\mu}_q, \mathbf{A}_q)$.
Тогда вариационная оценка имеет вид:

$$\int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathbf{Y}|\mathbf{X}, \mathbf{w}, \mathbf{f}) d\mathbf{w} + D_{\mathrm{KL}}\bigl(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})\bigr) \simeq$$

$$\sum_{i=1}^{m} \log p(\mathbf{y}_i|\mathbf{x}_i, \mathbf{w}_i) + D_{\mathrm{KL}}\bigl(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})\bigr) \to \max_{\mathbf{A}_q, \boldsymbol{\mu}_q},$$

В случае, если априорное распределение параметров $p(\mathbf{w}|\mathbf{f})$ является нормальным:

$$p(\mathbf{w}|\mathbf{f}) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{A}),$$

дивергенция $D_{\mathrm{KL}}\bigl(q(\mathbf{w})||p(\mathbf{w}|\mathbf{f})$ вычисляется аналитически:

$$D_{\mathrm{KL}}\bigl(q(\mathbf{w}||p(\mathbf{w}|\mathbf{f})\bigr) = \frac{1}{2}\bigl(\mathrm{tr}(\mathbf{A}^{-1}\mathbf{A}_q) + (\boldsymbol{\mu} - \boldsymbol{\mu}_q)^{\mathsf{T}}\mathbf{A}^{-1}(\boldsymbol{\mu} - \boldsymbol{\mu}_q) - n + \ln|\mathbf{A}| - \ln|\mathbf{A}_q|\bigr).$$

Градиентный спуск для оценки правдоподобия

Проведем оптимизацию нейросети в режиме мультистарта из r различных начальных приближений $\mathbf{w}_1, \dots, \mathbf{w}_r$ с использованием градиентного спуска:

$$\mathbf{w}' = \mathbf{w} - \alpha \nabla \sum_{\mathbf{x} \in \mathbf{X}} \log p(\mathbf{x}, \mathbf{w}|\mathbf{f}) = \sum_{\mathbf{x} \in \mathbf{X}} \log p(\mathbf{x}|\mathbf{w}, \mathbf{f}) p(\mathbf{w}|\mathbf{f}).$$

Векторы параметров $\mathbf{w}_1, \dots, \mathbf{w}_r$ соответствуют некоторому скрытому распределению $q(\mathbf{w})$.

Энтропия

Формулу вариационной оценки можно переписать с использованием энтропии:

$$\log p(\mathbf{X}|\mathbf{f}) \geq \int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{X}, \mathbf{w}|\mathbf{f})}{q(\mathbf{w})} d\mathbf{w} =$$

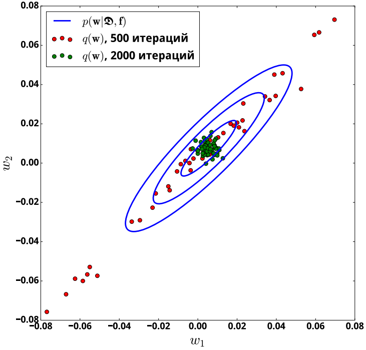
$$\mathsf{E}_{q(\mathbf{w})}[\log p(\mathbf{X}, \mathbf{w}|\mathbf{f})] - \mathsf{S}(q(\mathbf{w})),$$

где $\mathsf{S}(q(\mathbf{w}))$ — энтропия:

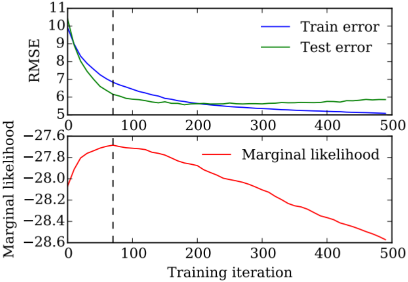
$$\mathsf{S}(q(\mathbf{w})) = -\int_{\mathbf{w}} q(\mathbf{w}) \log q(\mathbf{w}) d\mathbf{w}.$$

Переобучение

Градиентный спуск не минимизирует дивергенцию $\mathrm{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{X}))$. При приближении к моде распределения снижается оценка Evidence, что интерпретируется как переобучение модели.



Схождение распределения к моде

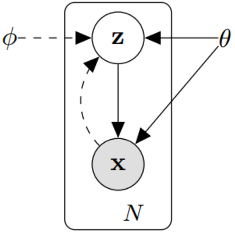


Оценка начала переобучения

Вариационный автокодировщик

Пусть объекты выборки $\mathbf{X}$ порождены при условии скрытой переменной $\mathbf{h} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$\mathbf{x} \sim p(\mathbf{x}|\mathbf{h}, \mathbf{w}).$$



$p(\mathbf{h}|\mathbf{x}, \mathbf{w})$ — неизвестно.
Будем максимизировать вариационную оценку правдоподобия выборки:

$$\log p(\mathbf{x}|\mathbf{w}) \geq \mathsf{E}_{q_\phi(\mathbf{h}|\mathbf{x})} \log\, p(\mathbf{x}|\mathbf{h}, \mathbf{w}) - D_{\mathrm{KL}}(q_\phi(\mathbf{h}|\mathbf{x})||p(\mathbf{h})) \to \max.$$

Распределения $q_\phi(\mathbf{h}|\mathbf{x})$ и $p(\mathbf{x}|\mathbf{h}, \mathbf{w})$ моделируются нейросетью:

$$q_\phi(\mathbf{h}|\mathbf{x}) \sim \mathcal{N}(\boldsymbol{\mu}_\phi(\mathbf{x}), \boldsymbol{\sigma}^2_\phi(\mathbf{x})),$$

$$p(\mathbf{x}|\mathbf{h}, \mathbf{w}) \sim \mathcal{N}(\boldsymbol{\mu}_w(\mathbf{h}), \boldsymbol{\sigma}^2_w(\mathbf{h})),$$

где функции $\boldsymbol{\mu}$,$\boldsymbol{\sigma}$ — выходы нейросети.

Вариационный автокодировщик: evidence

Оценка evidence получается двойным применением вариационной техники:

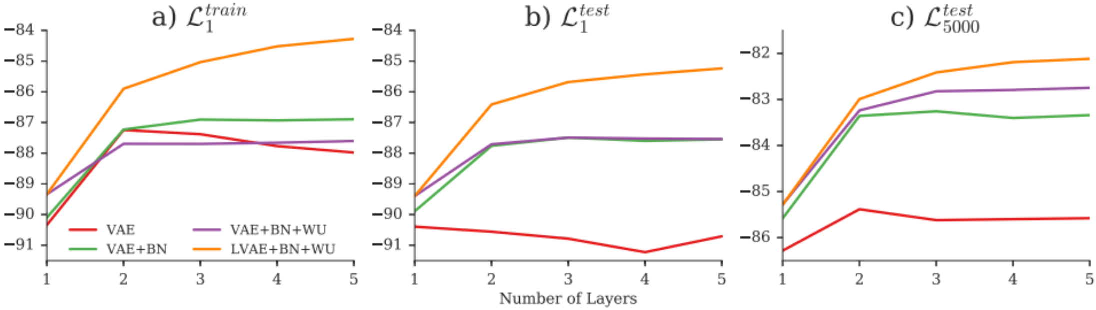
$$\log\, p(\mathbf{X}|\mathbf{f}) \geq \mathsf{E}_{q_\mathbf{w}} \log \hat{p}(\mathbf{x}|\mathbf{w}) + \log\, p(\mathbf{w}|\mathbf{f}) - \log\, q(\mathbf{w}),$$

где $q_\mathbf{w}$ — распределение, аппроксимирующее $p(\mathbf{w}|\mathbf{x}, \mathbf{f})$, $\log\hat{p}(\mathbf{x}|\mathbf{w})$ — вариационная оценка правдоподобия выборки.
Для оптимизации вариационных параметров применяется следующая параметризация:

$$\hat{\mathbf{w}} = \boldsymbol{\mu}_\mathbf{w} + \boldsymbol{\sigma}_\mathbf{w} \odot \epsilon_1, \quad \hat{\mathbf{h}} = \boldsymbol{\mu}_\mathbf{h} + \boldsymbol{\sigma}_\mathbf{h}(\mathbf{h}) \odot \epsilon_2,$$

$$\epsilon_1, \epsilon_2 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}).$$

Стэк автокодировщиков



Normalizing flow

$$\mathbf{h}_0 \sim q(\mathbf{h}_0|\mathbf{x}),$$

$$\mathbf{h}_t = \mathbf{f}(\mathbf{h}_{t-1}).$$

Итоговое распределение вычислимо с помощью якобианов трансформации $\mathbf{f}$:

$$\log q(\mathbf{h}_t|\mathbf{x}) = \log q(\mathbf{h}_0|\mathbf{x}) - \sum_{t=1}^{T} \log \det \left| \frac{d\mathbf{h}_t}{d\mathbf{h}_{t-1}} \right|$$

Подход расширяет множество рассматриваемых функций $\mathbf{f}$. Например, такой функцией может быть функция оптимизации.

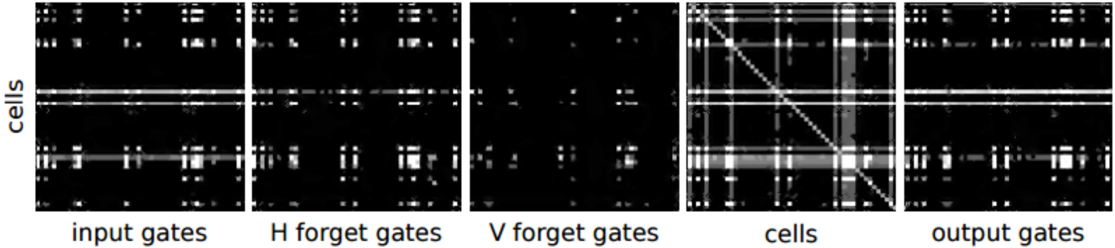
Еще интересные работы

- “Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning” (Gal et al) — Dropout как вариационная аппроксимация гауссового процесса.
- “Semi-Supervised Learning with Deep Generative Models” (Kingma et al) — semi-supervised VAE, полностью генеративный.
- “Markov Chain Monte Carlo and Variational Inference: Bridging the Gap” (Salimans et al) — MCMC + Variation inference.
- “Bayesian Learning via Stochastic Gradient Langevin Dynamics” (Welling et al) — динамика Ланжевена.
- “Stick-breaking variational autoencoders” (Nalisnick et al) — встраивание Дирихле-процесса в VAE.

Practical Variational Inference for Neural Networks, Graves

Применяется вариационный вывод для модели классификации/регрессии. Рассматривается LSTM.
Предлагается метод прунинга, зависящий от вариационных параметров модели:

$$\left|\frac{m_i}{\sigma_i}\right| < \alpha.$$



Neural Variational Inference for Text Processing, Miao

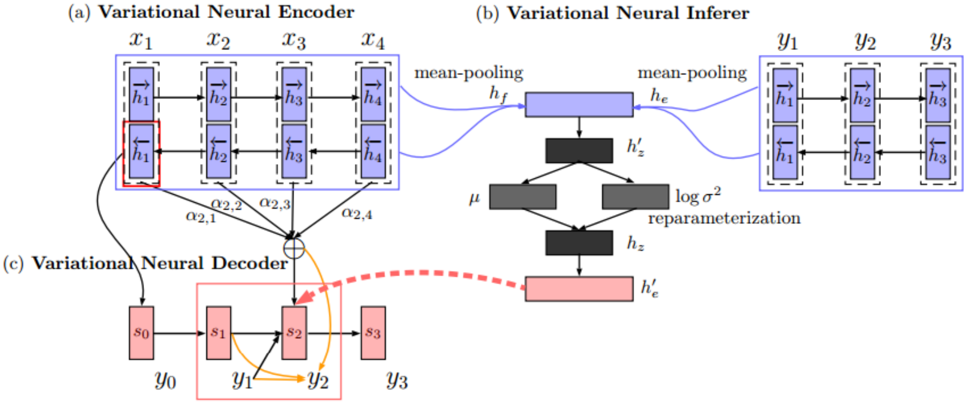
Стохастические предполдожения о переменных добавляются поверх LSTM-функции:

$$\mathbf{u} = g(\mathsf{LSTM}(\mathbf{x})),$$

$$p(\mathbf{h}|\mathbf{x}) = \mathcal{N}(\mathbf{l}_1(\mathbf{u}), \mathbf{l}_2(\mathbf{u})).$$

Применяется Attention, зависящий только от последнего состояния Encoder'a.

Variational Neural Machine Translation, Zhang et al

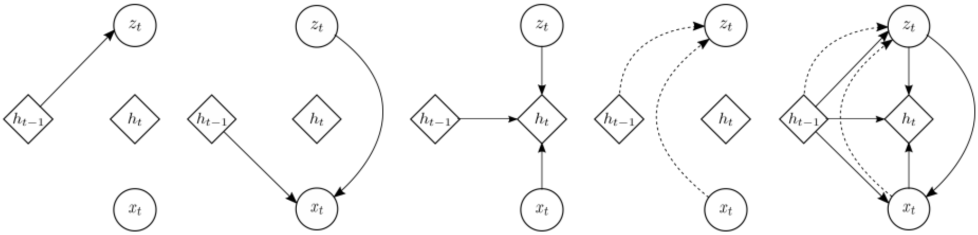


A Recurrent Latent Variable Model for Sequential Data, Chung et al

Последовательность $\mathbf{z}_t$ моделируется как стохастическая, скрытое состояние $\mathbf{h}_t$ зависит от нее:

$$\mathbf{x}_t|\mathbf{z}_t \sim \mathcal{N}(\mathbf{f}^{\text{DEC}}(\mathbf{f}^{\mathbf{Z}}(\mathbf{z}_t), \mathbf{h}_{t-1})),$$

$$\mathbf{h}_t = (\mathbf{f}^{\text{ENC}}((\mathbf{f}^{\mathbf{X}}(\mathbf{x}_t, \mathbf{f}^{\mathbf{Z}}(\mathbf{z}_t), \mathbf{h}_{t-1}).$$



(a) Prior (b) Generation (c) Recurrence (d) Inference (e) Overall

Learning Hard Alignments with Variational Inference, Lawson et al