Байесовское мультимоделирование: вариационный вывод

Московский Физико-Технический Институт

2021

Основа вариационного вывода

Задача вариационного исчисления — найти функцию, на которой заданный функционал достигает экстремального значения.

Пример

Найти плотность распределения p, доставляющую максимум энтропии $H = -\int_w \log p(w) p(w) dw$.

- p — функция
- H — функционал

Если функция задается из явно ограниченного семейства функций, то задачу вариационного исчисления можно рассматривать как задачу аппроксимации.

Выбор модели: связанный байесовский вывод

Первый уровень: выбираем оптимальные параметры:

$$\boldsymbol{w} = \arg\max \frac{p(\mathfrak{D}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{h})}{p(\mathfrak{D}|\boldsymbol{h})},$$

Второй уровень: выбираем модель, доставляющую максимум обоснованности модели.
Обоснованность модели (“Evidence”):

$$p(\mathfrak{D}|\boldsymbol{h}) = \int_{\boldsymbol{w}} p(\mathfrak{D}|\boldsymbol{w})p(\boldsymbol{w}|\boldsymbol{h})d\boldsymbol{w}.$$

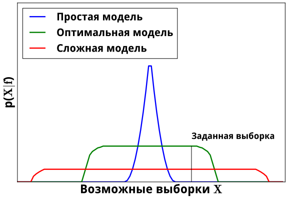
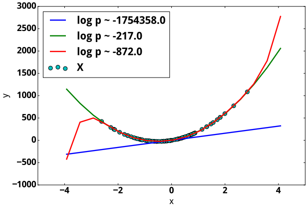


Схема выбора модели

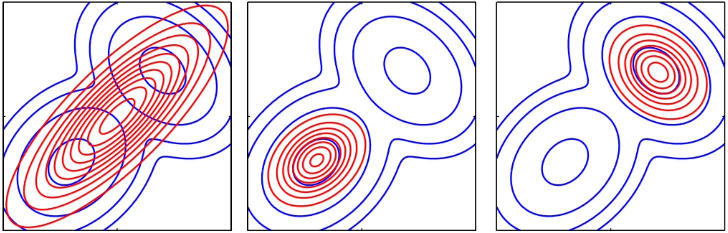


Пример: полиномы

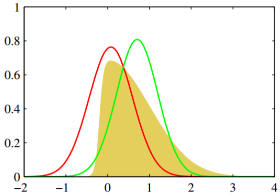
Вариационная оценка, ELBO

Вариационная оценка Evidence, Evidence lower bound — метод нахождения приближенного значения аналитически невычислимого распределения $p(\mathbf{w}|\mathfrak{D}, \mathbf{h})$ распределением $q(\mathbf{w}) \in \mathfrak{Q}$. Получение вариационной нижней оценки обычно сводится к задаче минимизации

$$\mathrm{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathfrak{D})) = -\int_{\mathbf{w}} q(\mathbf{w}) \log \frac{p(\mathbf{w}|\mathfrak{D})}{q(\mathbf{w})} d\mathbf{w} = \mathsf{E}_{\mathbf{w}} \log p(\mathfrak{D}|\mathbf{w}) - \mathrm{KL}(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})).$$



Вариационный вывод и expectation propogation (Bishop)



Аппроксимация Лапласа и вариационная оценка

Получение вариацонной нижней оценки

Максимизация вариационной нижней оценки

$$\int_{\boldsymbol{w}} q(\boldsymbol{w}) \log \frac{p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{X}, \boldsymbol{h})}{q(\boldsymbol{w})} d\boldsymbol{w}$$

эквивалентна минимизации расстояния Кульбака–Лейблера между распределением $q(\boldsymbol{w}) \in \mathfrak{Q}$ и апостериорным распределением параметров $p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{h})$:

$$\hat{q} = \arg\max_{q \in \mathfrak{Q}} \int_{\boldsymbol{w}} q(\boldsymbol{w}) \log \frac{p(\boldsymbol{y}, \boldsymbol{w}|\boldsymbol{X}, \boldsymbol{h})}{q(\boldsymbol{w})} d\boldsymbol{w} \Leftrightarrow$$

$$\hat{q} = \arg\min_{q \in \mathfrak{Q}} \mathrm{D}_{\mathrm{KL}}\big(q(\boldsymbol{w}) || p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{h})\big),$$

$$\mathrm{D}_{\mathrm{KL}}\big(q(\boldsymbol{w}) || p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{h})\big) = \int_{\boldsymbol{w}} q(\boldsymbol{w}) \log \left(\frac{q(\boldsymbol{w})}{p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{h})} \right) d\boldsymbol{w}.$$

Вариационная оценка и эффективный размер выборки

Утверждение 2

Пусть $m \gg 0$, $\lambda > 0$, $\frac{m}{\lambda} \in \mathbb{N}$, $\frac{m}{\lambda} \gg 0$. Тогда оптимизация функции

$$\mathsf{E}_q \log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w}) - \lambda \mathsf{D}_{\mathsf{KL}}\big(q(\boldsymbol{w})||p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{h})\big)$$

эквивалентна оптимизации вариационной оценки обоснованности для произвольной случайной подвыборки $\hat{\boldsymbol{y}}, \hat{\boldsymbol{X}}$ мощности $\frac{m}{\lambda}$ из генеральной совокупности.

См. также [β-VAE, Fixing Broken ELBO].

Использование вариационной нижней оценки

Для чего используют вариационный вывод?

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

Почему вариационный вывод?

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

Вариационный вывод может давать сильно заниженную оценку.

ELBO: нормальное распределение

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

$$\int_{\mathbf{w}} q(\mathbf{w}) \log p(\mathbf{Y}|\mathbf{X}, \mathbf{w}, \mathbf{h}) d\mathbf{w} - D_{\text{KL}}\bigl(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})\bigr) \simeq$$

$$\sum_{i=1}^{m} \log p(\mathbf{y}_i|\mathbf{x}_i, \hat{\mathbf{w}}) - D_{\text{KL}}\bigl(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})\bigr) \to \max_{\mathbf{A}_q, \boldsymbol{\mu}_q}, \quad \hat{\mathbf{w}} \sim q.$$

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

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

дивергенция $D_{\text{KL}}\bigl(q(\mathbf{w})||p(\mathbf{w}|\mathbf{h})\bigr)$ вычисляется аналитически:

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

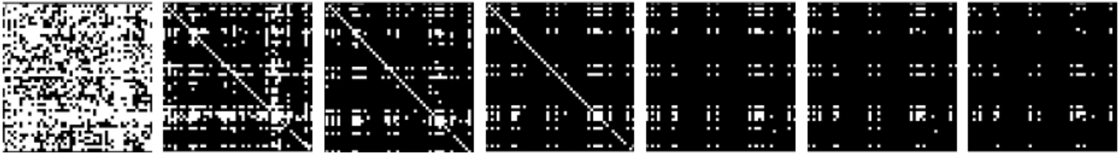
Graves, 2011

Априорное распределение: $p(\boldsymbol{w}|\sigma) \sim \mathcal{N}(\boldsymbol{\mu}, \sigma \boldsymbol{I})$.
Вариационное распределение: $q(\boldsymbol{w}) \sim \mathcal{N}(\boldsymbol{\mu}_q, \sigma_q \boldsymbol{I})$.
Жадная оптимизация гиперпараметров:

$$\boldsymbol{\mu} = \hat{\mathsf{E}}\boldsymbol{w}, \quad \sigma = \hat{\mathsf{D}}\boldsymbol{w}.$$

Прунинг параметра w_i определяется относительной плотностью:

$$\lambda = \frac{q(0)}{q(\boldsymbol{\mu}_{i,q})} = \exp(-\frac{\mu_i^2}{2\sigma_i^2}).$$



ELBO: нормальное распределение

"Обычная" функция потерь:

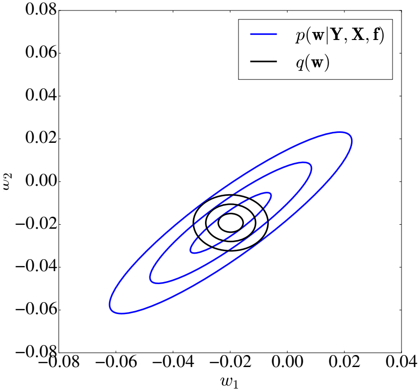
$$L = \sum_{\boldsymbol{x},\boldsymbol{y}\in\mathfrak{D}} -\log p(\boldsymbol{y}|\boldsymbol{x},\boldsymbol{w}) + \lambda||\boldsymbol{w}||_2^2.$$

Вариационный вывод при $(p(\boldsymbol{w}|\boldsymbol{h}) \sim \mathcal{N}(0,1))$:

$$L = \sum_{\boldsymbol{x},\boldsymbol{y}} \log p(\boldsymbol{y}|\boldsymbol{x},\hat{\boldsymbol{w}}) +$$

$$+\frac{1}{2}\left(\mathrm{tr}(\boldsymbol{A}_q) + \boldsymbol{\mu}_q^{\mathsf{T}}\boldsymbol{A}^{-1}\boldsymbol{\mu}_q - \ln|\boldsymbol{A}_q|\right).$$

Пример грубой аппроксимации нормальным диагональным распределением q



Локальная репараметризация

Как считать $\mathsf{E}_q \log p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{w})$?

- Graves, 2011: сэмплируем одну реализацию параметра на каждом шаге оптимизации. Для сэмлирования пользуемся свойством:

$$w \sim \mathcal{N}(\mu, \sigma^2) \to w \sim \varepsilon\sigma + \mu, \quad \varepsilon \sim \mathcal{N}(0, 1).$$

  - Плохое приближение матожидания
- Наивный вариант: засэмплировать столько реализаций параметра, сколько у нас объектов в батче
  - BackProp будет очень медленным

Локальная репараметризация, Kingma et al., 2015

Пусть $y = \mathrm{ReLU}(\boldsymbol{XW})$ и матрица параметров $\boldsymbol{W}$ распределена нормально: $w_{i,j} \sim \mathcal{N}(\mu_{i,j}, \sigma^2_{i,j})$.
Тогда результатом линейной операции $\boldsymbol{XW}$ будет гауссовая матрица:

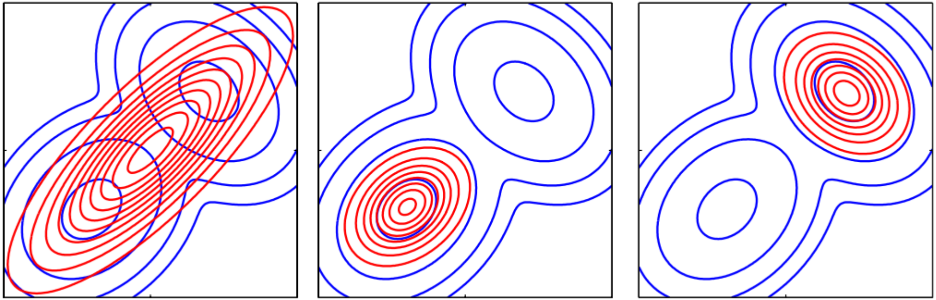
$$\boldsymbol{G} = \boldsymbol{XW}, \quad G_{i,j} \sim \mathcal{N}(\sum_k x_{i,k}\mu_{k,j}, \sum_k x^2_{i,k}\sigma^2_{k,j}).$$

Вместо сэмлирования полноценного вектора параметров для каждого элемента батча на каждом шаге оптимизации, сэмплируем элементы из $\boldsymbol{G}$ (то, что идет перед ReLU).

Пример

Пусть размер батча $= 64$, матрица $\boldsymbol{W}$ имеет размер 64×64.

- Graves: сэмплируем параметры один раз, $64 \times 64 = 4096$ элемента. Приближение матожидания одной реализацией.
- Наивный вариант: сэмплируем параметры 64 раза, $64 \times 64 \times 64 = 262144$ элемента.Приближение матожидания 64 реализациями.
- Локальная репараметризация: сэмплируем $\boldsymbol{G}$, $64 \times 64 = 4096$ элемента. Приближение матожидания одной реализациями.



Expectation propagation

Minka, 2001: представим апостериорное и аппроксимирующее распределение через произведение факторов: $p(\boldsymbol{w}|\mathfrak{D}) = \prod_i f_i$, $q(\boldsymbol{w}) = \prod_i \tilde{f}_i$.
Основная идея метода — минимизация $KL(p(\boldsymbol{w}|\mathfrak{D})|q(\boldsymbol{w}))$.

- Выбираем фактор $\tilde{f}_i$, который будем приближать, «Убираем» его из рассмотрения и заменяем на истинный фактор:
$$q^i \propto f_i \prod_{j \neq i} \tilde{f}_j$$
- Приравниваем моменты q^i к моментам приближенного распределения (корректное приближение, если q из экспоненциального семейства)
- Повторять, пока не сойдемся

Expectation propagation: плюсы и минусы

Минусы:

- Вводится предположение на апостериорное распределение (достаточно мягкое)
- В оригинальной версии работает только для q из экспоненциального семейства
- Нет гарантий сходимости

Плюсы:

- Приближает KL, а не его нижнюю оценку

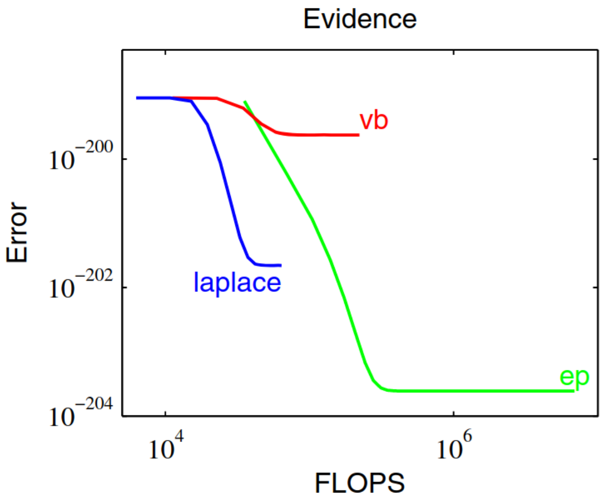


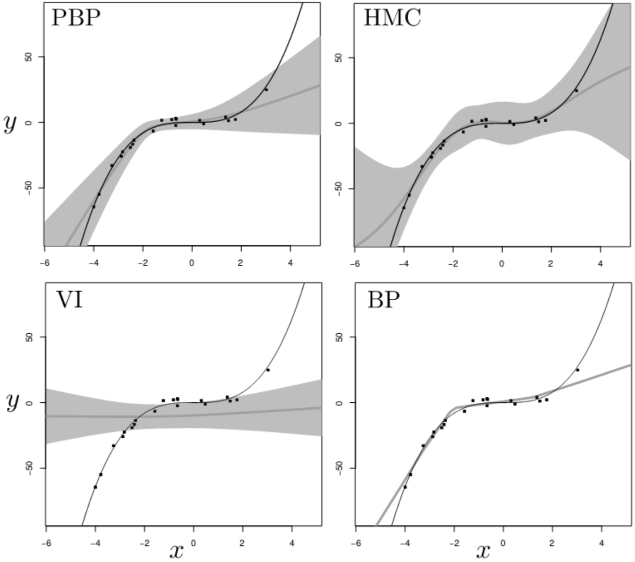
График для модели 2-х компонентной гауссовой смеси.

Probabilistic backpropagation

Комбинация Expectiation propagation и backpropagation.
Backward pass:
Обновляем параметры для каждого параметру по правилу Байеса:

$$p(w_i|\mathfrak{D}) = Z^{-1}p(\mathfrak{D}|w_i, \boldsymbol{w}^i)p(w) \to p(w_i|\mathfrak{D}) = Z^{-1}p(\mathfrak{D}|w_i, \boldsymbol{w}^i)\mathcal{N}(w|\mu, \sigma^2).$$

Проблема: вычисление Z.
Backward pass:
Z можно посчитать приближенно, в предположении $\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{w}) \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{v})$.
Для подсчета $\boldsymbol{m}, \boldsymbol{v}$ для сетей с ReLU-активацией в статье предлагается итеративный алгоритм.


PBP
HMC
VI
BP
y
x
50
0
-50
-6
-4
-2
0
2
4

Литература и прочие ресурсы

- Bishop C. M. Pattern recognition //Machine learning. – 2006. – Т. 128. – №. 9.
- MacKay D. J. C., Mac Kay D. J. C. Information theory, inference and learning algorithms. – Cambridge university press, 2003.
- Бахтеев О. Ю., Стрижов В. В. Выбор моделей глубокого обучения субоптимальной сложности //Автоматика и телемеханика. – 2018. – №. 8. – С. 129-147.
- Graves A. Practical variational inference for neural networks //Advances in neural information processing systems. – 2011. – Т. 24.
- Louizos C., Ullrich K., Welling M. Bayesian compression for deep learning //arXiv preprint arXiv:1705.08665. – 2017.
- Kingma D. P., Salimans T., Welling M. Variational dropout and the local reparameterization trick //Advances in neural information processing systems. – 2015. – Т. 28. – С. 2575-2583.
- Higgins I. et al. beta-vae: Learning basic visual concepts with a constrained variational framework. – 2016.
- Alemi A. et al. Fixing a broken ELBO //International Conference on Machine Learning. – PMLR, 2018. – С. 159-168.
- Minka T. P. Expectation propagation for approximate Bayesian inference //arXiv preprint arXiv:1301.2294. – 2013.
- Hernández-Lobato J. M., Adams R. Probabilistic backpropagation for scalable learning of bayesian neural networks //International conference on machine learning. – PMLR, 2015. – С. 1861-1869.