ВАРИАЦИОННЫЕ ПРИБЛИЖЕНИЯ II

Сергей Николенко

СПбГУ — Санкт-Петербург

25 февраля 2022 г.

Random facts:

- 25 февраля в Грузии — День советской оккупации: 25 февраля 1921 года Красная Армия вошла в Тбилиси во время Советско-грузинской войны; первый раз этот день отмечался 25 февраля 2011 года, в этот день в Грузии были приспущены государственные флаги и объявлена минута молчания
- 25 февраля 1831 г. в рамках подавления польского восстания 1830 года состоялось сражение при Грохове между русскими и войсками повстанцев; потери русских составили 9400 человек, поляки лишились 12 000 человек и трёх орудий; польские войска проиграли битву, но взять Варшаву одним ударом не удалось, и фельдмаршал Дибич-Забайкальский был вынужден отступить к базам снабжения
- 25 февраля 1956 г. в ходе XX съезда КПСС Никита Хрущёв выступил с закрытым докладом «О культе личности и его последствиях»; один из очевидцев доклада А.Н. Яковлев вспоминал: «В зале стояла глубокая тишина. Не слышно было ни скрипа кресел, ни кашля, ни шёпота. Никто не смотрел друг на друга — то ли от неожиданности случившегося, то ли от смятения и страха. Шок был невообразимо глубоким»

Байесовский вывод для гауссиана

Нормальное распределение: фиксируем σ

- Сделаем небольшое лирическое отступление, проведём байесовский вывод для нормального распределения:

$$p(x_1, \ldots, x_n \mid \mu, \sigma^2) \propto \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2\right).$$

- Хотим: найти сопряжённое априорное распределение, подсчитать правдоподобие, решить задачу предсказания.
- Для начала зафиксируем σ^2 и будем в качестве параметра рассматривать только μ.

Нормальное распределение: фиксируем σ

- Сопряжённое априорное распределение для μ при фиксированном σ^2 тоже нормальное и выглядит как

$$p(\mu \mid \mu_0, \sigma_0^2) \propto \frac{1}{\sigma_0^n} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right).$$

- Обычно выбирают $\mu_0 = 0$, $\sigma_0^2 \to \infty$ (порой буквально).
- Давайте рассмотрим сначала случай ровно одного наблюдения x и найдём $p(\mu \mid x)$.

Нормальное распределение: фиксируем σ

- При нашем априорном распределении у μ и x совместное нормальное распределение:

$$x = \mu + \sigma\epsilon, \quad \mu = \mu_0 + \sigma_0\delta, \quad \epsilon, \delta \sim \mathcal{N}(0,1).$$

Упражнение. Пусть (z_1, z_2) – случайные величины с совместным нормальным распределением. Докажите, что случайная величина $z_1 \mid z_2$ распределена нормально с параметрами

$$E(z_1 \mid z_2) = E(z_1) + \frac{Cov(z_1, z_2)}{Var(z_2)} \left(z_2 - E(z_2)\right),$$

$$Var(z_1 \mid z_2) = Var(z_1) - \frac{Cov^2(z_1, z_2)}{Var(z_2)}$$

($Var(x) = E\left[(x - Ex)^2\right]$, $Cov(x, y) = E\left[(x - Ex)(y - Ey)\right]$).

Нормальное распределение: фиксируем σ

- В нашем случае:

$$
\begin{aligned}
x = \mu + \sigma\epsilon, \quad \mu = \mu_0 + \sigma_0\delta, \quad \epsilon, \delta \sim \mathcal{N}(0,1),\\
E(x) &= \mu_0,\\
Var(x) &= E(Var(x \mid \mu)) + Var(E(x \mid \mu)) = \sigma^2 + \sigma_0^2,\\
Cov(x,\mu) &= E\left[(x-\mu_0)(\mu-\mu_0)\right] = \sigma_0^2.
\end{aligned}
$$

- Применив упражнение, получаем:

$$
\begin{aligned}
E(\mu \mid x) &= \mu_0 + \frac{\sigma_0^2}{\sigma_0^2+\sigma^2}(x-\mu_0) = \frac{\sigma_0^2}{\sigma_0^2+\sigma^2}x + \frac{\sigma^2}{\sigma_0^2+\sigma^2}\mu_0,\\
Var(\mu \mid x) &= \frac{\sigma^2\sigma_0^2}{\sigma_0^2+\sigma^2} = \frac{1}{\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2}}.
\end{aligned}
$$

Нормальное распределение: фиксируем σ

- Итого:

$$p(\mu \mid x) \sim \mathcal{N}\left(\frac{\sigma_0^2}{\sigma_0^2+\sigma^2}x + \frac{\sigma^2}{\sigma_0^2+\sigma^2}\mu_0, \left(\frac{1}{\sigma_0^2}+\frac{1}{\sigma^2}\right)^{-1}\right).$$

- Опять же, сложные вычисления можно забыть и пользоваться этими формулами.
- Замечание: часто используют $\tau = \frac{1}{\sigma^2}$ как параметр нормального распределения (precision). Тогда

$$\tau_{\mu|x} = \tau_\mu + \tau.$$

Нормальное распределение: фиксируем σ

- А что, если данных больше, $x_1, \ldots, x_n$?
- Тогда можно повторить всё то же самое, а можно заметить, что набор данных описывается своим средним.

Упражнение. Докажите, что если $p(x_i \mid \mu) \sim \mathcal{N}(\mu, \sigma^2)$ и x_i независимы, то $p(\bar{x} \mid \mu) \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

- Для апостериорной вероятности будет

$$p(\mu \mid x_1, \dots, x_n) \propto p(x_1, \dots, x_n \mid \mu) p(\mu) \propto p(\bar{x} \mid \mu) p(\mu) \propto p(\mu \mid \bar{x}).$$

- Подставляя в наш предыдущий результат, получим:

$$p(\mu \mid x_1, \dots, x_n) \sim \mathcal{N}\left(\frac{\sigma_0^2}{\sigma_0^2 + \frac{\sigma^2}{n}} x + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0, \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1} \right).$$

- Если зафиксировать μ и менять σ^2, то сопряжённым априорным распределением будет обратное гамма-распределение:

$$p(\sigma^2 \mid \alpha, \beta) \propto IG(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} \exp\left(\frac{-\beta}{z}\right).$$

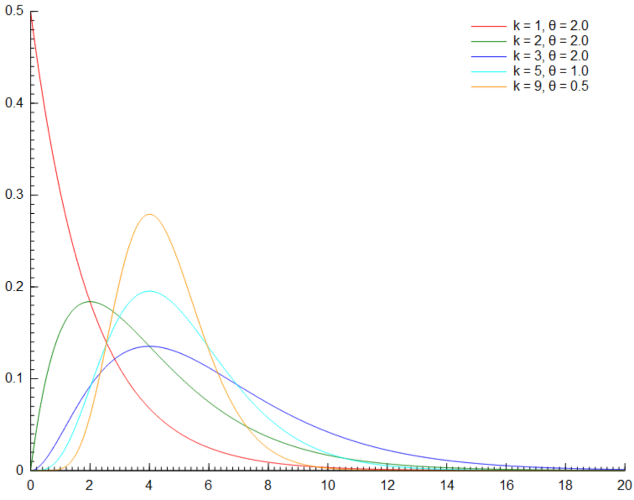
- Тогда в апостериорном распределении будет

$$p(\sigma^2 \mid x_1, \dots, x_n, \alpha, \beta) \propto IG\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum(x_i - \mu)\right).$$

- А в терминах $\tau = \frac{1}{\sigma^2}$ будет обычное гамма-распределение:

$$p(\tau \mid x_1, \dots, x_n, \alpha, \beta) \propto \mathrm{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum(x_i - \mu)\right).$$

Гамма--распределение



Когда и μ, и σ^2 меняются

- Что делать, когда и μ, и σ^2 меняются?
- Можно было бы предположить, что μ и σ^2 независимы; тогда просто априорное распределение будет

$$p(\mu, \sigma \mid \mu_0, \sigma_0, \alpha, \beta) \propto \mathcal{N}(\mu_0, \sigma_0^2) \cdot IG(\alpha, \beta).$$

- К сожалению, это распределение не будет сопряжённым к нормальному. Почему?

Когда и μ, и σ^2 меняются

- Что делать, когда и μ, и σ^2 меняются?
- Можно было бы предположить, что μ и σ^2 независимы; тогда просто априорное распределение будет

$$p(\mu, \sigma \mid \mu_0, \sigma_0, \alpha, \beta) \propto \mathcal{N}(\mu_0, \sigma_0^2) \cdot IG(\alpha, \beta).$$

- К сожалению, это распределение не будет сопряжённым к нормальному. Почему?
- Потому что μ и σ^2 зависимы. :) Новая точка x вводит зависимость между ними.
- В результате получается распределение Стьюдента.

- Вообще говоря, всё, о чём мы говорили – частные случаи *экспоненциального семейства* распределений:

$$p(\mathbf{x} \mid \eta) = h(\mathbf{x}) g(\eta) e^{\eta^\top \mathbf{u}(\mathbf{x})}.$$

- η называются *естественными параметрами* (natural parameters).

- Например, распределение Бернулли:

$$p(x \mid \mu) = \mu^x (1-\mu)^{1-x} = e^{x \ln \mu + (1-x)\ln(1-\mu)} =$$
$$= (1-\mu) e^{\ln\left(\frac{\mu}{1-\mu}\right) x},$$

и естественный параметр получился $\eta = \ln\left(\frac{\mu}{1-\mu}\right)$:

$$p(x \mid \eta) = \sigma(-\eta) e^{-\eta x},$$

где $\sigma(y) = \frac{1}{1+e^{-y}}$ – *сигмоид-функция*.

- Для мультиномиального распределения с параметрами $\mu_1, \dots, \mu_{M-1}$ получаются

$$\eta_k = \ln\left(\frac{\mu_k}{1 - \sum_j \mu_j}\right) \text{ и}$$

$$p(\mathbf{x} \mid \eta) = \left(1 + \sum_{k=1}^{M-1} e^{\eta_k}\right)^{-1} e^{\eta^\top \mathbf{x}}.$$

Упражнение. Проверьте!

- Так вот, для распределений из экспоненциального семейства

$$p(\mathbf{x} \mid \eta) = h(\mathbf{x})g(\eta)e^{\eta^\top \mathbf{u}(\mathbf{x})}$$

можно сразу оптом найти сопряжённые априорные распределения:

$$p(\eta \mid \chi, \nu) = f(\chi, \nu)g(\eta)^\nu e^{\nu\eta^\top \chi},$$

где χ – гиперпараметры, а g то же самое, что в исходном распределении.

Упражнение. Проверьте это и получите вышеописанные примеры как частные случаи.

Когда и μ, и σ^2 меняются

- В настоящем сопряжённом априорном распределении будут:

$$\begin{aligned} x \mid \mu, \tau &\sim \mathcal{N}(\mu, \tau), \\ \mu \mid \tau &\sim \mathcal{N}(\mu_0, n_0\tau), \\ \tau &\sim G(\alpha, \beta). \end{aligned}$$

- Давайте выясним, как изменятся параметры, и заодно докажем.

Когда и μ, и σ^2 меняются

- Самое простое – это, по уже известным результатам,

$$\mu \mid x, \tau \sim \mathcal{N}\left(\frac{n\tau}{n\tau + n_0\tau}\bar{x} + \frac{n_0\tau}{n\tau + n_0\tau}\mu_0, n\tau + n_0\tau\right).$$

- Затем давайте разберёмся с $\tau \mid x$:

$$p(\tau, \mu \mid x) \propto p(\tau) \cdot p(\mu \mid \tau) \cdot p(x \mid \tau, \mu),$$

  и мы хотим это распределение маргинализовать по μ...

Когда и μ, и σ^2 меняются

- Подсчитаем:

$$
\begin{aligned}
p(\tau, \mu \mid x) &\propto p(\tau) \cdot p(\mu \mid \tau) \cdot p(x \mid \tau, \mu) \\
&\propto \tau^{\alpha-1} e^{-\tau\beta} \cdot \tau^{\frac{1}{2}} e^{-\frac{n_0 \tau}{2}(\mu-\mu_0)^2} \cdot \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\sum(x_i-\mu)^2} \\
&\propto \tau^{\alpha+\frac{n}{2}-\frac{1}{2}} e^{-\tau\left(\beta+\frac{1}{2}\sum(x_i-\bar{x})^2\right)} e^{-\frac{\tau}{2}\left(n_0(\mu-\mu_0)^2+n(\bar{x}-\mu)^2\right)}
\end{aligned}
$$

(простой трюк: $x_i - \mu = x_i - \bar{x} + \bar{x} - \mu$).

Когда и μ, и σ^2 меняются

- Теперь надо проинтегрировать

$$\int_\mu e^{-\frac{\tau}{2}\left(n_0(\mu-\mu_0)^2+n(\bar{x}-\mu)^2\right)}d\mu.$$

Упражнение. Проинтегрируйте. :) Должна получиться нормировочная константа

$$\tau^{-\frac{1}{2}}e^{\frac{-nn_0\tau}{2(n+n_0)}(\bar{x}-\mu_0)^2}.$$

Когда и μ, и σ^2 меняются

- Таким образом, получается апостериорное распределение

$$p(\tau \mid x) \propto \tau^{\alpha+\frac{n}{2}-1} e^{-\tau\left(\beta+\frac{1}{2}\sum(x_i-\bar{x})^2+\frac{nn_0}{2(n+n_0)}(\bar{x}-\mu_0)^2\right)}.$$

- Итого результаты такие:

$$\begin{aligned}\mu \mid \tau, x &\sim \mathcal{N}\left(\frac{n\tau}{n\tau+n_0\tau}\bar{x}+\frac{n_0\tau}{n\tau+n_0\tau}\mu_0, n\tau+n_0\tau\right),\\ \tau \mid x &\sim G\left(\alpha+\frac{n}{2}, \beta+\frac{1}{2}\sum\left(x_i-\bar{x}\right)^2+\frac{nn_0}{2(n+n_0)}\left(\bar{x}-\mu_0\right)^2\right).\end{aligned}$$

- Теперь предсказание нового x_{new}:

$$p(x_{\text{new}} \mid x) = \int \int \underbrace{\text{Gamma}}_{\tau|x} \cdot \underbrace{\text{Gaussian}}_{\mu|\tau,x} \cdot \underbrace{\text{Gaussian}}_{x_{\text{new}}|\tau,\mu} d\tau d\mu =$$

$$= \int \underbrace{\text{Gamma}}_{\tau|x} \int \underbrace{\text{Gaussian}}_{\mu|\tau,x} \cdot \underbrace{\text{Gaussian}}_{x_{\text{new}}|\tau,\mu} d\tau d\mu =$$

$$= \int \underbrace{\text{Gamma}}_{\tau|x} \cdot \underbrace{\text{Gaussian}}_{x_{\text{new}}|\tau,x} d\tau = \ldots$$

- В результате получится распределение Стьюдента.

Вариационное приближение для гауссиана

Одномерный гауссиан

- И ещё пример: давайте найдём параметры одномерного гауссиана по точкам $\mathbf{X} = \{x_1, \dots, x_N\}$. Правдоподобие:

$$p(\mathbf{X} \mid \mu, \tau) = \left(\frac{\tau}{2\pi}\right)^{N/2} e^{-\frac{\tau}{2}\sum_{n=1}^{N}(x_n-\mu)^2}.$$

- Вводим сопряжённые априорные распределения:

$$p(\mu \mid \tau) = N(\mu \mid \mu_0, (\lambda_0\tau)^{-1}),$$
$$p(\tau) = \mathrm{Gamma}(\tau \mid a_0, b_0).$$

- Мы это только что подсчитали точно, но давайте приблизим теперь апостериорное распределение как

$$q(\mu, \tau) = q_\mu(\mu) q_\tau(\tau).$$

- На самом деле так не раскладывается!
- Это то, что мы делали для $q(\mathbf{Z}) = \prod_{i=1}^{M} q_i(\mathbf{Z}_i)$. Посчитаем...

- ...$q_\mu(\mu)$ – гауссиан с параметрами

$$\mu_N = \frac{\lambda_0\mu_0 + N\bar{x}}{\lambda_0 + N}, \quad \lambda_N = (\lambda_0 + N)\mathrm{E}[\tau].$$

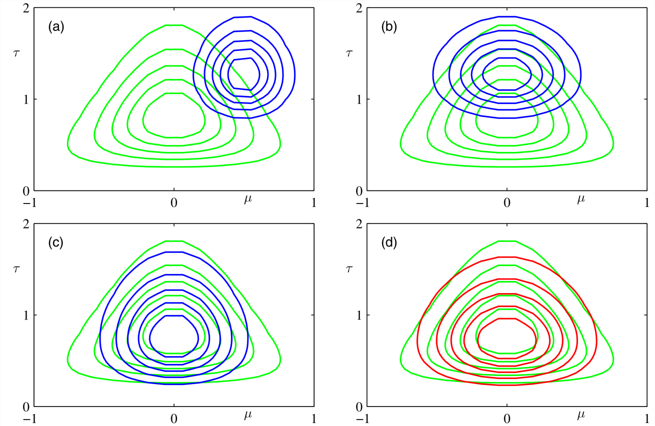
- А $q_\tau(\tau)$ – гамма-распределение с параметрами

$$a_N = a_0 + \frac{N}{2}, \quad b_N = b_0 + \frac{1}{2}\mathrm{E}_\mu\left[\sum_n (x_n - \mu)^2 + \lambda_0(\mu - \mu_0)^2\right].$$

- Всё получилось как надо, но без предположений о форме q_τ и q_μ.

Одномерный гауссиан

- Вот такой вывод в пространстве (μ, τ):



- А для $\mu_0 = a_0 = b_0 = \lambda_0 = 0$ (non-informative priors) можно и точно посчитать...

Одномерный гауссиан

- Получатся моменты для μ

$$\mathrm{E}[\mu] = \bar{x}, \quad \mathrm{E}[\mu^2] = \bar{x}^2 + \frac{1}{N\mathrm{E}[\tau]}.$$

- Это можно подставить и найти $\mathrm{E}[\tau]$:

$$\frac{1}{\mathrm{E}[\tau]} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \bar{x})^2.$$

- Автоматически получили несмещённую оценку дисперсии!

Спасибо!

Спасибо за внимание!