Априорные распределения

Сергей Николенко

Computer Science Club, Екатеринбург, 2011

Outline

Напоминание

- Напоминаю, что основная наша задача – как обучить параметры распределения и/или предсказать следующие его точки по имеющимся данным.
- В байесовском выводе участвуют:
  - $p(x \mid \theta)$ – правдоподобие данных;
  - $p(\theta)$ – априорное распределение;
  - $p(x) = \int_\Theta p(x \mid \theta) p(\theta) d\theta$ – маргинальное правдоподобие;
  - $p(\theta \mid x) = \frac{p(x|\theta)p(\theta)}{p(x)}$ – апостериорное распределение;
  - $p(x' \mid x) = \int_\Theta p(x' \mid \theta) p(\theta \mid x) d\theta$ – предсказание нового x'.
- Задача обычно в том, чтобы найти $p(\theta \mid x)$ и/или $p(x' \mid x)$.

Априорные распределения

- Когда мы проводим байесовский вывод, у нас, кроме правдоподобия, должно быть ещё *априорное распределение* (prior distribution) по всем возможным значениям параметров.
- Мы раньше к ним специально не присматривались, но они очень важны.
- Задача байесовского вывода – как подсчитать $p(\theta \mid x)$ и/или $p(x' \mid x)$.
- Но чтобы это сделать, сначала надо выбрать $p(\theta)$.

Субъективные и объективные априорные распределения

- Априорное распределение может быть
  - субъективным: поговорили с экспертами, поняли, что они говорят, выбрали $p(\theta)$;
  - объективным: априорное распределение берётся из имеющихся (имевшихся ранее) данных и получается тоже байесовскими методами.
- Про субъективные мне, в общем, больше нечего сказать, так что будем говорить об объективных.

Сопряжённые априорные распределения

- Разумная цель: давайте будем выбирать распределения так, чтобы они оставались такими же и *a posteriori*.
- До начала вывода есть априорное распределение $p(\theta)$.
- После него есть какое-то новое апостериорное распределение $p(\theta \mid x)$.
- Я хочу, чтобы $p(\theta \mid x)$ тоже имело тот же вид, что и $p(\theta)$, просто с другими параметрами.

Сопряжённые априорные распределения

- Не слишком формальное определение: семейство распределений $p(\theta \mid \alpha)$ называется семейством *сопряжённых априорных распределений* для семейства правдоподобий $p(x \mid \theta)$, если после умножения на правдоподобие апостериорное распределение $p(\theta \mid x, \alpha)$ остаётся в том же семействе: $p(\theta \mid x, \alpha) = p(\theta \mid \alpha')$.
- α называются *гиперпараметрами* (hyperparameters), это «параметры распределения параметров».
- Тривиальный пример: семейство всех распределений будет сопряжённым чему угодно, но это не очень интересно.

Сопряжённые априорные распределения

- Разумеется, вид хорошего априорного распределения зависит от вида распределения собственно данных, $p(x \mid \theta)$.
- Сопряжённые априорные распределения подсчитаны для многих распределений, мы приведём несколько примеров.

Outline

Испытания Бернулли

- Каким будет сопряжённое априорное распределение для бросания нечестной монетки (испытаний Бернулли)?
- Ответ: это будет бета-распределение; плотность распределения нечестности монетки θ

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

Испытания Бернулли

- Плотность распределения нечестности монетки θ

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

- Тогда, если мы посэмплируем монетку, получив s орлов и f решек, получится

$$p(s, f \mid \theta) = \binom{s+f}{s} \theta^s (1-\theta)^f, \text{ и}$$

$$p(\theta|s, f) = \frac{\binom{s+f}{s}\theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}/B(\alpha, \beta)}{\int_0^1 \binom{s+f}{s}x^{s+\alpha-1}(1-x)^{f+\beta-1}/B(\alpha, \beta)dx} =$$

$$= \frac{\theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}}{B(s+\alpha, f+\beta)}.$$

Испытания Бернулли

- Итого получается, что сопряжённое априорное распределение для параметра нечестной монетки θ – это

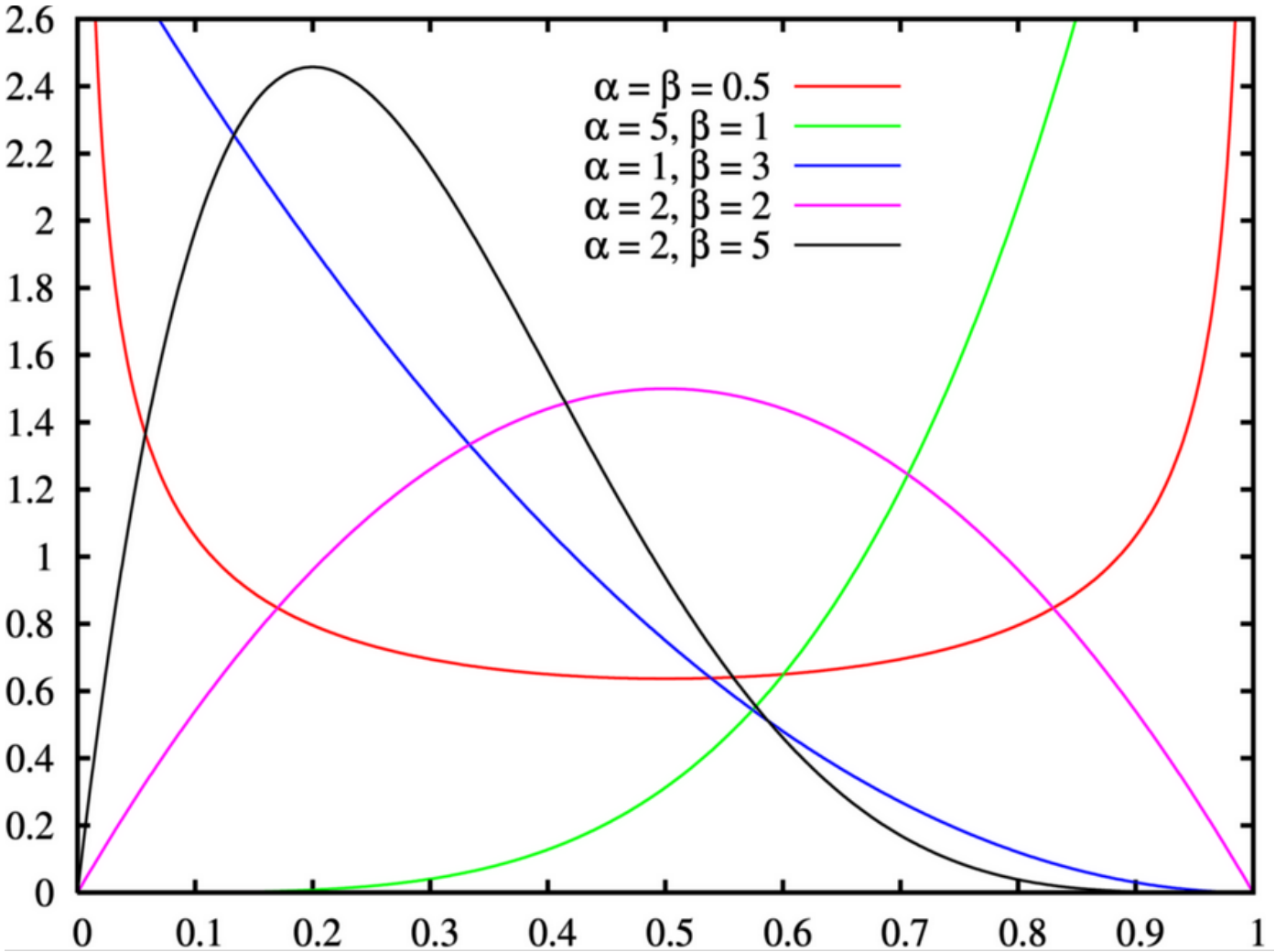
$$p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

- После получения новых данных с s орлами и f решками гиперпараметры меняются на

$$p(\theta \mid s+\alpha, f+\beta) \propto \theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}.$$

- На этом этапе можно забыть про сложные формулы и выводы, получилось очень простое правило обучения (под обучением теперь понимается изменение гиперпараметров).

Бета–распределение



Мультиномиальное распределение

- Простое обобщение: рассмотрим мультиномиальное распределение с n испытаниями, k категориями и по x_i экспериментов дали категорию i.
- Параметры θ_i показывают вероятность попасть в категорию i:

$$p(x \mid \theta) = \binom{n}{x_1, \ldots, x_n} \theta_1^{x_1} \theta_2^{x_2} \ldots \theta_k^{x_k}.$$

- Сопряжённым априорным распределением будет распределение Дирихле:

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_k^{\alpha_k - 1}.$$

Мультиномиальное распределение

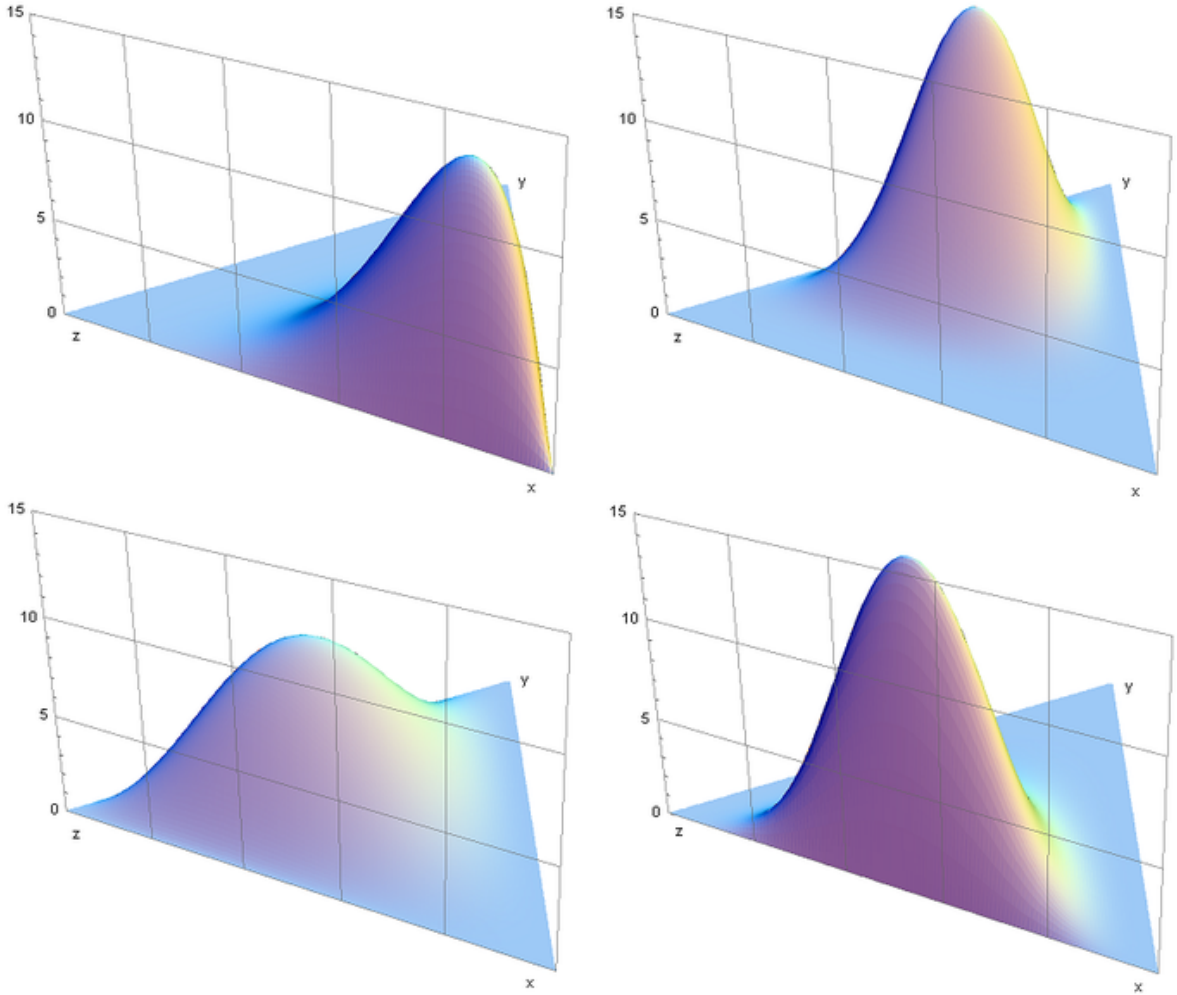
- Сопряжённым априорным распределением будет распределение Дирихле:

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \dots \theta_k^{\alpha_k - 1}.$$

Упражнение. Докажите, что при получении данных $x_1, \dots, x_k$ гиперпараметры изменятся на

$$p(\theta \mid x, \alpha) = p(\theta \mid x + \alpha) \propto \theta_1^{x_1 + \alpha_1 - 1} \theta_2^{x_2 + \alpha_2 - 1} \dots \theta_k^{x_k + \alpha_k - 1}.$$

Распределение Дирихле



Нормальное распределение: фиксируем σ

- Теперь давайте займёмся нормальным распределением:

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

- 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:

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

- Применив упражнение, получаем:

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

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 $p(\bar{x} \mid \mu) \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Нормальное распределение: фиксируем σ

- Для апостериорной вероятности будет

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

- Подставляя в наш предыдущий результат, получим:

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

Нормальное распределение: фиксируем μ

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

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

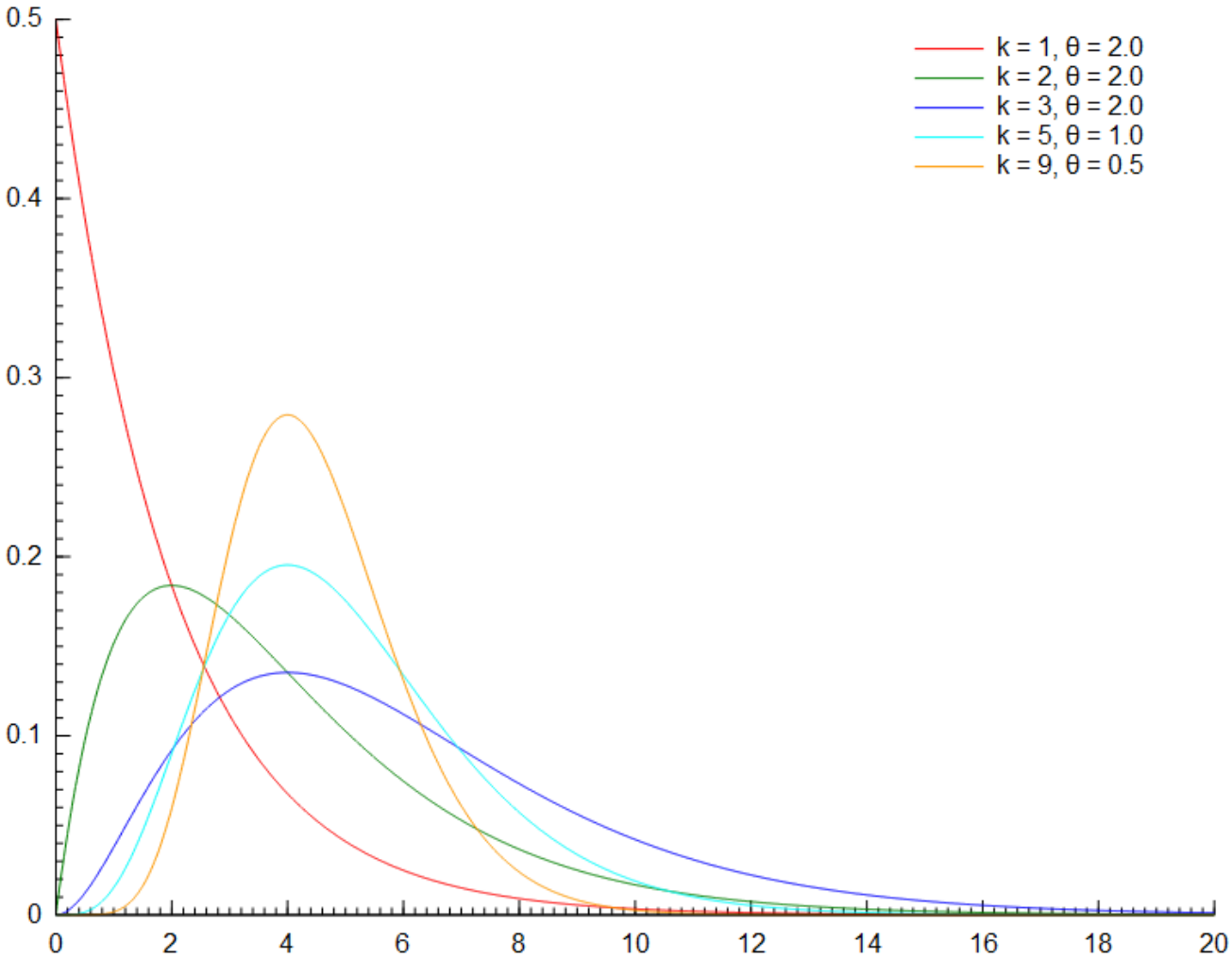
- Тогда в апостериорном распределении будет

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

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

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

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.

Экспоненциальное семейство

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

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

- $\boldsymbol{\eta}$ называются *естественными параметрами* (natural parameters).

Экспоненциальное семейство

- 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*.

Экспоненциальное семейство

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

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

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

Упражнение. Проверьте!

Экспоненциальное семейство

- Так вот, для распределений из экспоненциального семейства

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

можно сразу оптом найти сопряжённые априорные распределения:

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

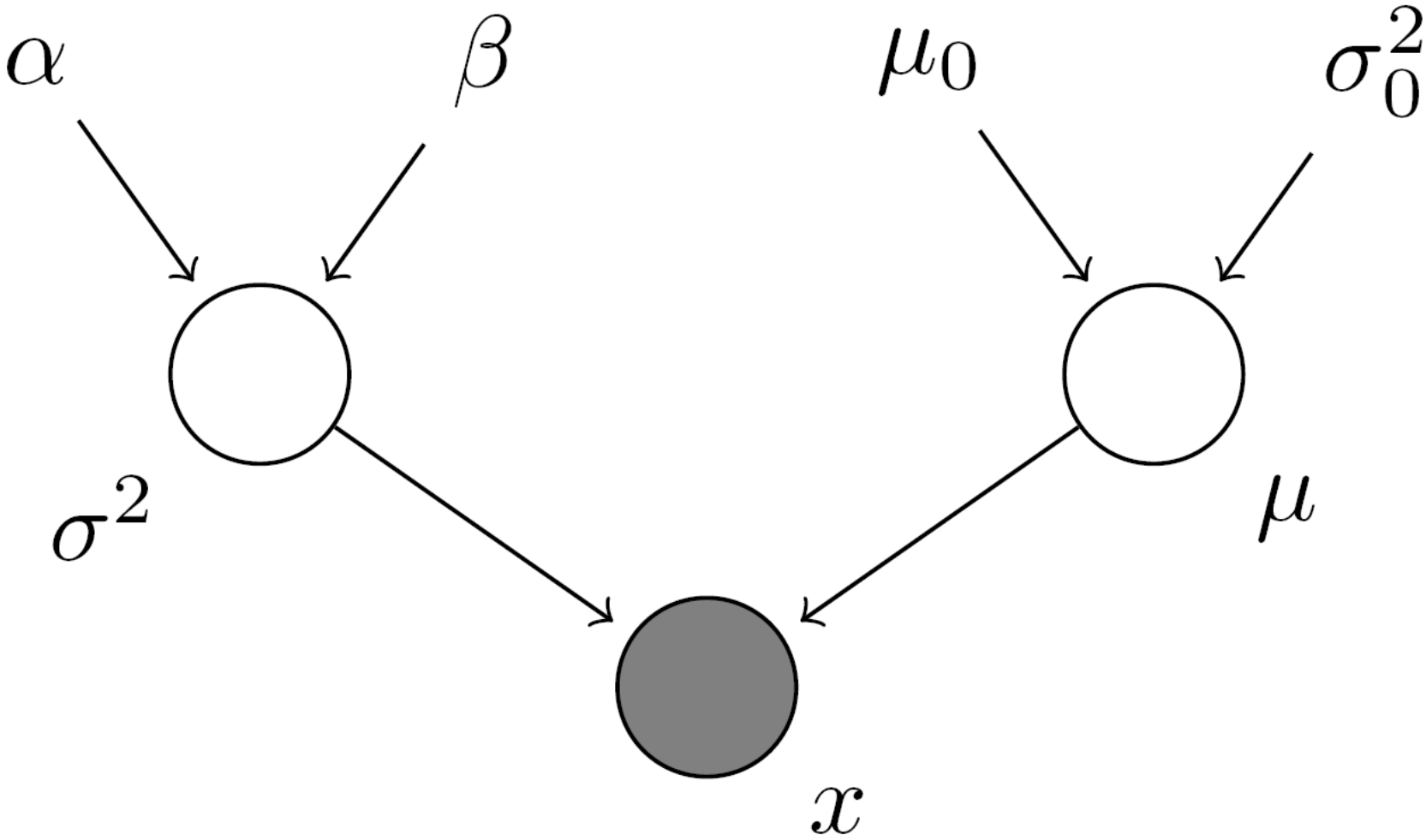
где $\boldsymbol{\chi}$ – гиперпараметры, а g то же самое, что в исходном распределении.

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

Thank you!

Спасибо за внимание!

Графическая модель



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:

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

Предсказание

- 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.