

# Априорные распределения

Сергей Николенко

Академический Университет, весенний семестр 2011

# Outline

- 1 Априорные распределения
  - Задача байесовского вывода
  - Сопряжённые априорные распределения
- 2 Сопряжённые априорные распределения
  - Монетка и мультиномиальное распределение
  - Нормальное распределение с байесовских позиций

# Напоминание

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

# 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  $p(\theta | x)$  и/или  $p(x' | x)$ .
- 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  $p(\theta | x)$ .
- Я хочу, чтобы  $p(\theta | x)$  тоже имело тот же вид, что и  $p(\theta)$ , просто с другими параметрами.

# Сопряжённые априорные распределения

- Не слишком формальное определение: семейство распределений  $p(\theta | \alpha)$  называется семейством *сопряжённых априорных распределений* для семейства правдоподобий  $p(x | \theta)$ , если после умножения на правдоподобие апостериорное распределение  $p(\theta | x, \alpha)$  остаётся в том же семействе:  $p(\theta | x, \alpha) = p(\theta | \alpha')$ .
- $\alpha$  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,  $p(x | \theta)$ .
- Сопряжённые априорные распределения подсчитаны для многих распределений, мы приведём несколько примеров.



# Outline

- 1 Априорные распределения
  - Задача байесовского вывода
  - Сопряжённые априорные распределения
- 2 Сопряжённые априорные распределения
  - Монетка и мультиномиальное распределение
  - Нормальное распределение с байесовских позиций

# Испытания Бернулли

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

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

# Испытания Бернулли

- Плотность распределения нечестности монетки  $\theta$

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

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

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

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

# Испытания Бернулли

- Итого получается, что сопряжённое априорное распределение для параметра нечестной монетки  $\theta$  – это

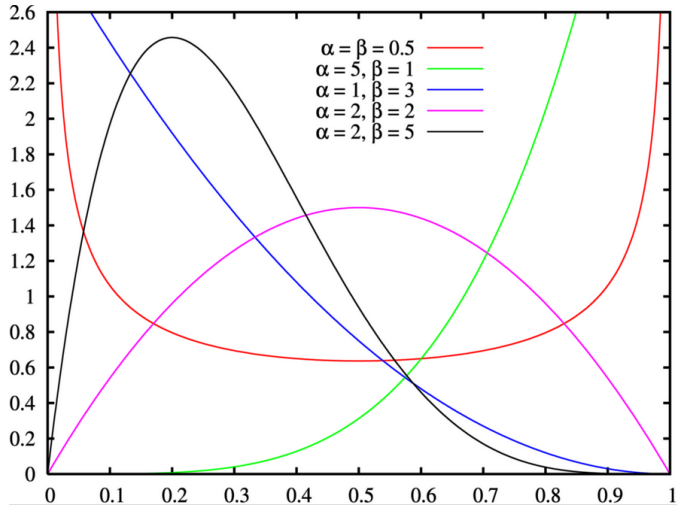
$$p(\theta | \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

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

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

- 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  $\theta_i$  показывают вероятность попасть в категорию  $i$ :

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

- Сопряжённым априорным распределением будет распределение Дирихле:

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

# Мультиномиальное распределение

- Сопряжённым априорным распределением будет распределение Дирихле:

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

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

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

## Нормальное распределение: фиксируем $\sigma$

- Теперь давайте займёмся нормальным распределением:

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

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



# Нормальное распределение: фиксируем $\sigma$

- Сопряжённое априорное распределение для  $\mu$  при фиксированном  $\sigma^2$  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  $\mu_0 = 0$ ,  $\sigma_0^2 \rightarrow \infty$  (порой буквально).
- Давайте рассмотрим сначала случай ровно одного наблюдения  $x$  и найдём  $p(\mu \mid x)$ .

## Нормальное распределение: фиксируем $\sigma$

- При нашем априорном распределении у  $\mu$  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  $z_1 | z_2$  распределена нормально с параметрами

$$E(z_1 | z_2) = E(z_1) + \frac{\text{Cov}(z_1, z_2)}{\text{Var}(z_2)} (z_2 - E(z_2)),$$

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

$$(\text{Var}(x) = E[(x - E_x)^2], \text{Cov}(x, y) = E[(x - E_x)(y - E_y)]).$$

# Нормальное распределение: фиксируем $\sigma$

- В нашем случае:

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

$$E(x) = \mu_0,$$

$$\text{Var}(x) = E(\text{Var}(x | \mu)) + \text{Var}(E(x | \mu)) = \sigma^2 + \sigma_0^2,$$

$$\text{Cov}(x, \mu) = E[(x - \mu_0)(\mu - \mu_0)] = \sigma_0^2.$$

- Применив упражнение, получаем:

$$E(\mu | x) = \mu_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}(x - \mu_0) = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}x + \frac{\sigma^2}{\sigma_0^2 + \sigma^2}\mu_0,$$

$$\text{Var}(\mu | x) = \frac{\sigma^2\sigma_0^2}{\sigma_0^2 + \sigma^2} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}.$$

# Нормальное распределение: фиксируем $\sigma$

- Итого:

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

- 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 $\sigma$

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

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

# Нормальное распределение: фиксируем $\sigma$

- Для апостериорной вероятности будет

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

- Подставляя в наш предыдущий результат, получим:

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

## Нормальное распределение: фиксируем $\mu$

- Если зафиксировать  $\mu$  и менять  $\sigma^2$ , 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

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

**Упражнение.** Докажите это.

# Нормальное распределение: фиксируем $\mu$

- Упражнения к следующему разу.

**Упражнение.** Переформулируйте предыдущий результат в терминах  $\tau$ .

**Упражнение.** Подсчитайте распределение для предсказания при фиксированном  $\mu$ :

$$\begin{aligned} p(x' | x, \mu, \alpha, \beta) &= \int_{\tau} p(x' | x, \mu, \tau, \alpha, \beta) p(\tau | x, \alpha, \beta) d\tau = \\ &= \int_{\tau} p(x' | x, \mu, \tau) p(\tau | \alpha_{post}, \beta_{post}) d\tau. \end{aligned}$$



Thank you!

**Спасибо за внимание!**