Априорные распределения

Сергей Николенко

Казанский Федеральный Университет, 2014

Outline

ML vs. MAP

- Мы остановились на том, что в статистике обычно ищут *гипотезу максимального правдоподобия* (maximum likelihood):

$$\theta_{ML} = \arg\max_\theta p(D \mid \theta).$$

- В байесовском подходе ищут *апостериорное распределение* (posterior)

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

и, возможно, *максимальную апостериорную гипотезу* (maximum a posteriori):

$$\theta_{MAP} = \arg\max_\theta p(\theta \mid D) = \arg\max_\theta p(D \mid \theta)p(\theta).$$

Постановка задачи

- Простая задача вывода: дана нечестная монетка, она подброшена N раз, имеется последовательность результатов падения монетки. Надо определить её «нечестность» и предсказать, чем она выпадет в следующий раз.

Постановка задачи

- Если у нас есть вероятность p_h того, что монетка выпадет решкой (вероятность орла $p_t = 1 - p_h$), то вероятность того, что выпадет последовательность s, которая содержит n_h решек и n_t орлов, равна

$$p(s|p_h) = p_h^{n_h}(1 - p_h)^{n_t}.$$

- Сделаем предположение: будем считать, что монетка выпадает равномерно, т.е. у нас нет априорного знания p_h.
- Теперь нужно использовать теорему Байеса и вычислить скрытые параметры.

Пример применения теоремы Байеса

- Правдоподобие: $p(p_h|s) = \frac{p(s|p_h)p(p_h)}{p(s)}$.
- Здесь $p(p_h)$ следует понимать как непрерывную случайную величину, сосредоточенную на интервале $[0, 1]$, коей она и является. Наше предположение о равномерном распределении в данном случае значит, что априорная вероятность $p(p_h) = 1$, $p_h \in [0, 1]$ (т.е. априори мы не знаем, насколько нечестна монетка, и предполагаем это равновероятным). А $p(s|p_h)$ мы уже знаем.
- Итого получается:

$$p(p_h|s) = \frac{p_h^{n_h}(1 - p_h)^{n_t}}{p(s)}.$$

Пример применения теоремы Байеса

- Итого получается:

$$p(p_h|s) = \frac{p_h^{n_h}(1-p_h)^{n_t}}{p(s)}.$$

- $p(s)$ можно подсчитать как

$$p(s) = \int_0^1 p_h^{n_h}(1-p_h)^{n_t}\,dp_h = \\ = \frac{\Gamma(n_h+1)\Gamma(n_t+1)}{\Gamma(n_h+n_t+2)} = \frac{n_h!n_t!}{(n_h+n_t+1)!},$$

но найти $\arg\max_{p_h} p(p_h \mid s) = \frac{n_h}{n_h+n_t}$ можно и без этого.

Пример применения теоремы Байеса

- Итого получается:

$$p(p_h|s) = \frac{p_h^{n_h}(1-p_h)^{n_t}}{p(s)}.$$

- Но это ещё не всё. Чтобы предсказать следующий исход, надо найти $p(\text{heads}|s)$:

$$p(\text{heads}|s) = \int_0^1 p(\text{heads}|p_h)p(p_h|s)dp_h = \int_0^1 \frac{p_h^{n_h+1}(1-p_h)^{n_t}}{p(s)}dp_h =$$

$$= \frac{(n_h+1)!n_t!}{(n_h+n_t+2)!} \cdot \frac{(n_h+n_t+1)!}{n_h!n_t!} = \frac{n_h+1}{n_h+n_t+2}.$$

- Получили правило Лапласа.

Пример применения теоремы Байеса

- Итого получается:

$$p(p_h|s) = \frac{p_h^{n_h}(1-p_h)^{n_t}}{p(s)}.$$

- Это была иллюстрация двух основных задач байесовского вывода:
  1. найти апостериорное распределение на гипотезах/параметрах:

$$p(\theta \mid D) \propto p(D|\theta)p(\theta)$$

     (и/или найти гипотезу максимального правдоподобия $\arg\max_\theta p(\theta \mid D)$);
  2. найти апостериорное распределение исходов дальнейших экспериментов:

$$p(x \mid D) \propto \int_{\theta\in\Theta} p(x \mid \theta)p(D|\theta)p(\theta)\mathrm{d}\theta.$$

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.

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

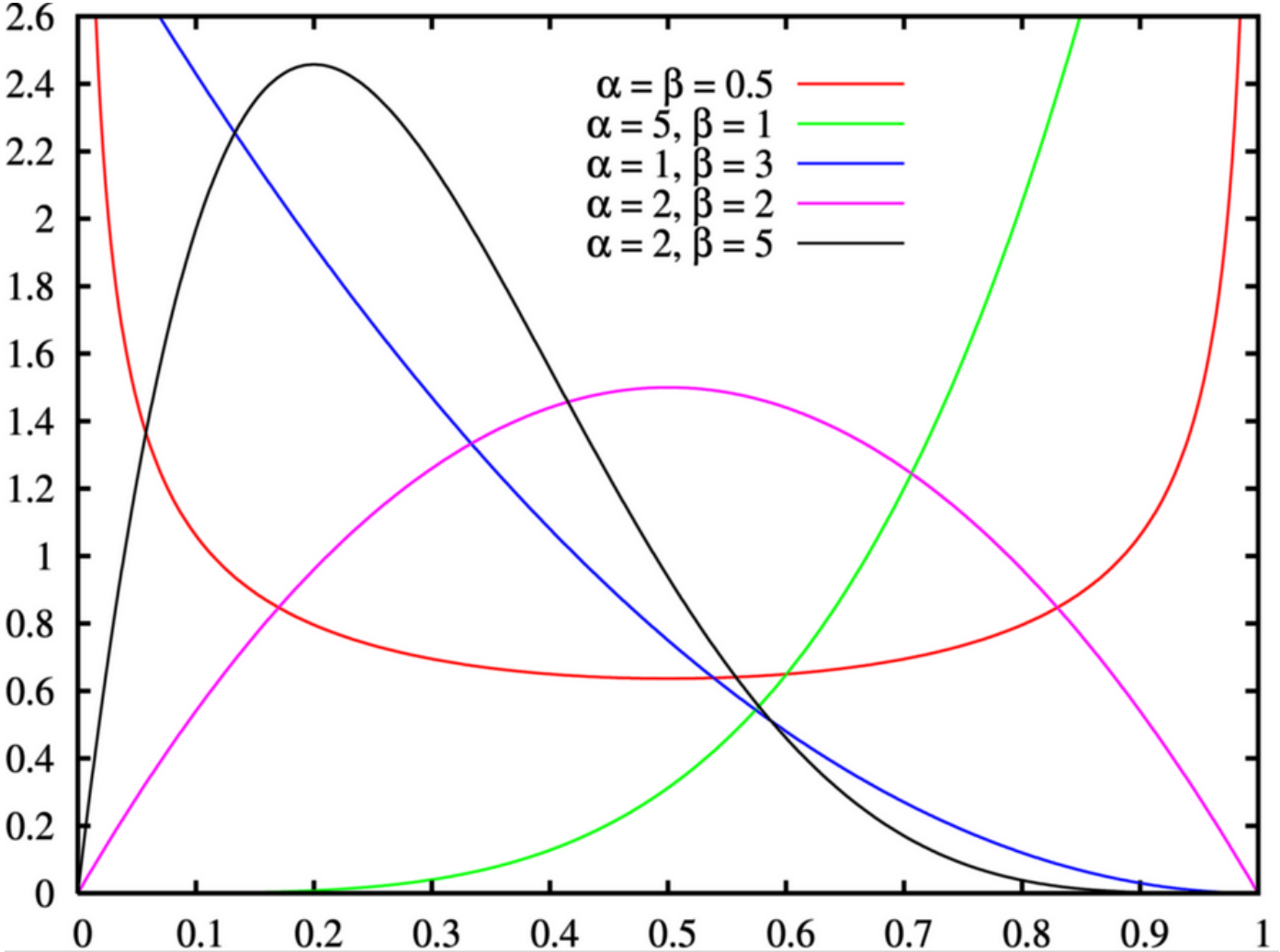
$$p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

- 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

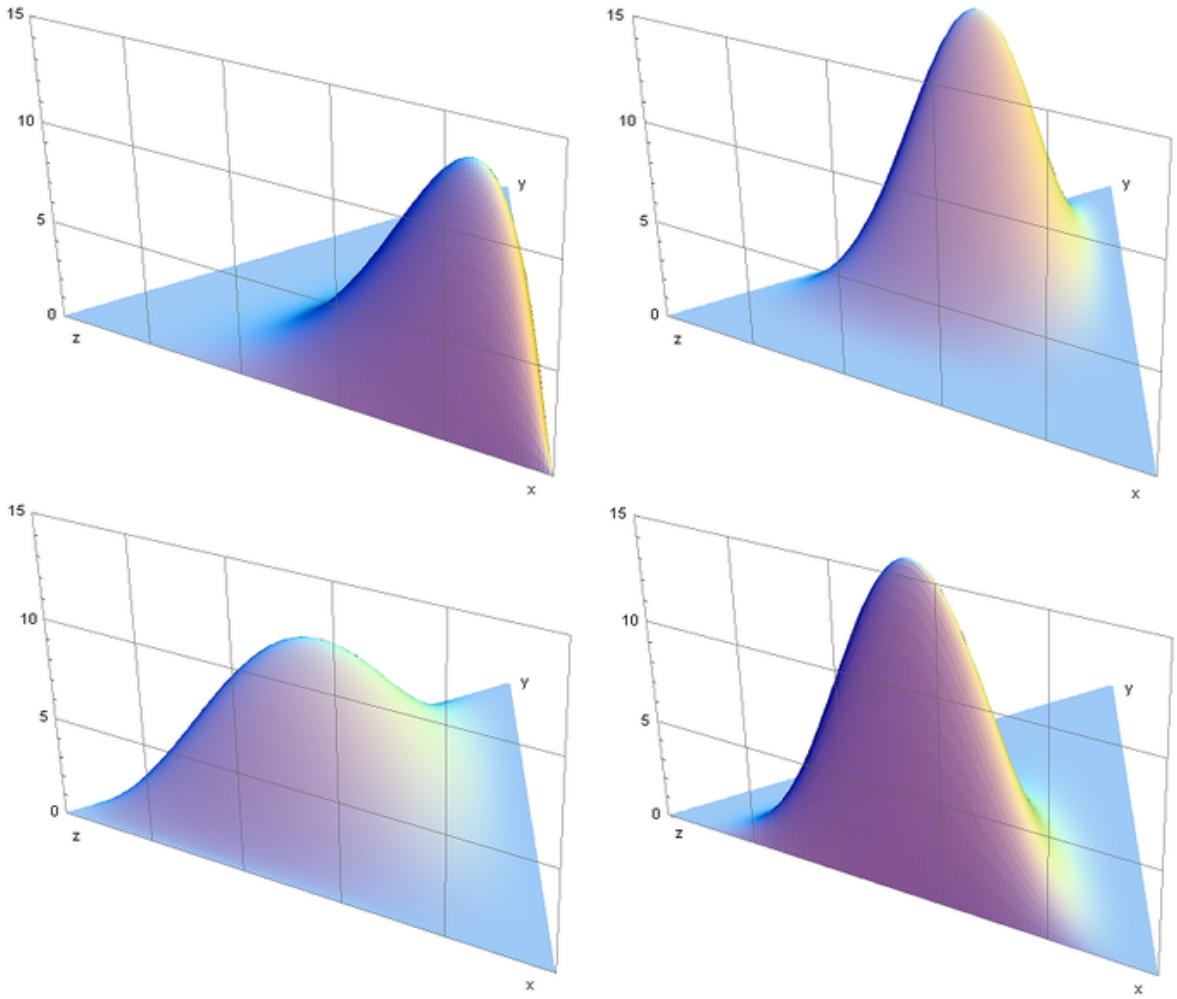
- Сопряжённым априорным распределением будет распределение Дирихле:

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

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



Outline

Нормальное распределение

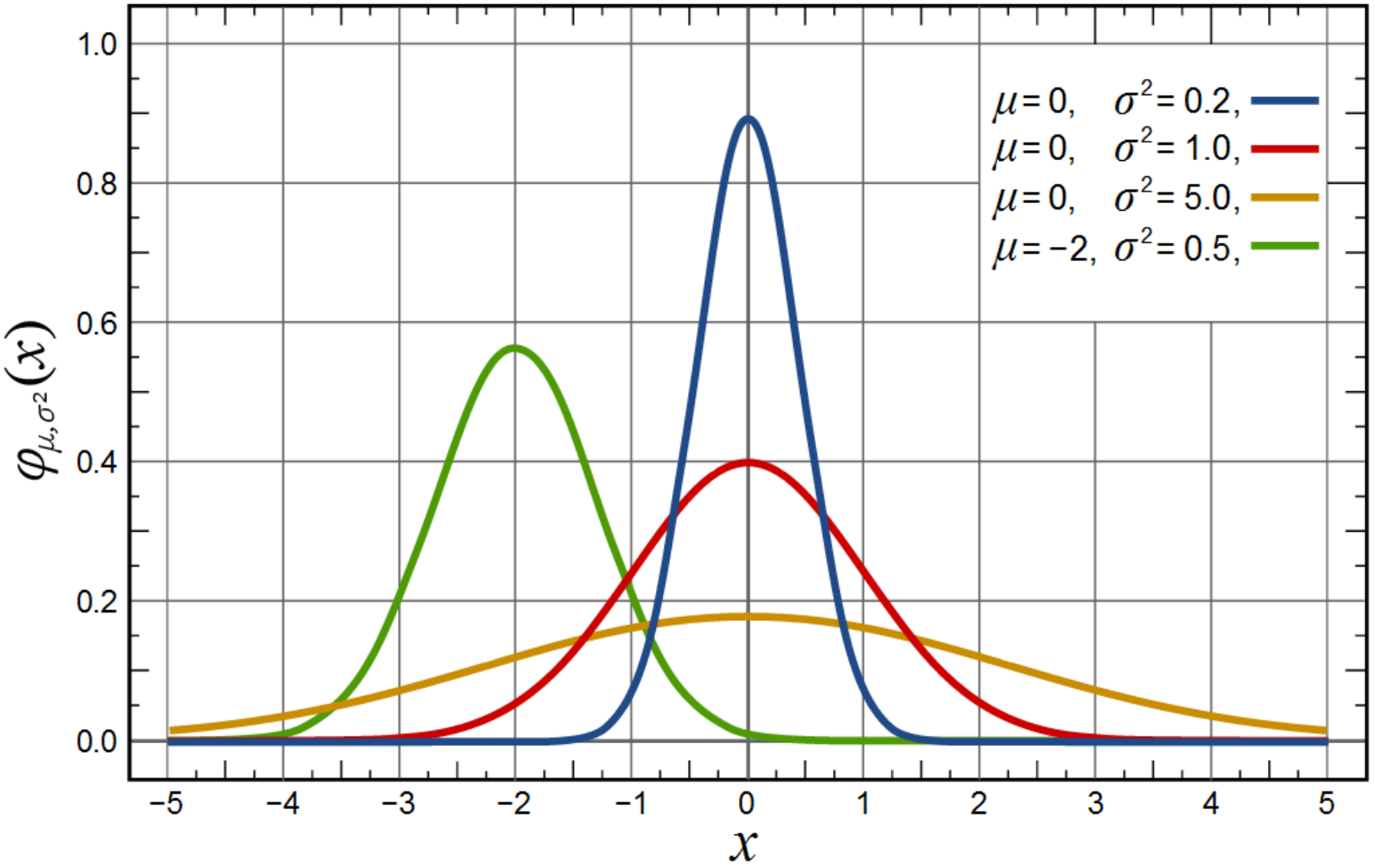
- Мы уже давно знаем нормальное распределение:

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

- Очень многие процессы могут моделироваться нормальным (гауссовским) распределением; обычно возникает, когда есть некое среднее значение μ и шум вокруг него.
- Функция правдоподобия данных $x_1, \ldots, x_n$:

$$p(x_1,\ldots,x_n|\mu,\sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\sum_{i=1}^{n}\frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Нормальное распределение



Гауссиан: достаточные статистики

- Заметим, что функция эта зависит от двух параметров, а не от n:

$$p(x_1, \ldots, x_n|\mu, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{S+n(\bar{x}-\mu)^2}{2\sigma^2}},$$

где

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}, \quad S = \sum_{i=1}^{n} (\bar{x} - x_n)^2.$$

- Параметры $\bar{x}$ и S называются *достаточными статистиками* (sufficient statistics).

Гауссиан: ГМП

- Какие параметры лучше всего описывают данные?
- Перейдём, как водится, к логарифму:

$$\ln p(x_1, \ldots, x_n|\mu, \sigma) = -n \ln(\sigma\sqrt{2\pi}) - \frac{S + n(\bar{x} - \mu)^2}{2\sigma^2}.$$

- Как выяснить, при каких параметрах функция правдоподобия максимизируется?

Гауссиан: ГМП

- Какие параметры лучше всего описывают данные?
- Перейдём, как водится, к логарифму:

$$\ln p(x_1, \ldots, x_n|\mu, \sigma) = -n \ln(\sigma\sqrt{2\pi}) - \frac{S + n(\bar{x} - \mu)^2}{2\sigma^2}.$$

- Как выяснить, при каких параметрах функция правдоподобия максимизируется?
- Взять частные производные и приравнять нулю.

Гауссиан: ГМП

- По μ:
$$\frac{\partial \ln p}{\partial \mu} = -\frac{n}{\sigma^2}(\mu - \bar{x}).$$
- То есть в гипотезе максимального правдоподобия $\mu_{ML} = \bar{x}$, независимо от S.
- Теперь нужно найти σ из гипотезы максимального правдоподобия.
- Для этого мы продифференцируем по $\ln \sigma$ — полезный приём на будущее. Кстати, $\frac{\mathrm{d}x^n}{\mathrm{d}(\ln x)} = nx^n$.

Гауссиан: ГМП

- $$\frac{\partial \ln p}{\partial \ln \sigma} = -n + \frac{\sum_{i=1}^{n} (x_i - \mu)^2}{\sigma^2}.$$
- Следовательно, в гипотезе максимального правдоподобия

$$\sigma_{ML} = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)^2}{n}}.$$

Упражнение. Докажите, что это смещённая оценка, т.е. ожидание этой оценки по настоящему нормальному распределению не равно σ^2.

Thank you!

Спасибо за внимание!