ПРИНЦИП МАКСИМУМА ЭНТРОПИИ

Сергей Николенко

СПбГУ — Санкт-Петербург

27 ноября 2025 г.

Random facts:

- 27 ноября 1095 г. Урбан II на Клермонском соборе по просьбе византийского императора Алексея I провозгласил Первый крестовый поход
- 27 ноября 1838 г. произошла битва при Сан-Хуан-де-Улуа в ходе Кондитерской войны между Мексикой и Францией; война началась по заявлению французского кондитера, якобы ограбленного мексиканскими мародёрами, Франция послала флот, чтобы вернуть долг, при бомбардировке Сан-Хуан-де-Улуа погибли более 60 человек, и президенту Мексики пришлось пообещать выплатить компенсацию Франции; впрочем, в итоге так и не выплатили
- 27 ноября 1895 г. Альфред Нобель подписал завещание, по которому бóльшая часть его состояния поступала в фонд Нобелевской премии
- 27 ноября — день авиакатастроф: в 1962 г. Boeing 707 врезался в гору при заходе на посадку под Лимой (97 погибших), в 1970 Douglas DC-8 выкатился со взлётной полосы и загорелся (47 погибших), в 1983 Boeing 747 под Мадридом зацепил несколько холмов и разрушился (181 погибших), а в 1989 г. в окрестностях Боготы террористы по указанию Пабло Эскобара взорвали Boeing 727 (110 погибших)
- 27 ноября 1992 г. была создана Высшая школа экономики

Принцип максимума энтропии

Принцип максимума энтропии

- Обычно энтропию выводят из аксиомы непротиворечивости

$$H_3(p_1, p_2, p_3) = H_2(p_1, p_2 + p_3) + qH_2\left(\frac{p_2}{p_2 + p_3}, \frac{p_3}{p_2 + p_3}\right)$$

- Но давайте посмотрим на понятие энтропии с другой стороны
- Пусть у нас имеется некоторая информация I, которой должно соответствовать распределение вероятностей $(p_1, \dots, p_n)$
- "Информация"" здесь понимается в самом общем виде: это просто значит, что не любое распределение нам подойдёт

Принцип максимума энтропии

- Будем порождать распределения так: разобьём общую вероятность (единицу) на N маленьких частей (квантов) размера $\delta = \frac{1}{N}$ и будем случайно распределять их между n исходами: N раз бросать честную игральную кость с n гранями
- Тогда, если $p_i = \frac{N_i}{N}$, нам нужно будет распределить ровно N_i этих квантов в i-й исход
- Каждая последовательность из N бросаний выпадает с вероятностью n^{-N}, и вероятность получить распределение $(p_1, \dots, p_n)$ получится из числа сочетаний:

$$n^{-N} \frac{N!}{N_1! \cdot \dots \cdot N_n!}$$

Принцип максимума энтропии

- Если мы будем запускать этот процесс и выбрасывать распределения, не соответствующие I, какое распределение будет чаще всего попадаться?
- Это будет то из распределений, соответствующих I, которое максимизирует величину $\frac{N!}{N_1!\cdot\ldots\cdot N_n!}$; для $N \to \infty$

$$\log N! = N\log N - N + \sqrt{2\pi N} + \frac{1}{12N} + O\left(\frac{1}{N^2}\right)$$

- Поскольку $N_i = p_i N$ и $\sum_{i=1}^{n} N_i = N$, $\log \frac{N!}{N_1!\cdot\ldots\cdot N_n!}$ равен

$$\log N! - \sum_{i=1}^{n} \log N_i! = N\log N - N - \sum_{i=1}^{n} (N_i \log N_i - N_i) + O\left(\sqrt{N}\right) =$$

$$= N\log N - N - \sum_{i=1}^{n} p_i N(\log N + \log p_i) + \sum_{i=1}^{n} N_i + O\left(\sqrt{N}\right) =$$

$$= -N\sum_{i=1}^{n} p_i \log p_i + O\left(\sqrt{N}\right).$$

Принцип максимума энтропии

- Получили

$$\log \frac{N!}{N_1! \cdot \ldots \cdot N_n!} = -N \sum_{i=1}^{n} p_i \log p_i + O\left(\sqrt{N}\right)$$

- При $N \to \infty$ логарифм числа сочетаний растёт как N умножить на энтропию, то есть из такого (вполне естественного) порождающего процесса с наибольшей вероятностью будет получаться именно распределение, максимизирующее энтропию!
- Мы получили другое обоснование понятия энтропии, не опирающееся на аксиому непротиворечивости
- Более того, в этом подходе ещё и вполне естественно получается, что энтропию нужно именно *максимизировать*

Принцип максимума энтропии

- *Принцип максимума энтропии* (maximum entropy principle): для некоторой заданной априорной информации I наиболее характерным распределением вероятностей является распределение с максимальной энтропией среди тех распределений, которые удовлетворяют этой информации
- Рассмотрим кубик, у которого среднее значение составляет $m = 4.5$ вместо 3.5; давайте применим принцип максимума энтропии, то есть максимизируем

$$H\left(p_1, \ldots, p_6\right) = -\sum_{i=1}^{n} p_i \log p_i$$

при двух условиях: $p_1 + \ldots + p_6 = 1$ и $\mathbb{E}\left[i\right] = \sum_{i=1}^{6} i p_i = m$

Принцип максимума энтропии

- Мы можем записать функцию Лагранжа для этой задачи условной оптимизации:

$$L = -\sum_{i=1}^{n} p_i \log p_i - \lambda \sum_{i=1}^{6} i p_i - \mu \sum_{i=1}^{6} p_i,$$

взять от неё производные по p_i и приравнять их к нулю:

$$\frac{\partial L}{\partial p_i} = -\log p_i - 1 - \lambda i - \mu = 0,$$

то есть $p_i = e^{-\lambda i - \mu - 1}$

Принцип максимума энтропии

- Чтобы теперь найти множители Лагранжа λ и μ, нужно применить известные нам условия. Во-первых, $p_1 + ... + p_6 = 1$, то есть

$$\sum_{i=1}^{6} e^{-\lambda i - \mu - 1} = e^{-\mu - 1} \sum_{i=1}^{6} e^{-\lambda i} = 1.$$

- Во-вторых, $\sum_{i=1}^{6} i p_i = m$, то есть

$$\sum_{i=1}^{6} i e^{-\lambda i - \mu - 1} = e^{-\mu - 1} \sum_{i=1}^{6} i e^{-\lambda i} = m.$$

- Отсюда мы можем легко выразить m и μ через λ:

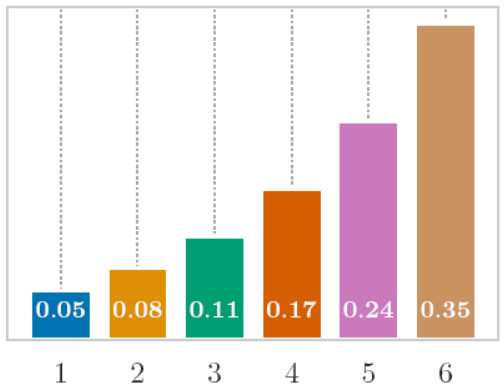
$$m = \frac{\sum_{i=1}^{6} i e^{-\lambda i}}{\sum_{i=1}^{6} e^{-\lambda i}}, \qquad \mu = \log \sum_{i=1}^{6} e^{-\lambda i} - 1$$

Принцип максимума энтропии

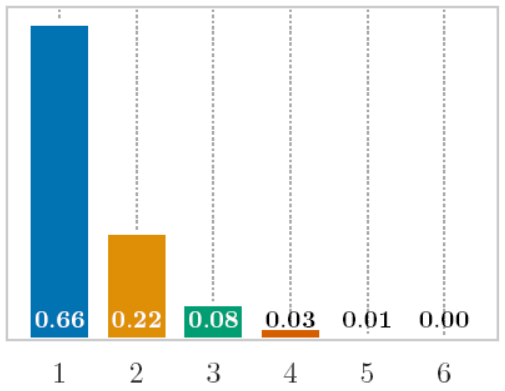
- Теперь осталось найти λ, подставив $m = 4.5$ в первое уравнение (видимо, аналитически не получится), а потом подставить λ во второе и в выражения для p_i
- Для нашего примера численно получается

$$\mathbf{p} = \begin{pmatrix} 0.0544 & 0.0788 & 0.1142 & 0.1654 & 0.2398 & 0.3475 \end{pmatrix}$$

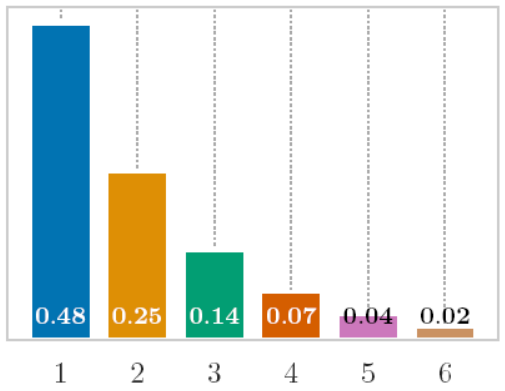
- В сумме они дают единицу, и это распределение скошено в сторону больших значений, как и следовало ожидать из имеющейся априорной информации о среднем



(а) $m = 4.5$



(б) $m = 1.5$



(в) $m = 2$

Принцип максимума энтропии

- Для непрерывного примера давайте попробуем найти распределение на $\mathbb{R}$, достигающее максимальной энтропии с заданным средним μ и дисперсией σ^2
- Мы хотим максимизировать функционал $\int_{-\infty}^{\infty} p(x) \log p(x) \, \mathrm{d}x$, причём среднее и дисперсия задают нам два условия на эту задачу оптимизации:

$$\int_{-\infty}^{\infty} xp(x) \, \mathrm{d}x = \mu, \qquad \int_{-\infty}^{\infty} (x-\mu)^2 p(x) \, \mathrm{d}x = \sigma^2;$$

кроме того, есть ещё базовое условие о том, что $p(x)$ представляет собой плотность вероятности, то есть $\int_{-\infty}^{\infty} p(x) \, \mathrm{d}x = 1$

Принцип максимума энтропии

- Давайте запишем функцию Лагранжа для этой задачи оптимизации, с множителями Лагранжа, соответствующими условиям выше:

$$L(x, \lambda_1, \lambda_2) = \int_{-\infty}^{\infty} p(x) \log p(x) \, \mathrm{d}x - \lambda_0 \left(1 - \int_{-\infty}^{\infty} p(x) \, \mathrm{d}x \right) -$$
$$- \lambda_1 \left(\mu - \int_{-\infty}^{\infty} x p(x) \, \mathrm{d}x \right) - \lambda_2 \left(\sigma^2 - \int_{-\infty}^{\infty} (x - \mu)^2 p(x) \, \mathrm{d}x \right).$$

Принцип максимума энтропии

- Теперь можно применить метод вариационного исчисления; добавим малую вариацию $\delta p\left(x\right)$ к плотности $p\left(x\right)$ и посмотрим, что произойдёт с функционалом:

$$\delta L = \int_{-\infty}^{\infty} \delta p\left(x\right) \left(\ln p\left(x\right) + 1 + \lambda_0 + \lambda_1 \left(x - \mu\right) + \lambda_2 \left(x - \mu\right)^2\right)$$

- В точке максимума вариация функционала должна быть равна нулю для любого малого $\delta p\left(x\right)$, то есть нулю должно быть равно выражение в скобках, и мы получаем, что

$$0 = \ln p\left(x\right) + 1 + \lambda_0 + \lambda_1 \left(x - \mu\right) + \lambda_2 \left(x - \mu\right)^2,$$

то есть

$$p\left(x\right) = e^{-1-\lambda_0-\lambda_1\left(x-\mu\right)-\lambda_2\left(x-\mu\right)^2}$$

Принцип максимума энтропии

- Получилось, что плотность распределения максимальной энтропии с заданными средним и дисперсией — это экспонента с квадратичной функцией в показателе, то есть нормальное распределение!
- Осталось только подставить полученную плотность в условия, чтобы найти значения λ_0, λ_1 и λ_2 (упражнение); получится, что это гауссиан с именно таким средним и именно такой дисперсией:

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

- Метод вариационного исчисления на этом не вполне заканчивается, нужно ещё подтвердить, что результат получился действительно тот, который нам нужен

Принцип максимума энтропии

- Давайте сделаем это другим способом
- Вариационное исчисление дало нам толстый намёк, что ответом должен быть гауссиан; воспользуемся этим намёком и запишем дивергенцию Кульбака–Лейблера между произвольным распределением $q(x)$ и гауссианом $p(x)$:

$$0 \leq \mathrm{KL}(q\|p) = \int_{-\infty}^{\infty} q(x) \log \frac{q(x)}{p(x)} \mathrm{d}x = \\ = -H(q) - \int_{-\infty}^{\infty} q(x) \log p(x) \, \mathrm{d}x.$$

Принцип максимума энтропии

- Второе слагаемое — перекрёстную энтропию между $q(x)$ и $p(x)$ — можно подсчитать примерно так же, как мы выше делали для двух гауссианов:

$$\int_{-\infty}^{\infty} q(x) \log p(x) \, \mathrm{d}x =$$

$$= -\int_{-\infty}^{\infty} q(x) \log \sqrt{2\pi\sigma^2} \mathrm{d}x - \int_{-\infty}^{\infty} q(x) \frac{(x-\mu)^2}{2\sigma^2} \mathrm{d}x =$$

$$= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \int_{-\infty}^{\infty} q(x) \frac{(x-\mu)^2}{2\sigma^2} \mathrm{d}x =$$

$$= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{\sigma_q^2}{2\sigma^2},$$

ведь у нас получилось буквально определение дисперсии распределения $q(x)$

Принцип максимума энтропии

- Но на дисперсию у нас как раз есть условие: $\sigma_q = \sigma$
- А значит,

$$\int_{-\infty}^{\infty} q(x) \log p(x) \, \mathrm{d}x = -\frac{1}{2} \log\left(2\pi\sigma^2\right) - \frac{1}{2} = -H(p),$$

и теперь дивергенция Кульбака–Лейблера равна

$$0 \leq \mathrm{KL}(q \| p) = -H(q) + H(p), \quad \text{то есть} \quad H(p) \geq H(q)$$

- В итоге мы получили, что для любого распределения $q(x)$ с заданной дисперсией σ^2 его энтропия $H(q)$ не превышает энтропии нормального распределения $H(p)$, что и требовалось
- Это уже полноценное формальное доказательство, и оно вовсе не апеллирует к вариационному исчислению

Априорные распределения Джеффриса

Априорные распределения Джеффриса

- Мы уже говорили о неинформативных априорных распределениях
- Равномерное распределение для шаров, которые достают из урны, оказалось не таким уж неинформативным: оно на самом деле соответствует данным $D = (2, 1)$, в которых мы уже достали два шара и увидели, что ровно один из них красный
- А равномерное распределение на параметр монетки θ можно рассматривать как результат «виртуального эксперимента» с двумя подбрасываниями; как обобщить эти рассуждения?
- Хотелось бы получить *объективные* априорные распределения, т.е. добиться того, чтобы два субъекта с одной и той же априорной информацией задали одни и те же априорные распределения
- Для этого нужно придумать целевую функцию, условие, которому должны априорные распределения удовлетворять

Априорные распределения Джеффриса

- Вернёмся для начала к монетке; равномерное распределение $p(\theta) = \mathrm{Unif}([0, 1])$ мы мотивировали тем, что «не предпочитаем никакого значения θ другому»
- Но это не единственный способ параметризовать монетку! Давайте, к примеру, перейдём к логарифму шансов (log-odds)

$$\eta = \log \frac{\theta}{1 - \theta}, \quad \text{то есть} \quad \theta = \frac{1}{1 + e^{-\eta}}$$

- Логарифм шансов параметризует монетку величиной, изменяющейся от $-\infty$ до ∞, и мы, по идее, не предпочитаем и никакого значения η другому
- Но если сделать в $p_\theta(\theta) = \mathrm{Unif}([0, 1])$ замену переменных $\theta = \frac{1}{1+e^{-\eta}}$, получится вовсе не равномерное распределение:

$$p_\eta(\eta) = p_\theta\left(\frac{1}{1 + e^{-\eta}}\right)\left(\frac{1}{1 + e^{-\eta}}\right)' = \frac{e^\eta}{(1 + e^\eta)^2}$$

Априорные распределения Джеффриса

- Получается, что если мы решаем «не предпочитать» разные значения θ, то мы тем самым очень даже предпочитаем разные значения η; и наоборот
- Равномерное распределение «ничего не предпочитает» только для конкретной параметризации — а это ведь одна и та же монетка, и у неё один и тот же параметр, как его ни преобразуй; по крайней мере, монотонные преобразования вроде перехода от θ к η и наоборот уж точно должны быть разрешены
- Долгое время это считалось серьёзным аргументом против байесовского подхода к статистике в целом; сам сэр Рональд Фишер критиковал байесовский подход с этих позиций

Априорные распределения Джеффриса

- Решение предложил сэр Гарольд Джеффрис: давайте считать это свойство требованием к «неинформативному априорному распределению»!
- Рассмотрим модель машинного обучения с правдоподобием $p_\theta(\mathbf{x}|\theta)$ (пусть пока один параметр $\theta \in \Theta$) и сделаем репараметризацию: применим к параметру θ произвольное гладкое монотонное преобразование $\eta = h(\theta)$
- Тогда заменой переменных мы получим новую модель

$$p_\eta(\mathbf{x}|\eta) = p_\theta\left(D|h^{-1}(\eta)\right)$$

Априорные распределения Джеффриса

- Предположим, что мы так или иначе придумали априорное распределение $p_\theta(\theta)$ для исходной модели. Тогда у нас есть два способа задать априорное распределение после репараметризации:
  - напрямую применить тот же принцип и взять то же распределение $p_\eta(\eta)$, но с параметром η;
  - сделать замену переменных $\eta = h(\theta)$ в априорном распределении $p_\theta(\theta)$ и получить распределение

$$\tilde{p}_\eta(\eta) = \frac{p_\theta\left(h^{-1}(\eta)\right)}{\left|h'\left(h^{-1}(\eta)\right)\right|}.$$

- Джеффрис предложил потребовать, чтобы для любой гладкой монотонной функции h эти два распределения совпадали: $p_\eta(\eta) = \tilde{p}_\eta(\eta)$, или, что то же самое,

$$\tilde{p}_\eta(h(\theta)) = \frac{p_\theta(\theta)}{\left|h'(\theta)\right|} \quad \text{для всех} \quad \theta \in \Theta$$

Априорные распределения Джеффриса

- Джеффрис дал и конструктивный ответ на вопрос, как построить такое априорное распределение
- В случае $\theta \in \mathbb{R}$ *априорное распределение Джеффриса* (Jeffreys prior) для модели $p_\theta\left(\mathbf{x}|\theta\right)$ определяется как

$$p^J\left(\theta\right) = \text{const} \cdot \sqrt{\mathcal{I}(\theta)}$$

Априорные распределения Джеффриса

- Проверим нужное свойство, записав определение $\mathcal{J}(\theta)$ и продифференцировав в этом определении $\log p_\eta\left(\mathbf{x}|h(\theta)\right)$ как сложную функцию:

$$\begin{aligned}
\mathcal{J}_\theta(\theta) &= \int \left(\frac{\partial}{\partial\theta}\log p_\theta\left(\mathbf{x}|\theta\right)\right)^2 p_\theta\left(\mathbf{x}|\theta\right)\mathrm{d}\mathbf{x} \\
&= \int \left(\frac{\partial}{\partial\theta}\log p_\eta\left(\mathbf{x}|h(\theta)\right)\right)^2 p_\eta\left(\mathbf{x}|h(\theta)\right)\mathrm{d}\mathbf{x} \\
&= \int \left(h'(\theta)\frac{\partial}{\partial\eta}\log p_\eta\left(\mathbf{x}|\eta\right)\right)^2 p_\eta\left(\mathbf{x}|h(\theta)\right)\mathrm{d}\mathbf{x} \\
&= h'(\theta)^2\mathcal{J}_\eta(h(\theta)), \qquad \text{а значит,}
\end{aligned}$$

$$p_\eta^J\left(h(\theta)\right) = \mathrm{const}\cdot\sqrt{\mathcal{J}_\eta(h(\theta))} = \mathrm{const}\cdot\frac{\mathcal{J}_\theta(\theta)}{|h'(\theta)|} = \frac{p_\theta^J\left(\theta\right)}{|h'(\theta)|},$$

что и требовалось

Априорные распределения Джеффриса

- Если интеграл $\int_\Theta \sqrt{\mathcal{I}(\theta)}$ конечен, то он даст нормировочную константу для плотности, и всё будет хорошо
- Но во многих практических случаях этот интеграл будет расходиться, и распределение Джеффриса будет задавать так называемое *некорректное априорное распределение* (improper prior), то есть такое априорное распределение, которое само по себе не является распределением вероятностей, но при подстановке в формулу Байеса даёт для любого непустого набора данных полноценное апостериорное распределение
- Грубо говоря, в таких случаях мы не будем обращать внимания на то, что интеграл расходится, а будем просто умножать правдоподобие данных на функцию $\sqrt{\mathcal{I}(\theta)}$ и нормировать результат — и вот у этого результата уже действительно интеграл обязательно должен будет сходиться

Априорные распределения Джеффриса

- Например, для бросания монетки модель в терминах θ выглядит как $p\left(x|\theta\right) = \theta^{[x=1]}(1-\theta)^{[x=0]}$ для данных, где $x = 1$ означает выпадение орла, а $x = 0$ — решки
- Априорное распределение Джеффриса можно найти как

$$p^J\left(\theta\right) \propto \sqrt{\mathcal{J}(\theta)} = \sqrt{\mathbb{E}\left[\left(\frac{\partial}{\partial\theta}\log p_\theta\left(D|\theta\right)\right)^2\right]} =$$

$$\sqrt{\mathbb{E}\left[\left(\frac{[x=1]}{\theta} - \frac{[x=0]}{1-\theta}\right)^2\right]} =$$

$$= \sqrt{\theta\left(\frac{1}{\theta}\right)^2 + (1-\theta)\left(\frac{1}{1-\theta}\right)^2} = \frac{1}{\sqrt{\theta(1-\theta)}}$$

Априорные распределения Джеффриса

- Получилось бета-распределение с дробными параметрами $\alpha = \beta = \frac{1}{2}$:

$$p^J(\theta) = \mathrm{Beta}\left(\theta \middle| \frac{1}{2}, \frac{1}{2}\right) \propto \theta^{-\frac{1}{2}} (1-\theta)^{-\frac{1}{2}}$$

- Аналогично, априорное распределение Джеффриса для игральной кости с k гранями – это распределение Дирихле с дробными параметрами $\alpha = \left(\frac{1}{2} \ \cdots \ \frac{1}{2} \right)$

Априорные распределения Джеффриса

- Другой пример — одномерное нормальное распределение $N\left(x\middle|\mu,\sigma^2\right)$, где σ^2 фиксирована, а вывод ведётся по μ
- Тогда информация Фишера для одной точки данных x равна

$$\mathcal{J}(\mu)=\mathbb{E}_{x|\mu}\left[\left(\frac{\partial}{\partial\mu}\log p\left(x|\mu\right)\right)^2\right]=\mathbb{E}_{x|\mu}\left[\left(\frac{x-\mu}{\sigma^2}\right)^2\right]=$$

$$=\frac{1}{\sigma^4}\mathbb{E}_{x|\mu}\left[(x-\mu)^2\right]=\frac{\sigma^2}{\sigma^4}=\frac{1}{\sigma^2};$$

а значит, информация Фишера для $D=\{x_1,\ldots,x_N\}$ равна N/σ^2, и априорное распределение Джеффриса получается пропорциональным $p^J\left(\mu\right)\propto\sqrt{N/\sigma^2}$

Априорные распределения Джеффриса

- Но σ^2 — константа, так что распределение Джеффриса просто равно одной и той же константе для всех μ
- Это характерный пример некорректного априорного распределения: нельзя взять равномерное распределение на всей прямой $\theta \in (-\infty, \infty)$, но для байесовского вывода это не страшно, потому что при умножении на правдоподобие пусть даже одной точки будет получаться суммируемая функция

Априорные распределения Джеффриса

- В многомерном случае распределение Джеффриса обобщается на векторный параметр $\theta \in \mathbb{R}^d$ так:

$$p^J(\theta) = \text{const} \cdot \sqrt{\det \mathcal{I}(\theta)}$$

- Базовое свойство проверяется так же, в многомерном случае замена переменных добавляет определитель матрицы частных производных:

$$\begin{aligned}
p^J(\eta) &= p^J(\theta) \left| \det \frac{\partial \theta_i}{\partial \eta_j} \right| \propto \sqrt{\det \mathcal{I}(\theta) \left(\det \frac{\partial \theta_i}{\partial \eta_j} \right)^2} \\
&= \sqrt{\det \frac{\partial \theta_k}{\partial \eta_i} \det \mathbb{E}\left[\frac{\partial \log p_\theta(\mathbf{x}|\theta)}{\partial \theta_k} \frac{\partial \log p_\theta(\mathbf{x}|\theta)}{\partial \theta_l} \right] \det \frac{\partial \theta_l}{\partial \eta_j}} \\
&= \sqrt{\det \mathbb{E}\left[\sum_{k,l} \frac{\partial \theta_k}{\partial \eta_i} \frac{\partial \log p_\theta(\mathbf{x}|\theta)}{\partial \theta_k} \frac{\partial \log p_\theta(\mathbf{x}|\theta)}{\partial \theta_l} \frac{\partial \theta_l}{\partial \eta_j} \right]} \\
&= \sqrt{\det \mathbb{E}\left[\frac{\partial \log p_\eta(\mathbf{x}|\eta)}{\partial \eta_i} \frac{\partial \log p_\eta(\mathbf{x}|\eta)}{\partial \eta_j} \right]} = \sqrt{\det \mathcal{I}(\eta)}.
\end{aligned}$$

Априорные распределения Джеффриса

- Для примера давайте найдём распределение Джеффриса $p^J(\theta) \propto \sqrt{\det \mathcal{I}(\theta)}$ для одномерного гауссиана с двумя параметрами
- Мы находимся в двумерном случае, где $\theta = \begin{pmatrix} \mu & \sigma^2 \end{pmatrix}^\top$, то есть матрица информации Фишера задаётся как

$$\mathcal{I}(\mu, \sigma^2) = -\mathbb{E}_{D|\mu,\sigma^2}\left[\begin{pmatrix} \frac{\partial^2}{\partial \mu^2}\log p\left(D\middle|\mu,\sigma^2\right) & \frac{\partial^2}{\partial \mu \partial \sigma^2}\log p\left(D\middle|\mu,\sigma^2\right) \\ \frac{\partial^2}{\partial \sigma^2 \partial \mu}\log p\left(D\middle|\mu,\sigma^2\right) & \frac{\partial^2}{\partial (\sigma^2)^2}\log p\left(D\middle|\mu,\sigma^2\right) \end{pmatrix}\right].$$

Априорные распределения Джеффриса

- Для гауссиана все эти производные легко подсчитать; для $D = \{x_1, \dots, x_n\}$ сначала перейдём к достаточным статистикам $\bar{x} = \frac{1}{n}\sum_i x_i$, $s^2 = \frac{1}{n-1}\sum_i (x_i - \bar{x})^2$:

$$\log p\left(D\middle|\mu, \sigma^2\right) = \text{const} - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left((n-1)s^2 + n(\bar{x}-\mu)^2\right),$$

а затем возьмём производные:

$$\frac{\partial \log p}{\partial \mu} = \frac{n(\bar{x}-\mu)}{\sigma^2}, \quad \frac{\partial \log p}{\partial \sigma^2} = -\frac{n}{2\sigma^2} - \frac{(n-1)s^2 + n(\bar{x}-\mu)^2}{4\sigma^4},$$

а значит,

$$\frac{\partial^2}{\partial \mu^2}\log p\left(D|\mu,\sigma^2\right) = -\frac{n}{\sigma^2},$$
$$\frac{\partial^2}{\partial \mu \partial \sigma^2}\log p\left(D|\mu,\sigma^2\right) = \frac{\partial^2}{\partial \sigma^2 \partial \mu}\log p\left(D|\mu,\sigma^2\right) = -\frac{n(\bar{x}-\mu)}{\sigma^4},$$
$$\frac{\partial^2}{\partial (\sigma^2)^2}\log p\left(D|\mu,\sigma^2\right) = \frac{n}{2\sigma^4} - \frac{(n-1)s^2 + n(\bar{x}-\mu)^2}{\sigma^2}.$$

Априорные распределения Джеффриса

- Теперь нужно взять ожидания; поскольку $\mathbb{E}[\bar{x}] = \mu$, $\mathbb{E}\left[(\bar{x}-\mu)^2\right] = \sigma^2/n$ и $\mathbb{E}\left[\sum_i (x_i - \bar{x})^2\right] = (n-1)\sigma^2$, матрица информации Фишера и её определитель равны

$$\mathcal{I}(\mu, \sigma^2) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}, \qquad \det \mathcal{I}(\mu, \sigma^2) = \frac{n^2}{2\sigma^6},$$

  и априорное распределение Джеффриса будет

$$p^J(\mu, \sigma^2) \propto \sqrt{\det \mathcal{I}(\theta)} \propto (\sigma^2)^{-3/2}$$

- Обратите внимание, что это распределение совсем не содержит μ — по μ опять получается некорректное равномерное априорное распределение на всей прямой
- А вот по σ^2 это полноценное распределение, интеграл по $(0, \infty)$ сходится, и нормировочную константу можно подсчитать

Априорные распределения Джеффриса

- Из априорного распределения Джеффриса получается такое апостериорное распределение:

$$p\left(\mu,\sigma^2\middle|x_1,\dots,x_n\right)\propto\left(\sigma^2\right)^{-3/2}\cdot\left(\sigma^2\right)^{-n/2}\cdot e^{-\frac{1}{2\sigma^2}\left((n-1)s^2+n(\bar{x}-\mu)^2\right)}=$$
$$=\left(\sigma^2\right)^{-(n+3)/2}e^{-\frac{1}{2\sigma^2}\left((n-1)s^2+n(\bar{x}-\mu)^2\right)}=$$
$$=\left(\frac{1}{\sqrt{\sigma^2}}e^{-\frac{(\bar{x}-\mu)^2}{2(\sigma/\sqrt{n})^2}}\right)\cdot\left(\left(\sigma^2\right)^{-(n+1)/2}e^{-\frac{(n-1)s^2}{2\sigma^2}}\right)$$

- Первый сомножитель — это (с точностью до константы) нормальное распределение на μ со средним $\bar{x}$ и дисперсией $(\sigma/\sqrt{n})^2$
- Получился вполне логичный результат: среднее постепенно уточняется с ростом n, то есть при получении новых данных

Спасибо!

Спасибо за внимание!