АПРИОРНЫЕ РАСПРЕДЕЛЕНИЯ ДЖЕФФРИСА

Сергей Николенко

СПбГУ — Санкт-Петербург

28 ноября 2024 г.

Random facts:

- 28 ноября 1443 г. Георгий Кастриоти, он же Скандербег, торжественно вступил в Крую и был провозглашён старшинами главой княжества Кастриоти и вождём всех албанцев
- 28 ноября 1660 г. двенадцать человек, в том числе Роберт Бойль, Кристофер Рен и сэр Роберт Морэй собрались в Грешем-колледже, первом высшем учебном заведении Лондона, и основали «Colledge for the Promoting of Physico-Mathematicall Experimentall Learning», который потом превратится в the Royal Society
- 28 ноября 1814 г. лондонская The Times стала первой газетой, производившейся на паровом печатном прессе, созданном немецкими инженерами Кёнигом и Бауэром
- 28 ноября 1918 г. в Эстонию вошли российские войска, части РККА заняли Нарву, и началась Освободительная Война (1918—1920)
- 28 ноября 1905 г. Артур Гриффит основал политическую партию под названием Sinn Féin
- 28 ноября 1990 г. Маргарет Тэтчер ушла в отставку как глава консервативной партии и, как следствие, как премьер-министр; её преемником на обеих позициях стал Джон Мейджор

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 $\mathcal{I}(\theta)$ и продифференцировав в этом определении $\log p_\eta\left(\mathbf{x}|h(\theta)\right)$ как сложную функцию:

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

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

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

Экспоненциальное семейство

Экспоненциальное семейство

- Мы видели общий паттерн: найти правдоподобие, посмотреть на его форму и догадаться, как должно выглядеть семейство сопряжённых априорных распределений.
- Это выглядит как достаточно несложная процедура, которая должна обобщаться.
- *Экспоненциальное семейство* распределений (exponential family): параметрическое семейство распределений принадлежит экспоненциальному семейству, если оно имеет вид

$$p\left(\mathbf{x}|\theta\right) = h(\mathbf{x})e^{\eta(\theta)^\top \mathbf{t}(\mathbf{x}) - a(\theta)} = h(\mathbf{x})g(\theta)e^{\eta(\theta)^\top \mathbf{t}(\mathbf{x})}$$

  для некоторого параметра θ; здесь $g(\theta) = e^{-a(\theta)}$.
- Векторная функция $\mathbf{t}(\mathbf{x})$ выделяет *достаточные статистики* (sufficient statistics), и она играет роль извлечения признаков из $\mathbf{x}$.

Экспоненциальное семейство

- Если $\eta(\theta) = \theta$, то такая параметризация называется *естественной*, а θ в таком случае называется *естественным параметром* (natural parameter):

$$p(\mathbf{x}|\theta) = h(\mathbf{x})e^{\theta^\top \mathbf{t}(\mathbf{x}) - a(\theta)} = h(\mathbf{x})g(\theta)e^{\theta^\top \mathbf{t}(\mathbf{x})}.$$

- Определение выглядит очень общим; главное предположение здесь в том, как θ и $\mathbf{x}$ разделяются в этом определении: в экспоненте они связаны друг с другом линейно, а вне экспоненты полностью разнесены по функциям $h(\mathbf{x})$ и $g(\theta)$, то есть единственная зависимость между $\mathbf{x}$ и θ — это скалярное произведение в экспоненте.
- Вообще говоря, почти всё, о чём мы говорили – частные случаи *экспоненциального семейства* распределений.

Экспоненциальное семейство

- Например, биномиальное распределение

$$\mathrm{Binom}\,(k|n,p) = \binom{n}{k} p^k\,(1-p)^{n-k} = \\ = \binom{n}{k} e^{k\log p + (n-k)\log(1-p)} = \binom{n}{k} e^{k\log\frac{p}{1-p} + n\log(1-p)}.$$

- В итоге получается, что биномиальное распределение принадлежит экспоненциальному семейству, и его естественный параметр — это

$$\theta = \log\frac{p}{1-p}, \qquad p = \frac{e^\theta}{1+e^\theta},$$

то есть в точности те самые log-odds; $t(k) = k$, $h(k) = \binom{n}{k}$,

$$a(\theta) = -n\log(1-p) = n\log(1+e^\theta), \quad g(\theta) = e^{n\log(1-p)} = (1+e^\theta)^{-n}.$$

Экспоненциальное семейство

- Аналогично, мультиномиальное распределение

$$\mathrm{Mult}\,(\mathbf{x}|n,p_1,\dots,p_k) = \begin{cases} \frac{n!}{x_1!x_2!\dots x_k!}p_1^{x_1}p_2^{x_2}\dots p_k^{x_k}, & \text{если } \sum_{i=1}^k x_i = n, \\ 0 & \text{в противном случае,} \end{cases}$$

можно переписать как

$$\mathrm{Mult}\,(\mathbf{x}|n,p_1,p_2,\dots,p_k) = \frac{n!}{x_1!x_2!\dots x_k!}e^{\sum_{i=1}^k x_i \log p_i},$$

то есть на первый взгляд кажется, что в экспоненциальном семействе здесь

$$\mathbf{t}(\mathbf{x}) = \mathbf{x}, \quad \theta = \log \mathbf{p}, \quad a(\theta) = 0, \quad h(\mathbf{x}) = \frac{n!}{x_1!x_2!\dots x_k!}.$$

Экспоненциальное семейство

- Но такое представление ведёт к техническим трудностям из-за того, что $a(\theta) = 0$, поэтому лучше выразить

$$e^{\sum_{i=1}^{k} x_i \log p_i} = e^{\sum_{i=1}^{k-1} x_i \log p_i + \left(n - \sum_{i=1}^{k-1} x_i\right) \log\left(1 - \sum_{i=1}^{k-1} p_i\right)} = \\ = e^{\sum_{i=1}^{k-1} x_i \log\left(\frac{p_i}{1 - \sum_{i=1}^{k-1} p_i}\right) + n \log\left(1 - \sum_{i=1}^{k-1} p_i\right)}.$$

- Таким образом, в итоге $\mathbf{t}(\mathbf{x}) = \mathbf{x}$,

$$\theta_i = \log\left(\frac{p_i}{1 - \sum_{i=1}^{k-1} p_i}\right) = \log\frac{p_i}{p_k}, \quad p_i = \frac{e^{\theta_i}}{\sum_{j=1}^{k} e^{\theta_j}},$$

  и теперь $a(\theta) = -n \log\left(1 - \sum_{i=1}^{k-1} p_i\right) = n \log\left(\sum_{j=1}^{k} e^{\theta_j}\right)$.
- В обратном выражении для p_i через θ у нас опять получилась как раз та самая softmax-функция.

Экспоненциальное семейство

- С распределением Пуассона совсем нет вопросов:

$$p\left(x|\lambda\right) = \frac{1}{x!}\lambda^x e^{-\lambda} = \frac{1}{x!}e^{x\log\lambda - \lambda}$$

  сразу же принадлежит экспоненциальному семейству с $t(x) = x$, $\theta = \log\lambda$, $h(x) = \frac{1}{x!}$, $a(\theta) = \lambda = e^\theta$.

- Редкий пример распределения, которое *не* принадлежит экспоненциальному семейству — это гипергеометрическое распределение

$$p\left(x|N,n,K\right) = \frac{1}{\binom{N}{n}}\binom{K}{x}\binom{N-K}{n-x};$$

  его преобразовать к нужной форме никак не получится.

Спасибо за внимание!