# ЭКСПОНЕНЦИАЛЬНОЕ СЕМЕЙСТВО

Сергей Николенко

СПбГУ — Санкт-Петербург

22 ноября 2022 г.

---

*Random facts*:

- 22 ноября 1307 г. Климент V издал буллу «Pastoralis Praeminentiae», в которой утверждал, что обвинения против ордена тамплиеров доказаны, и призывал государей Европы последовать примеру Филиппа Красивого
- 22 ноября 1604 г. король Испании Филипп III отдыхал у камина, и пламя сильно разгорелось; пока специального дворянина, имевшего право затворить заслонку или отодвинуть кресло короля, искали по всему дворцу, Филипп ждал не двигаясь с места и получил серьёзные ожоги (это, конечно, анекдот)
- 22 ноября 1621 г. настоятелем лондонского собора св. Павла был назначен Джон Донн; с тех пор он мог сам выбирать, по ком звонит колокол
- 22 ноября 1641 г. английский парламент принял «Великую ремонстрацию» (The Grand Remonstrance) — перечень из 204 статей, исчисляющих злоупотребления короля
- 22 ноября 1977 г. начались полёты «Конкорда» через Атлантику: из Лондона в Нью-Йорк можно было добраться менее чем за три с половиной часа, билет «туда-обратно» стоил $10500

# 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

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

  для некоторого параметра $\theta$; 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 $\theta$ 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 $\theta$ 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 $\theta$ — 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 $\theta$ 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.

# Экспоненциальное семейство

- Нормальное распределение:

$$N\left(x\middle|\mu,\sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} =$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{\begin{pmatrix} x^2 \\ x \end{pmatrix}^\top \begin{pmatrix} -1/2\sigma^2 \\ \mu/\sigma^2 \end{pmatrix} - \frac{\mu^2}{2\sigma^2} - \log\sigma}.$$

- Иначе говоря, одномерное нормальное распределение имеет две достаточные статистики, $\mathbf{t}(x) = \begin{pmatrix} x^2 \\ x \end{pmatrix}$, и естественный параметр размерности два:

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} -1/2\sigma^2 \\ \mu/\sigma^2 \end{pmatrix} = \begin{pmatrix} -\tau/2 \\ \mu\tau \end{pmatrix};$$

а остальные функции выглядят как $h(x) = \frac{1}{\sqrt{2\pi}}$,

$$a(\theta) = \frac{\mu^2}{2\sigma^2} + \log\sigma = \frac{\mu^2\tau}{2} - \frac{1}{2}\log\tau = -\frac{\theta_2^2}{4\theta_1^2} - \frac{1}{2}\log(-2\theta_1).$$

# Экспоненциальное семейство

- Многомерный гауссиан:

$$N\left(\mathbf{x}|\mu,\Sigma\right)=\frac{1}{\sqrt{2\pi\det\Sigma}}e^{-\frac{1}{2}\left(\mathbf{x}-\mu\right)^{\top}\Sigma^{-1}\left(\mathbf{x}-\mu\right)}=$$
$$=e^{-\frac{1}{2}\left(\mathbf{x}^{\top}\Sigma^{-1}\mathbf{x}-2\mathbf{x}^{\top}\Sigma^{-1}\mu+\mu^{\top}\Sigma^{-1}\mu+\log\left(2\pi\det\Sigma\right)\right)}.$$

- Нужно представить $\mathbf{x}^{\top}\Sigma^{-1}\mathbf{x}$ в виде скалярного произведения; здесь $\mathrm{vec}\left(A\right)$ обозначает разворачивание матрицы в плоский вектор:

$$\mathbf{x}^{\top}\Sigma^{-1}\mathbf{x}=\sum_{i,j=1}^{d}\left(\Sigma^{-1}\right)_{ij}x_{i}x_{j}=\mathrm{vec}\left(\mathbf{x}\mathbf{x}^{\top}\right)^{\top}\mathrm{vec}\left(\Sigma^{-1}\right).$$

- В итоге $h(\mathbf{x})=1$, $\mathbf{t}(\mathbf{x})=\begin{pmatrix}\mathrm{vec}\left(\mathbf{x}\mathbf{x}^{\top}\right)\\ \mathbf{x}\end{pmatrix}$, $\theta=\begin{pmatrix}-\frac{1}{2}\mathrm{vec}\left(\Sigma^{-1}\right)\\ \Sigma^{-1}\mu\end{pmatrix}$,

$$a(\theta)=\mu^{\top}\Sigma^{-1}\mu+\log\left(2\pi\det\Sigma\right).$$

# Экспоненциальное семейство

- Теперь интересные результаты. Первый — о среднем и дисперсии распределений из экспоненциального семейства.
- Интеграл от любого распределения равен единице:

$$a(\theta) = \log \int h(\mathbf{x}) e^{\theta^\top \mathbf{t}(\mathbf{x})} \mathrm{d}\mathbf{x}.$$

- Возьмём градиент по $\theta$ слева и справа:

$$\nabla_\theta a(\theta) = \nabla_\theta \log \int h(\mathbf{x}) e^{\theta^\top \mathbf{t}(\mathbf{x})} \mathrm{d}\mathbf{x} = \frac{\int \nabla_\theta h(\mathbf{x}) e^{\theta^\top \mathbf{t}(\mathbf{x})} \mathrm{d}\mathbf{x}}{\int h(\mathbf{x}) e^{\theta^\top \mathbf{t}(\mathbf{x})} \mathrm{d}\mathbf{x}} =$$

$$= \frac{\int h(\mathbf{x}) \nabla_\theta e^{\theta^\top \mathbf{t}(\mathbf{x})} \mathrm{d}\mathbf{x}}{e^{a(\theta)}} = \frac{\int h(\mathbf{x}) e^{\theta^\top \mathbf{t}(\mathbf{x})} \mathbf{t}(\mathbf{x}) \mathrm{d}\mathbf{x}}{e^{a(\theta)}} =$$

$$= \int \mathbf{t}(\mathbf{x}) \left( h(\mathbf{x}) e^{\theta^\top \mathbf{t}(\mathbf{x}) - a(\theta)} \right) \mathrm{d}\mathbf{x}.$$

# Экспоненциальное семейство

- Иначе говоря, мы видим, что $\nabla_\theta a(\theta)$ — это математическое ожидание достаточных статистик исходного распределения:

$$\mathbb{E}\left[\mathbf{t}(\mathbf{x})\right] = \nabla_\theta a(\theta).$$

- Это очень мощный результат, который нередко пригождается и в машинном обучении.
- Функцию $a(\theta)$ называют *кумулянтом* (cumulant).

# Экспоненциальное семейство

- Кроме того, для минимальных представлений распределений из экспоненциального семейства, когда в $\theta$ нет лишних параметров, это можно обратить, то есть выразить $\theta$ через $\mathbb{E}\left[\mathbf{t}(\mathbf{x})\right]$.
- Здесь будет естественно рассмотреть *параметризацию средним* (mean parametrization), в которой параметром будет

$$\mu = \mathbb{E}\left[\mathbf{t}(\mathbf{x})\right] = \nabla_\theta a(\theta).$$

- Например, привычная нам параметризация гауссиана именно такова.

# Экспоненциальное семейство

- Результат можно продолжить на другие моменты.
- Рассмотрим матрицу вторых производных $a(\theta)$:

$$\frac{\partial^2 a}{\partial\theta_i\partial\theta_j} = \frac{\partial\mathbb{E}\left[t_i(\mathbf{x})\right]}{\partial\theta_j} = \frac{\partial}{\partial\theta_j}\int t_i(\mathbf{x})h(\mathbf{x})e^{\theta^\top\mathbf{t}(\mathbf{x})-a(\theta)}\mathrm{d}\mathbf{x} =$$

$$= \int t_i(\mathbf{x})h(\mathbf{x})\frac{\partial e^{\theta^\top\mathbf{t}(\mathbf{x})-a(\theta)}}{\partial\theta_j}\mathrm{d}\mathbf{x} = \int t_i(\mathbf{x})h(\mathbf{x})e^{\theta^\top\mathbf{t}(\mathbf{x})-a(\theta)}\left(t_j(\mathbf{x}) - \frac{\partial a(\theta)}{\partial\theta_j}\right)\mathrm{d}\mathbf{x} =$$

$$= \int t_i(\mathbf{x})\left(t_j(\mathbf{x}) - \mathbb{E}\left[t_j(\mathbf{x})\right]\right)h(\mathbf{x})e^{\theta^\top\mathbf{t}(\mathbf{x})-a(\theta)}\mathrm{d}\mathbf{x} =$$

$$= \mathbb{E}\left[t_i(\mathbf{x})t_j(\mathbf{x})\right] - \mathbb{E}\left[t_i(\mathbf{x})\right]\mathbb{E}\left[t_j(\mathbf{x})\right].$$

- Иначе говоря, гессиан функции $a(\theta)$ — это в точности матрица ковариаций вектора достаточных статистик $\mathbf{t}(\mathbf{x})$:

$$\mathrm{Var}\left[\mathbf{t}(\mathbf{x})\right] = \mathbf{H}\left(a(\theta)\right) = \left(\frac{\partial^2 a}{\partial\theta_i\partial\theta_j}\right)_{i,j=1}^{k}.$$

- Аналогичные результаты верны и для других моментов.

# Экспоненциальное семейство

- Второй результат – о правдоподобии:

$$p\left(\mathbf{x}_1, \dots, \mathbf{x}_N | \theta\right) = \prod_{n=1}^{N} h(\mathbf{x}_n) e^{\eta(\theta)^\top \mathbf{t}(\mathbf{x}_n) - a(\theta)},$$

$$\log p\left(\mathbf{x}_1, \dots, \mathbf{x}_N | \theta\right) = \sum_{n=1}^{N} \log h(\mathbf{x}_n) + \eta(\theta)^\top \left( \sum_{n=1}^{N} \mathbf{t}(\mathbf{x}_n) \right) - N a(\theta).$$

- Всё, что в логарифме правдоподобия зависит от $\theta$, содержит $\mathbf{x}_n$ только в виде $\sum_{n=1}^{N} \mathbf{t}(\mathbf{x}_n)$.
- Т.е. достаточно сохранять суммы $\sum_{n=1}^{N} \mathbf{t}(\mathbf{x}_n)$ и число $N$, а сами точки $\mathbf{x}_n$ можно «забывать»; поэтому $\mathbf{t}(\mathbf{x}_n)$ называются *достаточными статистиками*.
- Например, для одномерного гауссиана достаточно хранить $\sum_{n=1}^{N} x_n$ и $\sum_{n=1}^{N} x_n^2$, и из них можно найти гипотезу максимального правдоподобия и для $\mu$, и для $\sigma^2$.

# Экспоненциальное семейство

- И саму гипотезу максимального правдоподобия можно найти, взяв градиент по $\theta$ и приравняв нулю:

$$\nabla_\theta a(\theta) = \frac{1}{N} \nabla_\theta \eta(\theta)^\top \left( \sum_{n=1}^{N} \mathbf{t}(\mathbf{x}_n) \right).$$

- Для естественного параметра $\eta(\theta) = \theta$ правая часть не зависит от $\theta$. А при параметризации средним, $\eta(\theta) = \theta$, сразу

$$\mu = \nabla_\theta a(\theta) = \frac{1}{N} \sum_{n=1}^{N} \mathbf{t}(\mathbf{x}_n).$$

- *Теорема Купмана–Питмана–Дармуа* (Pitman–Koopman–Darmois): при некоторых условиях регулярности экспоненциальное семейство — это *единственное* семейство распределений с конечным набором достаточных статистик, т.е. с набором достаточных статистик, размер которого не зависит от $N$.

# Экспоненциальное семейство

- Третий интересный результат – уже о байесовском выводе: оказывается, есть универсальный способ найти семейство сопряжённых априорных распределений для любого распределения из экспоненциального семейства.
- Для семейства распределений с параметром $\theta$ семейством сопряжённых априорных распределений будет

$$p\left(\theta|\chi,\nu\right) = f(\chi,\nu)e^{\chi^\top\eta(\theta)-\nu a(\theta)},$$

где $\chi$ и $\nu$ — гиперпараметры, функция $a(\theta)$ та же самая, что в исходном распределении, а функция $f(\chi,\nu)$ — это нормировочная константа.

# Экспоненциальное семейство

- Давайте найдём апостериорное распределение:

$$p\left(\theta\middle|\mathbf{x}_1,\dots,\mathbf{x}_N,\chi_0,\nu_0\right)\propto p\left(\theta\middle|\chi_0,\nu_0\right)p\left(\mathbf{x}_1,\dots,\mathbf{x}_N|\theta\right).$$

- Подставим:

$$\log p\left(\theta\middle|\mathbf{x}_1,\dots,\mathbf{x}_N,\chi_0,\nu_0\right)=\sum_{n=1}^{N}\log h(\mathbf{x}_n)+\eta(\theta)^\top\left(\sum_{n=1}^{N}\mathbf{t}(\mathbf{x}_n)\right)-Na(\theta)+$$
$$+\log f(\chi_0,\nu_0)+\chi_0^\top\eta(\theta)-\nu_0 a(\theta)+\text{const}=$$
$$=\text{const}+\left(\chi_0+\sum_{n=1}^{N}\mathbf{t}(\mathbf{x}_n)\right)^\top\eta(\theta)-\left(N+\nu_0\right)a(\theta),$$

и в итоге у нас получилось апостериорное распределение $p\left(\theta\middle|\chi_N,\nu_N\right)$ того же вида, но с параметрами

$$\nu_N=\nu_0+N,\qquad\chi_N=\chi_0+\sum_{n=1}^{N}\mathbf{t}(\mathbf{x}_n).$$

# Экспоненциальное семейство

- Снова видим, что для апостериорного распределения нужно знать только достаточные статистики $\sum_{n=1}^{N} \mathbf{t}(\mathbf{x}_n)$.
- Более того, можно и предсказательное распределение найти:

$$p\left(\mathbf{x}\middle|\mathbf{x}_1, \dots, \mathbf{x}_n, \chi_0, \nu_0\right) = \int p\left(\mathbf{x}|\theta\right) p\left(\theta\middle|\mathbf{x}_1, \dots, \mathbf{x}_N, \chi_0, \nu_0\right) \mathrm{d}\theta =$$

$$= \int h(\mathbf{x}) e^{\mathbf{t}(\mathbf{x})^\top \eta(\theta) - a(\theta)} f(\chi_N, \nu_N) e^{\chi_N^\top \eta(\theta) - \nu_N a(\theta)} \mathrm{d}\theta =$$

$$= f(\chi_N, \nu_N) \int h(\mathbf{x}) e^{\left(\mathbf{t}(\mathbf{x}) + \chi_N\right)^\top \eta(\theta) - (\nu_N + 1) a(\theta)} = \frac{f(\chi_N, \nu_N)}{f(\chi_N + \mathbf{t}(\mathbf{x}), \nu_N + 1)}.$$

- Получилось, что предсказательное распределение — это отношение нормировочных констант для сопряжённого априорного распределения с разными параметрами.
- Сравните это с выводом предсказательного распределения для испытаний Бернулли, который мы делали в начале курса.

# Спасибо!

Спасибо за внимание!