

# ЭКСПОНЕНЦИАЛЬНОЕ СЕМЕЙСТВО

Сергей Николенко

СПбГУ — Санкт-Петербург

04 декабря 2025 г.

*Random facts:*



- 4 декабря в России — День российской информатики, потому что 4 декабря 1948 г. в СССР за номером 10475 было зарегистрировано изобретение И.С. Брука и Б.И. Рамеева — цифровая электронная вычислительная машина
- 4 декабря 771 г. король франков Карломан умер в своём дворце Самусси, и Карл Великий стал единовластным правителем Франкского королевства
- 4 декабря 1783 г., через 9 дней после отплытия последнего корабля англичан, Джордж Вашингтон на ужине в таверне Френсиса (Fraunces Tavern) попрощался со своими офицерами, подал в отставку и удалился в своё поместье
- 4 декабря 1934 г. Леонид Николаев выстрелом из револьвера убил в Смольном Сергея Кирова; как это убийство было задумано и организовано, не ясно до сих пор
- 4 декабря 1956 г. на джем-сессию в Sun Studio в первый и последний раз собрался Million Dollar Quartet: Элвис Пресли, Джерри Ли Льюис, Карл Перкинс и Джонни Кэш
- 4 декабря 1971 г. во время концерта Фрэнка Заппы дотла сгорело многоэтажное казино в Монтрё; это событие увековечено в песне *Deep Purple* «Smoke on the Water»

# 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

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

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  $\mathbf{x}$ .

- 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.

- Например, биномиальное распределение

$$\begin{aligned}\text{Binom}(k|n, p) &= \binom{n}{k} p^k (1-p)^{n-k} = \\ &= \binom{n}{k} e^{k \log p + (n-k) \log(1-p)} = \binom{n}{k} e^{k \log \frac{p}{1-p} + n \log(1-p)}.\end{aligned}$$

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

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

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

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

- Аналогично, мультиномиальное распределение

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

можно переписать как

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

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

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

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

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

- 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.

- С распределением Пуассона совсем нет вопросов:

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

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

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

его преобразовать к нужной форме никак не получится.



- Нормальное распределение:

$$\begin{aligned} N(x|\mu, \sigma^2) &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} = \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{\mu^2}{2\sigma^2}} = \frac{1}{\sqrt{2\pi}} e^{\begin{pmatrix} x^2 \\ x \end{pmatrix}^\top \begin{pmatrix} -1/2\sigma^2 \\ \mu/\sigma^2 \end{pmatrix} - \frac{\mu^2}{2\sigma^2} - \log \sigma}. \end{aligned}$$

- 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  $h(x) = \frac{1}{\sqrt{2\pi}}$ ,

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

- Многомерный гауссиан:

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

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

$$\mathbf{x}^\top \Sigma^{-1} \mathbf{x} = \sum_{i,j=1}^d (\Sigma^{-1})_{ij} x_i x_j = \text{vec}(\mathbf{x}\mathbf{x}^\top)^\top \text{vec}(\Sigma^{-1}).$$

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

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

# ЭКСПОНЕНЦИАЛЬНОЕ СЕМЕЙСТВО

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

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

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

$$\begin{aligned} \nabla_\theta a(\theta) &= \nabla_\theta \log \int h(\mathbf{x}) e^{\theta^\top \mathbf{t}(\mathbf{x})} d\mathbf{x} = \frac{\int \nabla_\theta h(\mathbf{x}) e^{\theta^\top \mathbf{t}(\mathbf{x})} d\mathbf{x}}{\int h(\mathbf{x}) e^{\theta^\top \mathbf{t}(\mathbf{x})} d\mathbf{x}} = \\ &= \frac{\int h(\mathbf{x}) \nabla_\theta e^{\theta^\top \mathbf{t}(\mathbf{x})} d\mathbf{x}}{e^{a(\theta)}} = \frac{\int h(\mathbf{x}) e^{\theta^\top \mathbf{t}(\mathbf{x})} \mathbf{t}(\mathbf{x}) d\mathbf{x}}{e^{a(\theta)}} = \\ &= \int \mathbf{t}(\mathbf{x}) \left( h(\mathbf{x}) e^{\theta^\top \mathbf{t}(\mathbf{x}) - a(\theta)} \right) d\mathbf{x}. \end{aligned}$$

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

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

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

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

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

- Например, привычная нам параметризация гауссиана именно такова.

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

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

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

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

- Аналогичные результаты верны и для других моментов.

- Второй результат – о правдоподоби:

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

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

- 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  $\sigma^2$ .

- 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  $N$ .



- 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

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

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

- Давайте найдём апостериорное распределение:

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

- Подставим:

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

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

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

- 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:

$$\begin{aligned} p(\mathbf{x} | \mathbf{x}_1, \dots, \mathbf{x}_N, \chi_0, \nu_0) &= \int p(\mathbf{x} | \theta) p(\theta | \mathbf{x}_1, \dots, \mathbf{x}_N, \chi_0, \nu_0) d\theta = \\ &= \int h(\mathbf{x}) e^{\mathbf{t}(\mathbf{x})^\top \eta(\theta) - a(\theta)} f(\chi_N, \nu_N) e^{\chi_N^\top \eta(\theta) - \nu_N a(\theta)} d\theta = \\ &= f(\chi_N, \nu_N) \int h(\mathbf{x}) e^{(\mathbf{t}(\mathbf{x}) + \chi_N)^\top \eta(\theta) - (\nu_N + 1)a(\theta)} d\theta = \frac{f(\chi_N + \mathbf{t}(\mathbf{x}), \nu_N + 1)}{f(\chi_N, \nu_N)}. \end{aligned}$$

- 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.

# ОБОБЩЁННЫЕ ЛИНЕЙНЫЕ МОДЕЛИ

---

- Ранее мы обобщили многое из того, что узнали о разных распределениях и байесовском выводе для этих распределений в экспоненциальном семействе.
- Теперь давайте попробуем обобщить происходящее в линейных моделях для задач регрессии и классификации.
- И линейная, и логистическая регрессия имеют одну и ту же форму:

$$\hat{\mathbf{y}} = h(\mathbf{w}^\top \mathbf{x}),$$

только для разных функций  $h$ :  $h(a) = a$  для линейной регрессии и  $h(a) = \sigma(a)$  для логистической регрессии (в бинарном случае).

- Может быть, можно обобщить?

- *Обобщённые линейные модели* (generalized linear models, GLM): начнём с того, что определим линейную функцию от входов

$$c = \mathbf{w}^\top \mathbf{x},$$

а затем определим среднее интересующего нас распределения  $\mu$  как функцию от  $a$ .

- Обычно задают обратную к ней функцию

$$c = g(\mu), \quad \text{то есть} \quad \mu = g^{-1}(c);$$

- $g$  называется *функцией связи* (link function), и в качестве  $g$  можно выбрать практически любую обратимую функцию.

- Например, в линейной регрессии  $g = g^{-1} = \text{id}$ , а в бинарной логистической регрессии  $g^{-1}(a) = \sigma(a)$ , то есть  $g(\mu) = \log \frac{\mu}{1-\mu}$ .
- Далее нужно определить одномерное распределение  $p(y|\mathbf{x}, \mathbf{w})$  со средним  $\mu$ , которое бы использовало  $g^{-1}(\mathbf{w}^\top \mathbf{x})$  как достаточную статистику.
- Правда, мы хотим научиться контролировать не только среднее, но и дисперсию на выходе, поэтому вместо общего вида экспоненциального семейства будем рассматривать распределение вида

$$p(y|\theta, \sigma^2) = h(y, \sigma^2) e^{\frac{y\theta - a(\theta)}{\sigma^2}},$$

где  $\sigma^2$  — параметр дисперсии (dispersion parameter).

- Само это семейство иногда называют *дисперсным экспоненциальным семейством* (overdispersed exponential family); это не обобщение экспоненциального, а его частный случай, в котором  $t(y) = \frac{y}{\sigma^2}$ , а кумулянт равен  $\frac{1}{\sigma^2} a(\theta)$ .

- Тогда доказанный нами в предыдущей лекции результат  $\mathbb{E}[\mathbf{t}(\mathbf{x})] = \nabla_{\theta} a(\theta)$  сразу даёт среднее

$$\mathbb{E}\left[\frac{y}{\sigma^2}\right] = \frac{\partial\left(\frac{1}{\sigma^2}a(\theta)\right)}{\partial\theta}, \quad \text{то есть} \quad \mathbb{E}[y] = \frac{\partial a(\theta)}{\partial\theta}.$$

- А результат

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

даёт дисперсию

$$\text{Var}\left[\frac{y}{\sigma^2}\right] = \frac{1}{\sigma^4} \text{Var}[y] = \frac{\partial^2\left(\frac{1}{\sigma^2}a(\theta)\right)}{\partial\theta^2},$$

то есть

$$\text{Var}[y] = \sigma^2 \frac{\partial^2 a(\theta)}{\partial\theta^2}.$$



- Осталось только договориться, что такое здесь  $\theta$ ; обобщённые линейные модели потому и называются линейными, что  $\theta = \mathbf{w}^\top \mathbf{x}$ .
- Это значит, что функция связи здесь определяется как

$$g^{-1}(\mathbf{w}^\top \mathbf{x}) = \mu = \mathbb{E}[y \mid \mathbf{x}, \mathbf{w}] = a'(\mathbf{w}^\top \mathbf{x}), \quad g(\mu) = \mathbf{w}^\top \mathbf{x}.$$

- Так, для линейной регрессии

$$p(y \mid \mu, \sigma^2) = e^{-\frac{(y-\mu)^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2)} = e^{\frac{y\mu - \frac{1}{2}\mu^2}{\sigma^2} - \left(\frac{y^2}{2\sigma^2} + \frac{1}{2} \log(2\pi\sigma^2)\right)},$$

поэтому для линейной регрессии логично положить

$$g = g^{-1} = \text{id}, \quad \theta = \mu = \mathbf{w}^\top \mathbf{x},$$

$$a(\theta) = -\frac{1}{2}\mu^2, \quad h(y, \sigma^2) = e^{-\left(\frac{y^2}{2\sigma^2} + \frac{1}{2} \log(2\pi\sigma^2)\right)}.$$

- Второй пример — испытания Бернулли. Здесь среднее — это вероятность орла  $\mu = \mathbb{E}[y \mid \mathbf{x}]$ , и мы уже знаем, что

$$p(y|\mu) = \mu^y (1 - \mu)^{1-y} = e^{y \log \frac{\mu}{1-\mu} + \log(1-\mu)},$$

то есть в данном случае можно положить

$$\sigma^2 = 1, \quad \theta = \log \frac{\mu}{1-\mu}, \quad a(\theta) = -\log(1 - \mu), \quad h(y, \sigma^2) = 0.$$

- Получилось вложение логистической регрессии в обобщённые линейные модели:

$$\mu = \frac{e^\theta}{1 + e^\theta}, \quad a(\theta) = -\log \left( 1 - \frac{e^\theta}{1 + e^\theta} \right) = \log(1 + e^\theta), \quad \text{то есть}$$

$$g^{-1}(\theta) = a'(\theta) = \frac{e^\theta}{1 + e^\theta} = \sigma(\theta), \quad \text{и} \quad g(\mu) = \log \frac{\mu}{1 - \mu} = \sigma^{-1}(\mu).$$

- Но эти два примера мы знали и так; а что-нибудь новенькое?
- Но чтобы извлечь из примеров новых моделей что-то полезное, надо сначала научиться вести вывод.
- Это и в общем случае можно делать точно так же, как мы это делали для логистической регрессии; логарифм правдоподобия выглядит как

$$\log p(y|\mathbf{x}, \mathbf{w}, \sigma^2) = \log h(y, \sigma^2) + \frac{y\mathbf{w}^\top \mathbf{x} - a(\mathbf{w}^\top \mathbf{x})}{\sigma^2},$$

и, соответственно, логарифм правдоподобия набора данных  $D = \{(\mathbf{x}_n, y_n)\}_{n=1}^N$  равен

$$\begin{aligned}\ell(\mathbf{w}) &= \log p(y_1, \dots, y_N | \mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{w}, \sigma^2) = \\ &= \text{const} + \frac{1}{\sigma^2} \sum_{n=1}^N (y_n \mathbf{w}^\top \mathbf{x}_n - a(\mathbf{w}^\top \mathbf{x}_n)).\end{aligned}$$

- От него теперь можно взять градиент:

$$\nabla_{\mathbf{w}} \ell(\mathbf{w}) = \frac{1}{\sigma^2} \sum_{n=1}^N (y_n - a'(\mathbf{w}^\top \mathbf{x}_n)) \mathbf{x}_n = \frac{1}{\sigma^2} \sum_{n=1}^N (y_n - \mu_n) \mathbf{x}_n,$$

где мы подставили  $\mu_n = \mathbb{E}[y_n \mid \mathbf{x}_n, \mathbf{w}]$ .

- Иначе говоря, мы снова получили в градиенте сумму входных векторов  $\mathbf{x}_n$ , взвешенных их ошибками, то есть отклонениями от ожидаемого среднего.

- Результат можно использовать напрямую в (стохастическом) градиентном спуске, а можно, опять же в точности как в логистической регрессии, сделать следующий шаг и перейти к методу второго порядка:

$$\mathbf{H}(\ell) = -\frac{1}{\sigma^2} \sum_{n=1}^N \frac{\partial \mu_n}{\partial \theta_n} \mathbf{x}_n \mathbf{x}_n^\top = -\frac{1}{\sigma^2} \mathbf{X}^\top \mathbf{S} \mathbf{X},$$

где матрица  $\mathbf{S}$  — это диагональная матрица весов, составленная из производных обратной функции связи:

$$\mathbf{S} = \text{diag} \left( \frac{\partial \mu_1}{\partial \theta_1}, \frac{\partial \mu_2}{\partial \theta_2}, \dots, \frac{\partial \mu_N}{\partial \theta_N} \right).$$

- Аналогично логистической регрессии, в данном случае мы получим новый вариант метода итеративных взвешенных наименьших квадратов:

$$\mathbf{w}^{\text{new}} = \mathbf{w}^{\text{old}} - (\mathbf{X}^{\top} \mathbf{S} \mathbf{X})^{-1} \mathbf{X}^{\top} (\boldsymbol{\mu} - \mathbf{y}) = (\mathbf{X}^{\top} \mathbf{S} \mathbf{X})^{-1} \mathbf{X}^{\top} \mathbf{S} \mathbf{z},$$

где

$$\mathbf{z} = \mathbf{X} \mathbf{w}^{\text{old}} - \mathbf{S}^{-1} (\boldsymbol{\mu} - \mathbf{y}).$$

- Таким образом, мы можем использовать этот метод для того, чтобы найти гипотезу максимального правдоподобия в любой обобщённой линейной модели.
- В байесовский вывод углубляться не будем, но его тоже можно провести; он получится только приближённым, но здесь уже разумнее будет использовать общие методы приближённого вывода — MCMC-методы или вариационные приближения.

- А теперь можно и что-нибудь новенькое получить. Первый новый пример — пуассоновская регрессия (Poisson regression).
- Предположим, что целевая переменная  $y$  в нашей модели имеет смысл числа неких происходящих независимо друг от друга событий: звонки в колл-центр, лайки под новым постом в социальной сети, число мутаций в данном участке ДНК и так далее.
- Иначе говоря,  $y$  было бы неплохо описать распределением Пуассона; поскольку среднее здесь совпадает с интенсивностью  $\lambda$ , давайте сразу переобозначим её через  $\mu$ :

$$p(y|\mu) = \frac{1}{y!} \mu^y e^{-\mu}, \quad \text{то есть} \quad \log p(y|\mu) = y \log \mu - \mu - \log(y!).$$

- Мы видим, что естественным параметром здесь является  $\theta = \log \mu$ , и приняв, как обычно,  $\theta = \mathbf{w}^\top \mathbf{x}$ , мы получим обобщённую линейную модель

$$\log p(y|\mathbf{x}, \mathbf{w}) = y\mathbf{w}^\top \mathbf{x} - e^{\mathbf{w}^\top \mathbf{x}} - \log(y!),$$

то есть в данном случае  $\sigma^2 = 1$ ,  $a(\theta) = e^\theta = e^{\mathbf{w}^\top \mathbf{x}}$ ,  $h(y, \sigma^2) = \frac{1}{y!}$ .

- И теперь мы уже умеем обучать эту модель или градиентным спуском, или методом второго порядка.



- Однако распределение Пуассона часто оказывается недостаточно выразительным.
- У него дисперсия совпадает со средним, а в реальных данных это зачастую не так, и хотелось бы иметь похожее на пуассоновское распределение с переменной дисперсией.
- Возможный ответ на этот запрос — отрицательное биномиальное распределение

$$\text{NegBinom}(k|r, p) = \binom{k+r-1}{r-1} (1-p)^k p^r.$$

- Его среднее составляет  $\mu = \frac{r(1-p)}{p}$ , и в экспоненциальное семейство оно вкладывается как

$$\text{NegBinom}(k|r, p) = \binom{k+r-1}{r-1} e^{k \log p + r \log(1-p)}, \text{ то есть}$$

$$\theta = \log p, \quad t(k) = k, \quad a(\theta) = -r \log(1-e^\theta), \quad h(k) = \binom{k+r-1}{r-1}. \quad 5$$

- *Отрицательная биномиальная регрессия* (negative binomial regression) — это обобщённая линейная модель с отрицательным биномиальным распределением в качестве функции связи:

$$\mu = g^{-1}(\mathbf{w}^\top \mathbf{x}) = a'(\mathbf{w}^\top \mathbf{x}) = r \frac{e^{\mathbf{w}^\top \mathbf{x}}}{1 - e^{\mathbf{w}^\top \mathbf{x}}} = \frac{r}{e^{-\mathbf{w}^\top \mathbf{x}} - 1}.$$

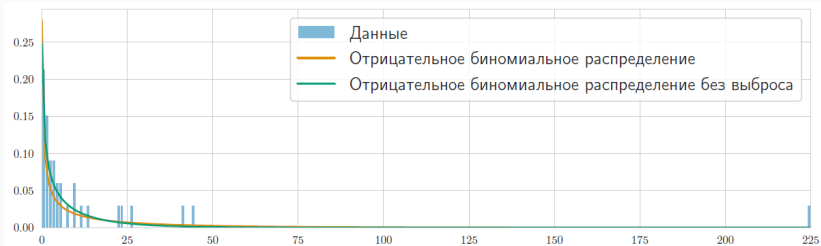
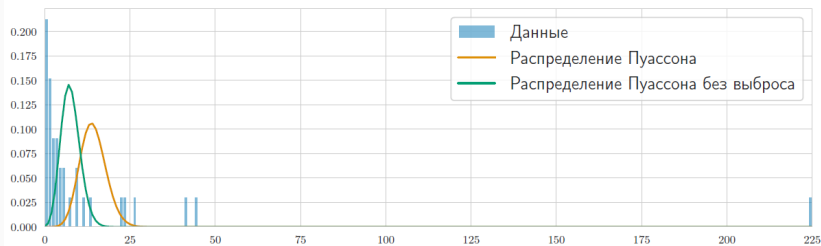
- В параметризации средним  $p = \frac{\mu}{\mu+r}$ ,  $1 - p = \frac{r}{\mu+r}$ , то есть

$$\text{NegBinom}(k|r, \mu) = \binom{k+r-1}{r-1} \left( \frac{r}{r+\mu} \right)^r \left( \frac{\mu}{r+\mu} \right)^k.$$

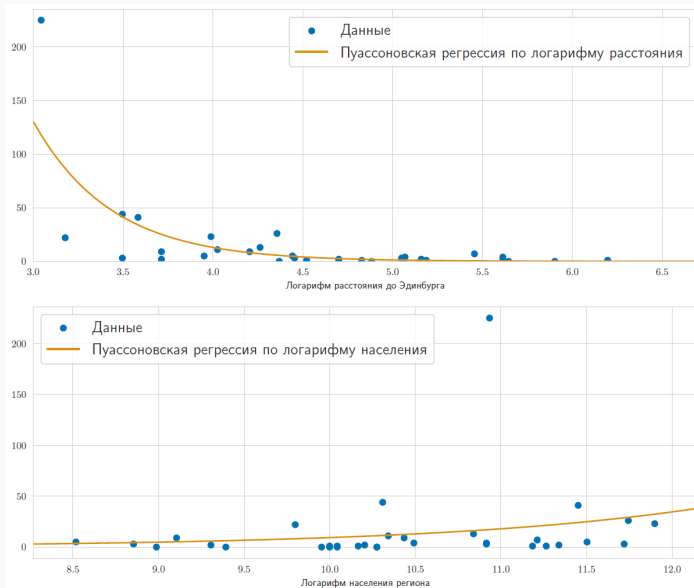
- И теперь мы уже тоже умеем обучать эту модель или градиентным спуском, или методом второго порядка.

Регион	Расстояние до Эдинбурга	Число подмастерий	Население (тыс.)	Степень урбанизации	Направление от Эдинбурга	Регион	Расстояние до Эдинбурга	Число подмастерий	Население (тыс.)	Степень урбанизации	Направление от Эдинбурга
Midlothian	21	225	56	18.8	S	Ayr	110	2	84	26.4	W
West Lothian	24	22	18	37.9	W	Kircudbright	110	0	29	11.3	S
East Lothian	33	44	30	43.4	S	Kincardine	125	1	26	12.3	N
Kinross	33	3	7	30.3	N	Bute	132	0	12	43.6	W
Fife	36	41	94	41.3	N	Aberdeen	156	3	123	23.1	N
Peebles	41	9	9	29.3	S	Wigtown	157	0	22	23.2	S
Clackmannan	41	2	11	47.4	N	Banff	159	4	36	12.9	N
Selkirk	52	5	5	41.9	S	Moray	174	2	27	27.6	N
Lanark	54	23	147	68.1	W	Nairn	175	0	8	45.3	N
Berwick	56	11	31	15.2	S	Argyll	179	1	72	12.7	W
Roxburgh	67	9	34	31.8	S	Inverness	234	7	74	10.8	N
Stirling	71	13	51	31.1	W	Ross	274	4	55	10.7	N
Perth	78	26	126	14.4	N	Caithness	274	1	23	28.3	N
Dunbarton	79	0	21	27.3	W	Sutherland	283	0	23	10.3	N
Angus	85	5	99	55.3	N	Orkney	366	0	29	9.0	N
Dumfries	86	3	55	25.9	S	Shetland	491	1	22	7.7	N
Renfrew	92	1	78	69.9	W						

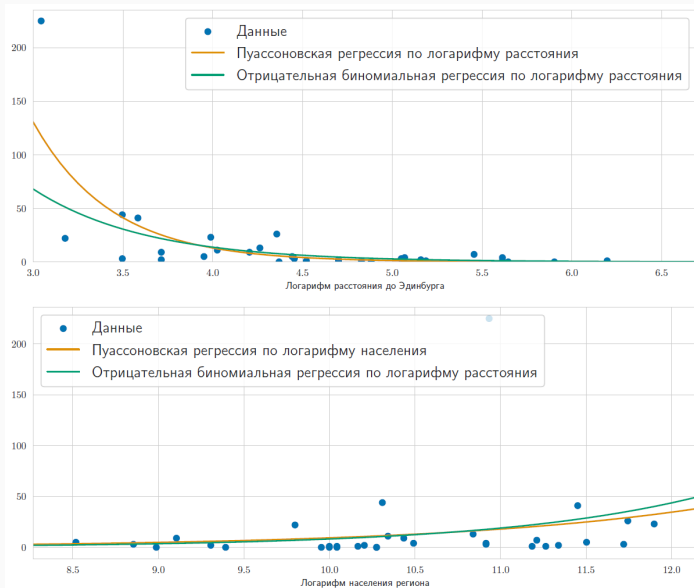
# ПРИМЕР



# ПРИМЕР



# ПРИМЕР



СПАСИБО!

Спасибо за внимание!

