ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Экспоненциальное семейство и GLM





Машинное обучение · Основы байесовского вывода

11–12

Сергей Николенко · Центральный университет, Москва · 7 ноября 2026 г.

Random facts:

- 7 ноября 1631 г. Пьер Гассенди впервые наблюдал прохождение Меркурия по диску Солнца, предсказанное Кеплером
- 7 ноября 1665 г. вышел первый номер Oxford Gazette — будущей London Gazette
- 7 ноября 1929 г. в Нью-Йорке открылся Музей современного искусства MoMA
- 7 ноября 1940 г. Такомский мост разрушился от аэроупругих колебаний всего через четыре месяца после открытия
- 7 ноября 1996 г. к Марсу стартовал Mars Global Surveyor, которому предстояло составить подробную карту планеты

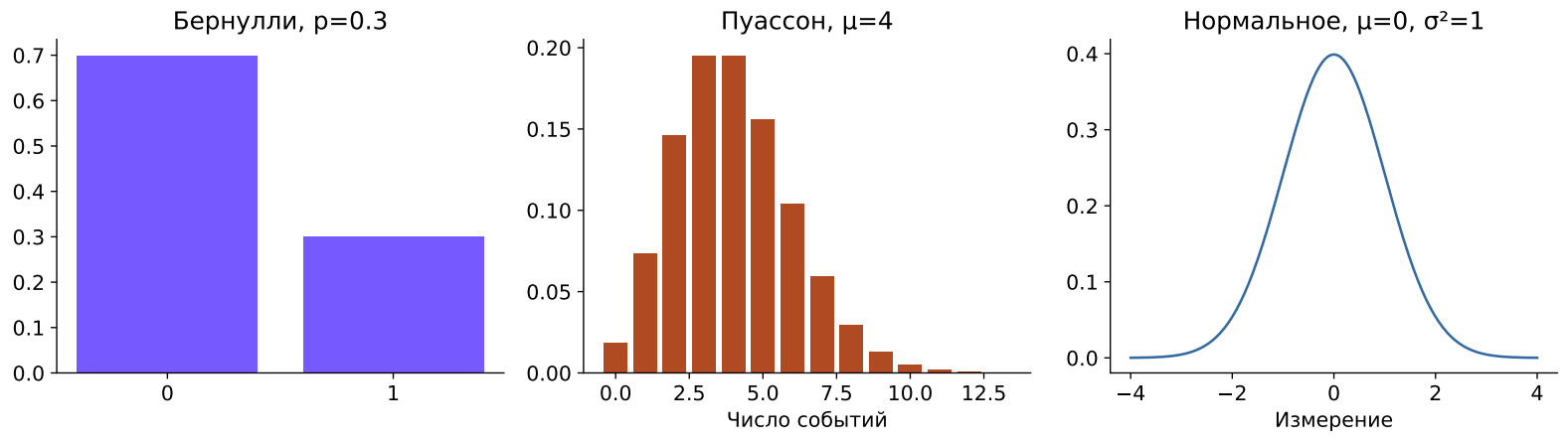
Краткое резюме предыдущей серии

- В прошлый раз мы выбирали между моделями разной сложности: считали правдоподобие данных для каждой модели и разбирались, откуда в информационных критериях берётся штраф за число параметров.
- Теперь объединим знакомые распределения в экспоненциальное семейство и поймём роль достаточных статистик.
- А потом зададим тот же вопрос про регрессию: что общего у линейной и логистической регрессии и как построить такую же модель, скажем, для числа звонков в колл-центр?
- В конце вернёмся к оценке размера неизвестной популяции и границ равномерного распределения.

Экспоненциальное семейство

Как читать общую запись

- Что общего у монетки, числа звонков в колл-центр и роста случайного прохожего? На первый взгляд, даже множества возможных ответов у них совершенно разные.
- Давайте попробуем записать все три распределения по одной формуле. Какие её части будут зависеть от наблюдения, а какие — от параметра?



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

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

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

Экспоненциальное семейство

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

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

- Определение выглядит очень общим; главное предположение здесь в том, как $\boldsymbol{\theta}$ 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 $g(\boldsymbol{\theta})$, то есть единственная зависимость между $\mathbf{x}$ и $\boldsymbol{\theta}$ — это скалярное произведение в экспоненте.
- Вообще говоря, почти всё, о чём мы говорили – частные случаи *экспоненциального семейства* распределений.

Экспоненциальное семейство

- Например, биномиальное распределение

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

Экспоненциальное семейство

- 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; $t(k) = k$, $h(k) = \binom{n}{k}$,

$$a(\theta) = -n \log(1-p) = n \log(1+e^\theta),$$

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

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

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

Экспоненциальное семейство

- Здесь параметры $\log p_i$ не независимы. Выберем категорию k опорной и выразим $p_k = 1 - \sum_{i<k} p_i$:

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

Экспоненциальное семейство

- Достаточные статистики теперь $(x_1, \dots, x_{k-1})$,

$$\theta_i = \log(p_i/p_k), \quad \theta_k = 0, \qquad p_i = \frac{e^{\theta_i}}{1 + \sum_{j<k} e^{\theta_j}}.$$

  И $a(\boldsymbol{\theta}) = n \log(1 + \sum_{j<k} e^{\theta_j})$.

- В обратном выражении для p_i через $\boldsymbol{\theta}$ 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 $t(x) = x$, $\theta = \log\lambda$, $h(x) = \frac{1}{x!}$, $a(\theta) = \lambda = e^\theta$.

- А вот у гипергеометрического распределения при неизвестном K и фиксированных N, n

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

  множество возможных x может меняться вместе с K. В регулярном экспоненциальном семействе носитель задаётся одной и той же функцией $h(x)$.

Экспоненциальное семейство

- Нормальное распределение:

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

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

Экспоненциальное семейство

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

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

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

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

Экспоненциальное семейство

- Многомерный гауссиан:

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

Экспоненциальное семейство

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

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

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

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

Экспоненциальное семейство

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

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

Экспоненциальное семейство

- Возьмём градиент по $\boldsymbol{\theta}$ слева и справа:

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

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

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

Экспоненциальное семейство

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

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

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

Экспоненциальное семейство

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

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

- Для гауссиана с неизвестными средним и дисперсией это $(\mathbb{E}[x^2], \mathbb{E}[x]) = (\mu^2 + \sigma^2, \mu)$; при известной дисперсии достаточно μ.

Экспоненциальное семейство

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

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

Экспоненциальное семейство

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

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

- Аналогичные результаты верны и для других моментов.

Экспоненциальное семейство

- Второй результат – о правдоподобии:

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

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

- Всё, что в логарифме правдоподобия зависит от $\boldsymbol{\theta}$, 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 μ, и для σ^2.

Экспоненциальное семейство

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

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

- Для естественного параметра $\boldsymbol{\eta}(\boldsymbol{\theta}) = \boldsymbol{\theta}$ правая часть не зависит от $\boldsymbol{\theta}$. Для естественной параметризации средний вектор достаточных статистик равен градиенту a; поэтому

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

Экспоненциальное семейство

- *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 $\boldsymbol{\theta}$ семейством сопряжённых априорных распределений будет

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

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

Экспоненциальное семейство

- Давайте найдём апостериорное распределение:

$$p\left(\boldsymbol{\theta}|\mathbf{x}_1, \dots, \mathbf{x}_N, \boldsymbol{\chi}_0, \nu_0\right) \propto p\left(\boldsymbol{\theta}|\boldsymbol{\chi}_0, \nu_0\right) p\left(\mathbf{x}_1, \dots, \mathbf{x}_N|\boldsymbol{\theta}\right).$$

Экспоненциальное семейство

- Подставим:

$$\log p\left(\boldsymbol{\theta}|\mathbf{x}_1,\dots,\mathbf{x}_N,\boldsymbol{\chi}_0,\nu_0\right) = \sum_{n=1}^{N}\log h(\mathbf{x}_n) + \\ + \boldsymbol{\eta}(\boldsymbol{\theta})^\top\left(\sum_{n=1}^{N}\mathbf{t}(\mathbf{x}_n)\right) - Na(\boldsymbol{\theta}) + \\ + \log f(\boldsymbol{\chi}_0,\nu_0) + \boldsymbol{\chi}_0^\top\boldsymbol{\eta}(\boldsymbol{\theta}) - \nu_0 a(\boldsymbol{\theta}) + \text{const} = \\ = \text{const} + \left(\boldsymbol{\chi}_0 + \sum_{n=1}^{N}\mathbf{t}(\mathbf{x}_n)\right)^\top\boldsymbol{\eta}(\boldsymbol{\theta}) - \left(N+\nu_0\right)a(\boldsymbol{\theta}),$$

Экспоненциальное семейство

- В итоге у нас получилось апостериорное распределение $p\left(\boldsymbol{\theta}|\boldsymbol{\chi}_N, \nu_N\right)$ того же вида, но с параметрами

$$\nu_N = \nu_0 + N, \qquad \boldsymbol{\chi}_N = \boldsymbol{\chi}_0 + \sum_{n=1}^{N} \mathbf{t}(\mathbf{x}_n).$$

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:

$$p\left(\mathbf{x}|\mathbf{x}_1, \dots, \mathbf{x}_n, \boldsymbol{\chi}_0, \nu_0\right) = \int p\left(\mathbf{x}|\boldsymbol{\theta}\right) p\left(\boldsymbol{\theta}|\mathbf{x}_1, \dots, \mathbf{x}_N, \boldsymbol{\chi}_0, \nu_0\right) \mathrm{d}\boldsymbol{\theta} =$$

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

$$= h(\mathbf{x}) f(\boldsymbol{\chi}_N, \nu_N) \int e^{\left(\mathbf{t}(\mathbf{x}) + \boldsymbol{\chi}_N\right)^\top \boldsymbol{\eta}(\boldsymbol{\theta}) - (\nu_N + 1) a(\boldsymbol{\theta})} \, d\boldsymbol{\theta}$$

$$= h(\mathbf{x}) \frac{f(\boldsymbol{\chi}_N, \nu_N)}{f(\boldsymbol{\chi}_N + \mathbf{t}(\mathbf{x}), \nu_N + 1)}.$$

Экспоненциальное семейство

- Получилось, что предсказательная плотность — это $h(\mathbf{x})$, умноженная на отношение нормировочных констант сопряжённого семейства.
- Сравните это с выводом предсказательного распределения для испытаний Бернулли, который мы делали в начале курса.

Обобщённые линейные модели

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:

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

  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

- *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 μ как функцию от c.
- 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.

Обобщённые линейные модели

- Например, в линейной регрессии $g = g^{-1} = \mathrm{id}$, а в бинарной логистической регрессии $g^{-1}(a) = \sigma(a)$, то есть $g(\mu) = \log \frac{\mu}{1-\mu}$.
- Далее нужно определить одномерное распределение $p\left(y|\mathbf{x}, \mathbf{w}\right)$, среднее которого равно $\mu = g^{-1}(\mathbf{w}^\top \mathbf{x})$.

Обобщённые линейные модели

- Правда, мы хотим научиться контролировать не только среднее, но и дисперсию на выходе, поэтому вместо общего вида экспоненциального семейства будем рассматривать распределение вида

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

  где σ^2 — параметр дисперсии (dispersion parameter).
- Само это семейство иногда называют *дисперсным экспоненциальным семейством* (exponential dispersion 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 только что доказанный нами результат $\mathbb{E}\left[\mathbf{t}(\mathbf{x})\right] = \boldsymbol{\nabla}_{\boldsymbol{\theta}} a(\boldsymbol{\theta})$ сразу даёт среднее

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

- А результат

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

даёт дисперсию

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

то есть

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

Обобщённые линейные модели

- Теперь выберем каноническую связь: $\theta = \mathbf{w}^\top \mathbf{x}$. Для других связей линейным остаётся $g(\mu)$, а не обязательно θ.
- Это значит, что функция связи здесь определяется как

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

Обобщённые линейные модели

- Так, для линейной регрессии

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

поэтому для линейной регрессии логично положить

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

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

- 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

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

Обобщённые линейные модели

- Получилось вложение логистической регрессии в обобщённые линейные модели:

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

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

Обобщённые линейные модели

- Но эти две модели мы знали и так. А как научиться обучать произвольную модель с канонической связью?
- Начнём с логарифма правдоподобия одного наблюдения:

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

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

$$\ell(\mathbf{w}) = \log p\left(y_1, \dots, y_N\middle|\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{w}, \sigma^2\right) =$$

$$= \text{const} + \frac{1}{\sigma^2}\sum_{n=1}^N \left(y_n\mathbf{w}^\top\mathbf{x}_n - a(\mathbf{w}^\top\mathbf{x}_n)\right).$$

- От него теперь можно взять градиент:

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

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

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

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

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

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

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

где

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

Обобщённые линейные модели

- Эта запись даёт IRLS для канонической связи. Для другой связи меняются рабочие веса и отклик; общий метод остаётся итеративным взвешенным МНК.
- 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.

Обобщённые линейные модели

- 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 λ, давайте сразу переобозначим её через μ:

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

Обобщённые линейные модели

- 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

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

  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

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

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

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

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

$$h(k) = \binom{k+r-1}{r-1}.$$

Обобщённые линейные модели

- *Отрицательная биномиальная регрессия* (negative binomial regression) — это обобщённая линейная модель с отрицательным биномиальным условным распределением при фиксированном r:

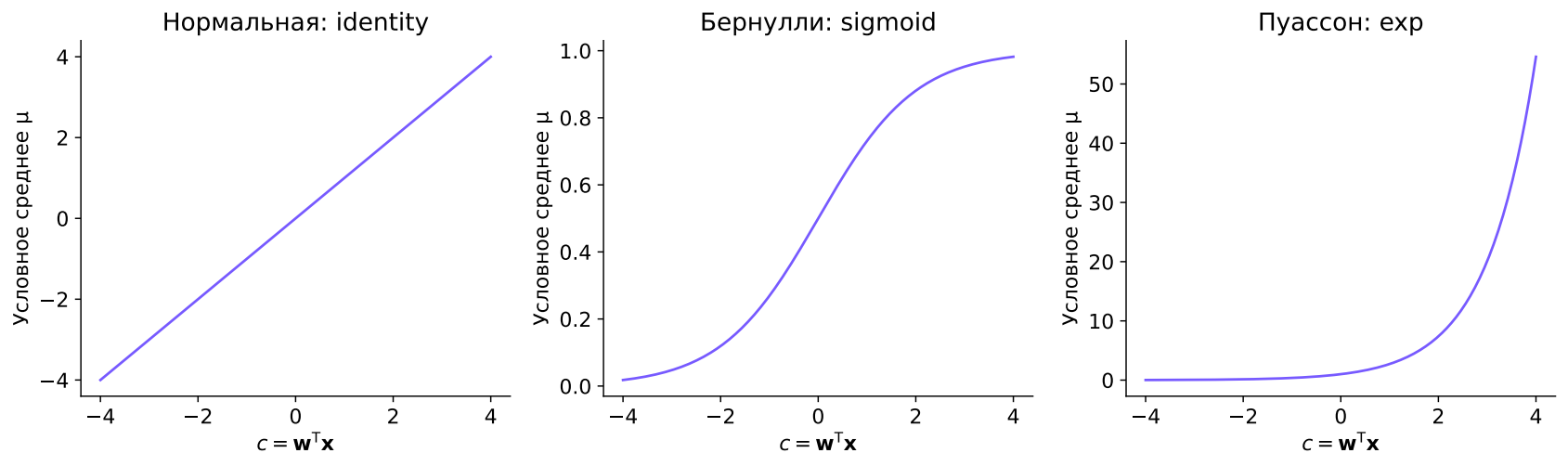
$$\mu = g^{-1}(\mathbf{w}^\top \mathbf{x}) = a'(\mathbf{w}^\top \mathbf{x}) = r\frac{e^{\mathbf{w}^\top \mathbf{x}}}{1 - e^{\mathbf{w}^\top \mathbf{x}}} = \frac{r}{e^{-\mathbf{w}^\top \mathbf{x}} - 1}.$$

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

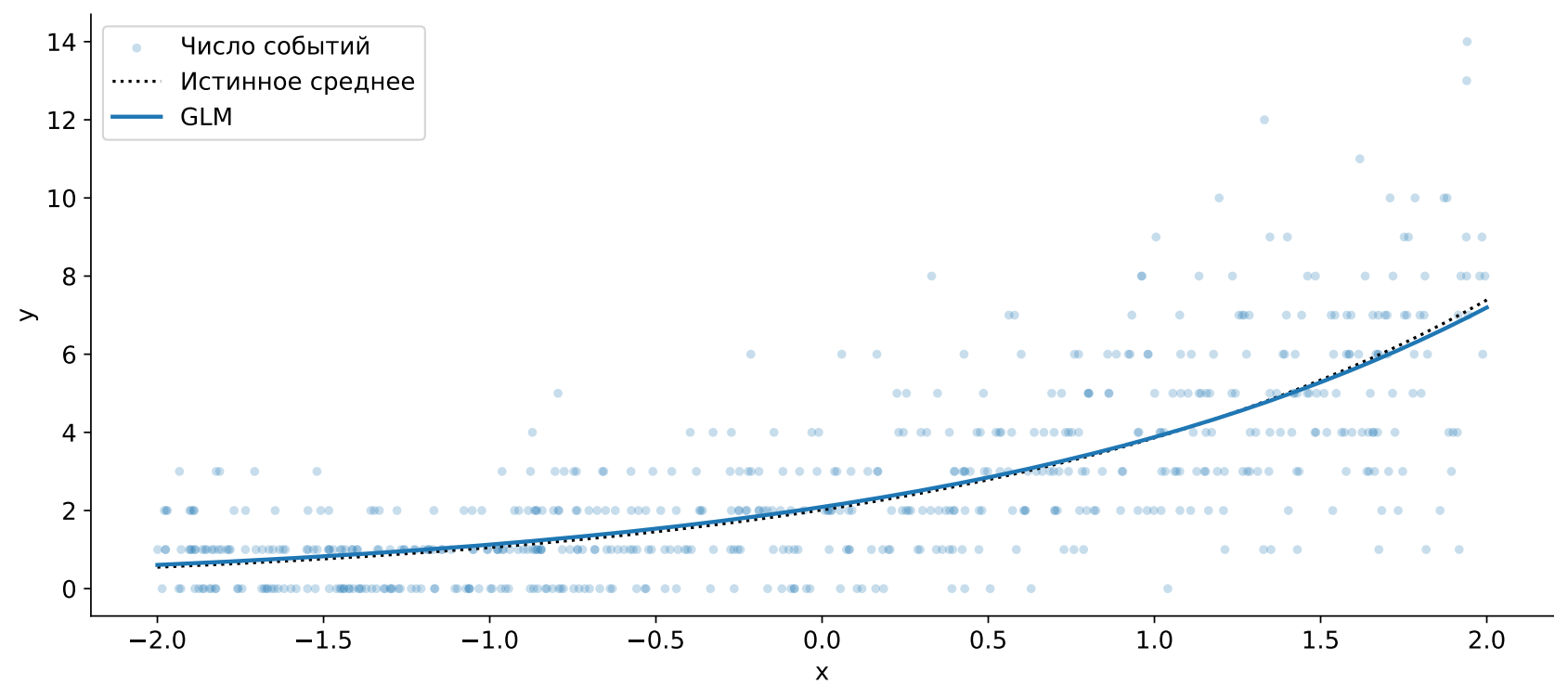
$$\text{NegBinom}\,(k|r, \mu) = \binom{k + r - 1}{r - 1} \left(\frac{r}{r + \mu}\right)^r \left(\frac{\mu}{r + \mu}\right)^k.$$

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

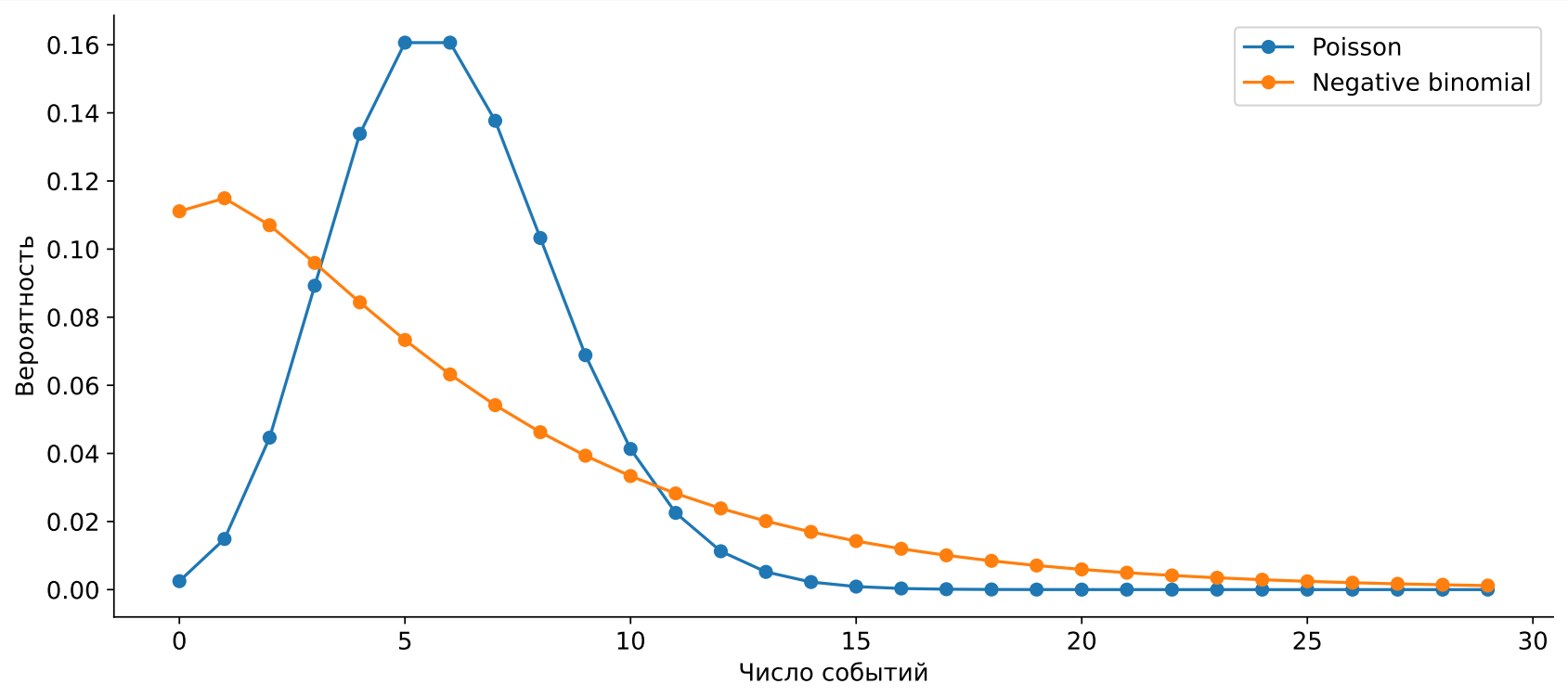
Один предиктор — разные области значений среднего



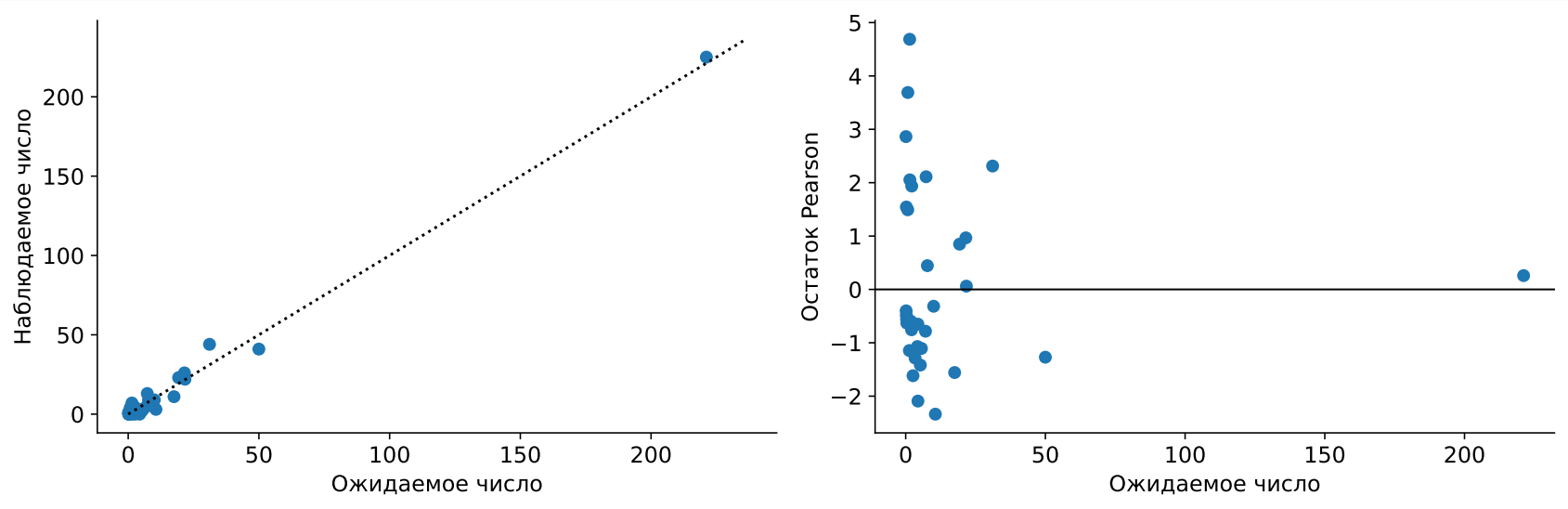
Пуассоновская регрессия



Среднее не определяет хвосты: сверхдисперсия



Сколько учеников в каждом округе?

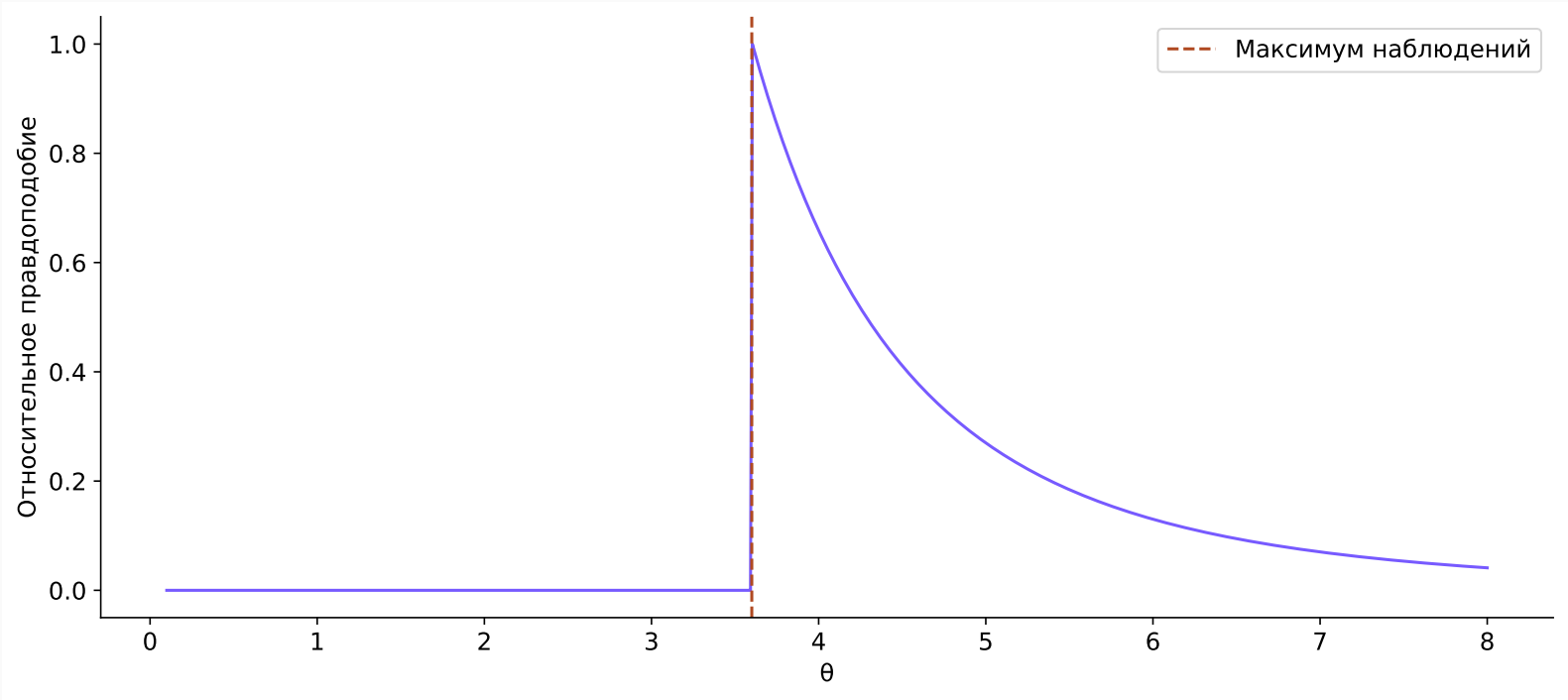


Две связи для отрицательной биномиальной регрессии

- При $P(Y=k) = \binom{k+r-1}{k} p^r (1-p)^k$ имеем $\mu = r(1-p)/p$ и $p = r/(r+\mu)$.
- Канонический параметр $\theta = \log(1-p) = \log(\mu/(\mu+r)) < 0$.
- На практике часто берут log-связь $\log \mu = \mathbf{w}^\top \mathbf{x}$: она не каноническая для NB, но обеспечивает положительность среднего.
- При фиксированном r это GLM; если r тоже оценивается, нужна дополнительная процедура для параметра дисперсии.

Почему равномерное распределение — отдельный случай?

- У $\mathrm{Uniform}(0, \theta)$ носитель $[0, \theta]$ зависит от параметра.
- Максимум выборки достаточен, но регулярные формулы экспоненциального семейства напрямую не применяются.



ЦЕНТРАЛЬНЫЙ УНИВЕРСИТЕТ

Спасибо за внимание!



