# ОБОБЩЕННЫЕ ЛИНЕЙНЫЕ МОДЕЛИ

Сергей Николенко

СПбГУ — Санкт-Петербург

06 декабря 2022 г.

---

*Random facts*:

- 6 декабря 1917 г. парламент Финляндии принял решение выйти из состава России, а 6 декабря 1920 г. была упразднена республика Северная Ингрия, которая была образована летом 1919 г. в северной части Петроградского уезда, со столицей в деревне Кирьясало
- 6 декабря 1928 г. в колумбийском городе Сьенага произошла «Банановая бойня»: власти вывели войска против бастующих сотрудников United Fruit Company и открыли огонь; оценки числа погибших расходятся от 47 до 2000 человек
- 6 декабря 1933 г. федеральный судья США Джон Вулзи решил, что «Улисс» Джеймса Джойса не является непристойным и может быть опубликован, а 6 декабря 1953 г. Владимир Набоков закончил «Лолиту»
- 6 декабря 1956 г. на Олимпийских играх в Мельбурне состоялся матч по водному поло между сборными СССР и Венгрии, получивший название «Кровь в бассейне»
- 6 декабря 2000 г. в Дубне в сотрудничестве с Ливерморской лабораторией был открыт ливерморий, 116-й элемент таблицы Менделеева

# 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 $\mu$ 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.

# 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 $p\left(y|\mathbf{x}, \mathbf{w}\right)$ 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

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

  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 $\frac{1}{\sigma^2} a(\theta)$.

# Обобщённые линейные модели

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

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

- А результат

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

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

- 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

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

- 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

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

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

# Обобщённые линейные модели

- 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:

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

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

# Обобщённые линейные модели

- Но эти 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

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

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

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

# Обобщённые линейные модели

- От него теперь можно взять градиент:

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

  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.

# Обобщённые линейные модели

- 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:

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

# Обобщённые линейные модели

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

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

  где

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

- 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.

# 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 $\lambda$, давайте сразу переобозначим её через $\mu$:

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

# 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.

# Обобщённые линейные модели

- 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

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

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

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

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

# Обобщённые линейные модели

- *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.

# Спасибо!

Спасибо за внимание!