ОБОБЩЕННЫЕ ЛИНЕЙНЫЕ МОДЕЛИ

Сергей Николенко

СПбГУ — Санкт-Петербург

04 ноября 2023 г.

Random facts:

- 4 ноября 1582 г. Ермак разбил хана Кучума в сражении на Чувашем мысу и через 3 дня вступил в столицу его ханства — Искер
- 4 ноября 1847 г. сэр Джеймс Янг Симпсон, шотландский врач, надышался хлороформом и открыл его анестезирующие свойства
- 4 ноября 1890 г. в Лондоне открылась первая в мире подземная электрическая дорога; метро на конной тяге существовало в Лондоне ещё с 1863 г.
- 4 ноября 1922 г. Говард Картер открыл в Долине Царей гробницу Тутанхамона
- 4 ноября 1937 г. в Москве было запущено первое в СССР производство пломбира, по привезённым Микояном американским рецептам и на американском оборудовании
- 4 ноября 1956 г. началась операция «Вихрь»: ввод в Венгрию советских воинских частей и штурм Будапешта под командованием маршала Жукова
- 4 ноября 1995 г. ультраправый экстремист Игаль Амир застрелил Ицхака Рабина на площади Царей Израиля в Тель-Авиве, после выступления на многотысячном митинге в поддержку мирного процесса

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 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

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

- 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.

Экспоненциальное семейство

- Нормальное распределение:

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

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

- Иначе говоря, одномерное нормальное распределение имеет две достаточные статистики, $\mathbf{t}(x) = \binom{x^2}{x}$, 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 θ 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 θ нет лишних параметров, это можно обратить, то есть выразить θ 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 θ, 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 θ 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 θ. 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 θ семейством сопряжённых априорных распределений будет

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

где χ и ν — 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:

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

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.

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 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 μ, 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 σ^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 θ; 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

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

- 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 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.

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

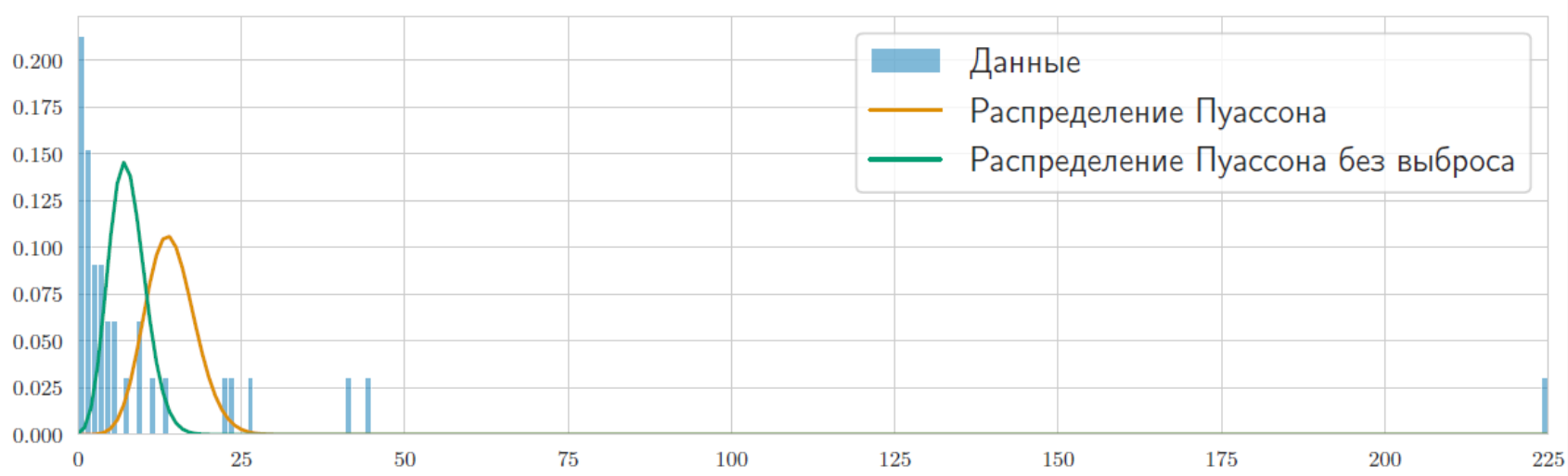
$$\text{NegBinom}\,(k|r, \mu) = \binom{k + r - 1}{r - 1}\left(\frac{r}{r + \mu}\right)^r \left(\frac{\mu}{r + \mu}\right)^k.$$

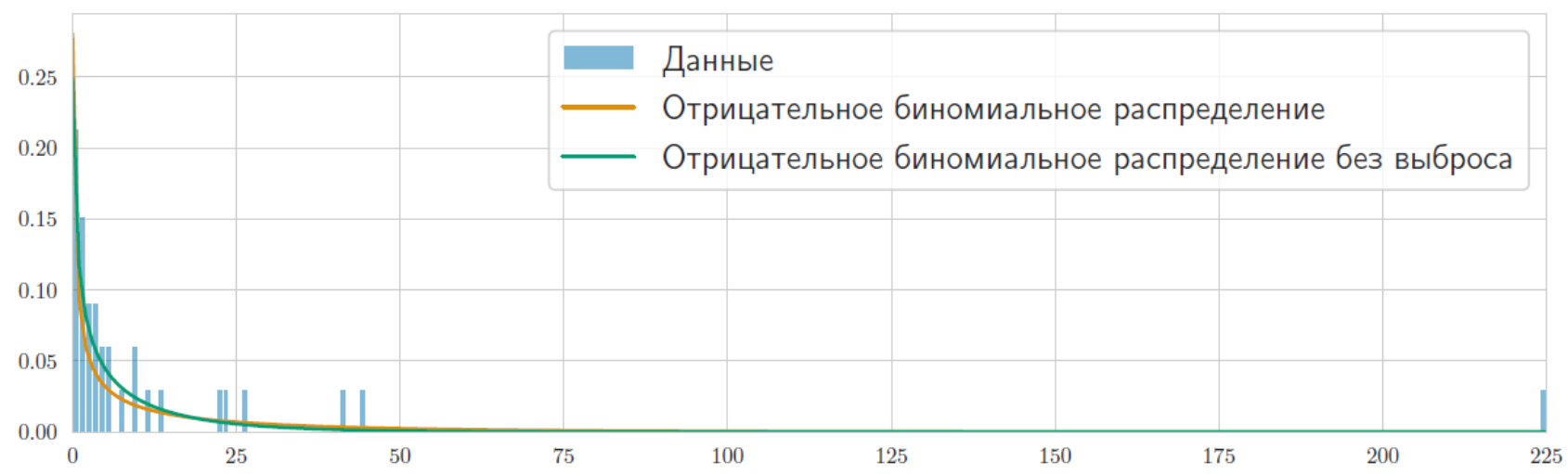
- И теперь мы уже тоже умеем обучать эту модель или градиентным спуском, или методом второго порядка.

Пример

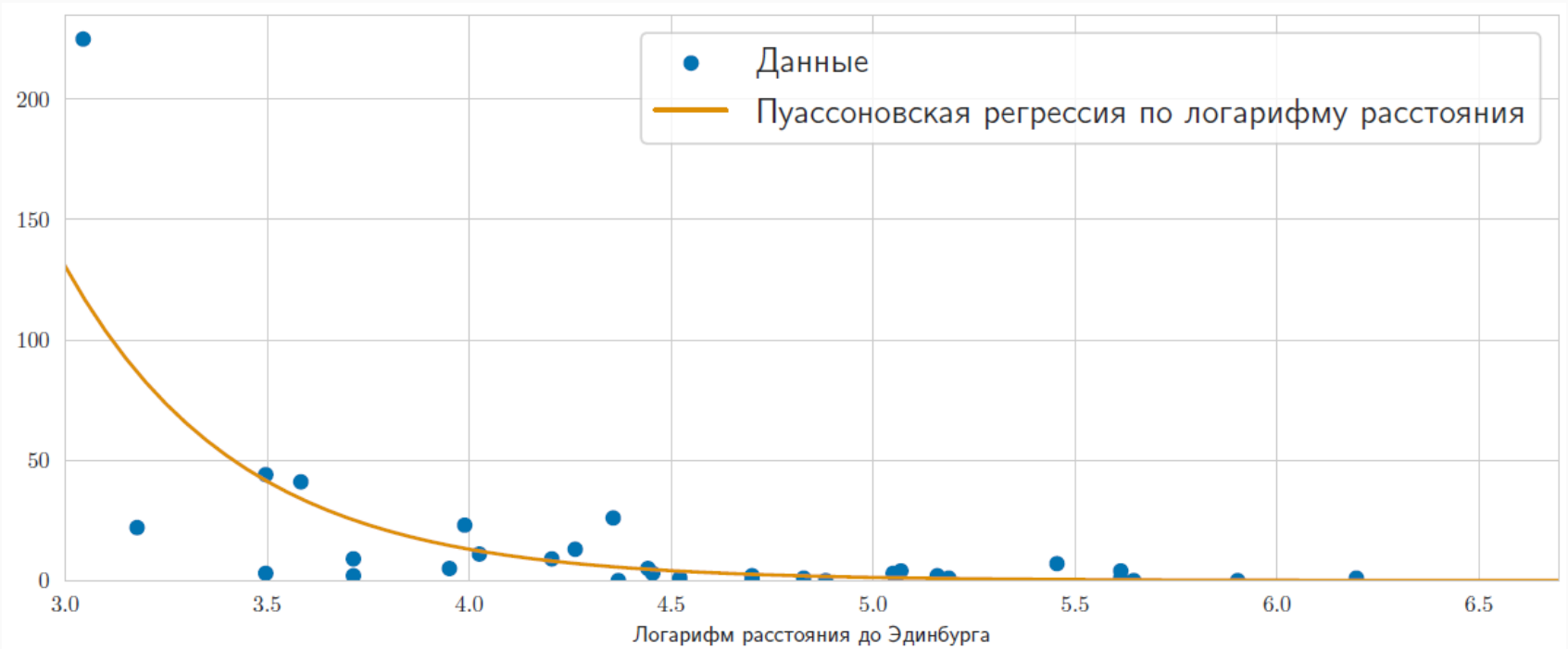
Регион	Расстояние до Эдинбурга	Число подмастерий	Население (тыс.)	Степень урбанизации	Направление от Эдинбурга	Регион	Расстояние до Эдинбурга	Число подмастерий	Население (тыс.)	Степень урбанизации	Направление от Эдинбурга
Midlothian	21	225	56	18.8	S	Ayr	110	2	84	26.4	W
West Lothian	24	22	18	37.9	W	Kircudbright	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


Данные
Распределение Пуассона
Распределение Пуассона без выброса
0.200
0.175
0.150
0.125
0.100
0.075
0.050
0.025
0.000
0
25
50
75
100
125
150
175
200
225

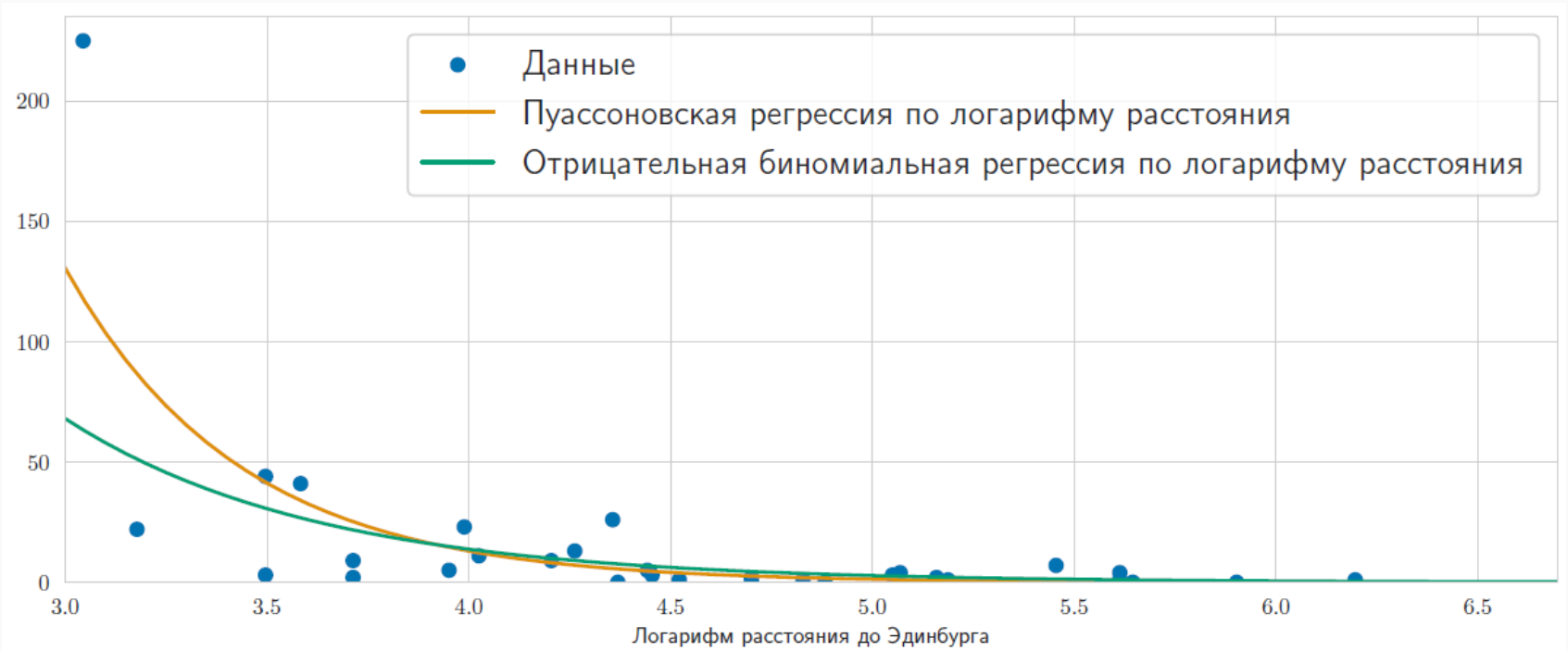

Данные
Отрицательное биномиальное распределение
Отрицательное биномиальное распределение без выброса
0.25
0.20
0.15
0.10
0.05
0.00
0
25
50
75
100
125
150
175
200
225

Пример


Данные
Пуассоновская регрессия по логарифму расстояния
200
150
100
50
0
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
Логарифм расстояния до Эдинбурга


Данные
Пуассоновская регрессия по логарифму населения
200
150
100
50

Пример


Данные
Пуассоновская регрессия по логарифму расстояния
Отрицательная биномиальная регрессия по логарифму расстояния
200
150
100
50
0
3.0
3.5
4.0
4.5
5.0
5.5
6.0
6.5
Логарифм расстояния до Эдинбурга


Данные
Пуассоновская регрессия по логарифму населения
Отрицательная биномиальная регрессия по логарифму расстояния
200
150
100
50

Эквивалентное ядро

Эквивалентное ядро

- Вспомним наши байесовские предсказания:

$$p(t \mid \mathbf{t}, \alpha, \beta) = N(t \mid \mu_N^\top \phi(\mathbf{x}), \sigma_N^2),$$
$$\text{где } \sigma_N^2 = \frac{1}{\beta} + \phi(\mathbf{x})^\top \Sigma_N \phi(\mathbf{x}).$$

- Давайте перепишем среднее апостериорного распределения в другой форме (вспомним, что $\mu_N = \beta \Sigma_N \Phi^\top \mathbf{t}$):

$$y(\mathbf{x}, \mu_N) = \mu_N^\top \phi(\mathbf{x}) = \beta \phi(\mathbf{x})^\top \Sigma_N \Phi^\top \mathbf{t} =$$
$$= \sum_{n=1}^{N} \beta \phi(\mathbf{x})^\top \Sigma_N \phi(\mathbf{x}_n) t_n.$$

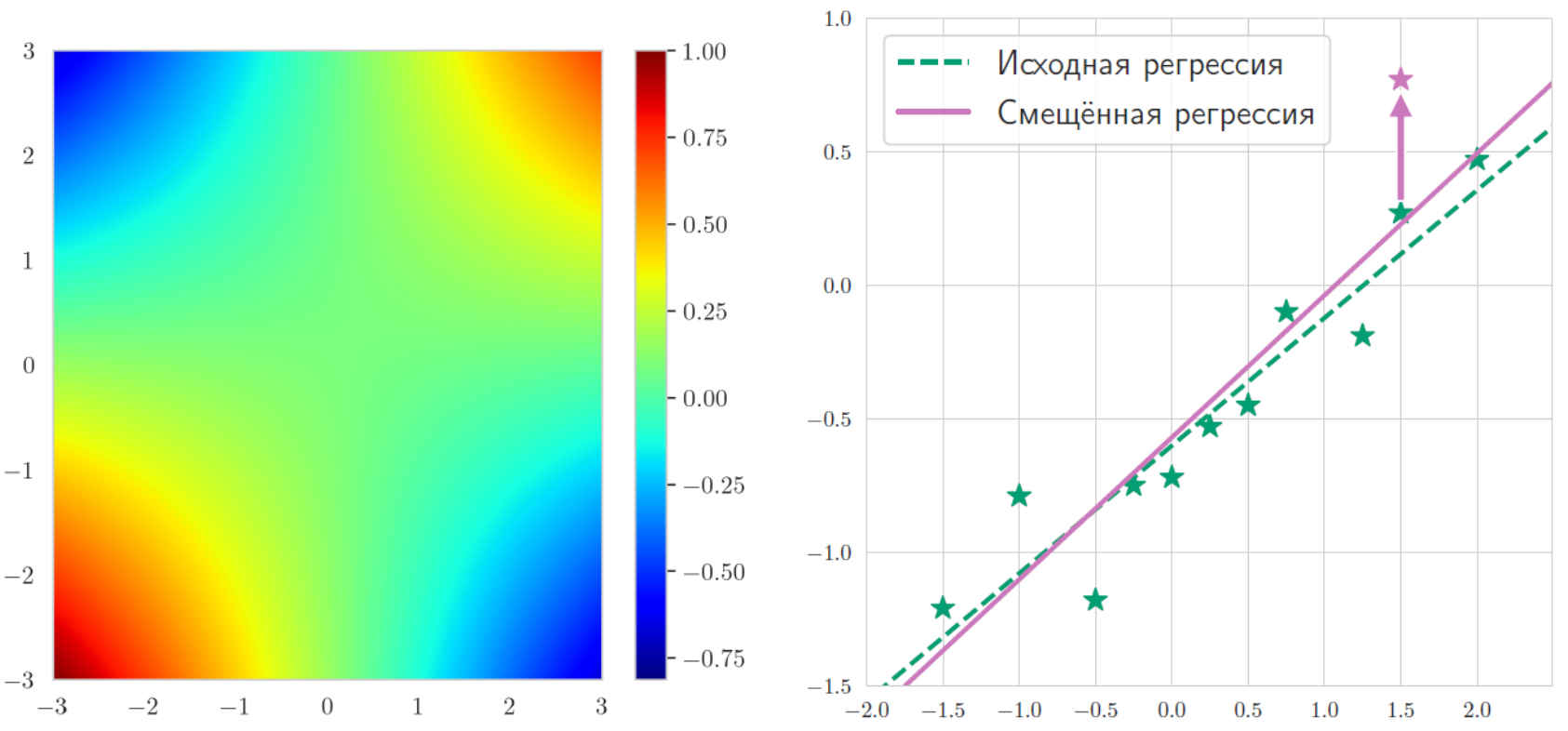
Эквивалентное ядро

- $y(\mathbf{x}, \mu_N) = \sum_{n=1}^{N} \beta\phi(\mathbf{x})^\top \Sigma_N \phi(\mathbf{x}_n) t_n$.
- Это значит, что предсказание можно переписать как

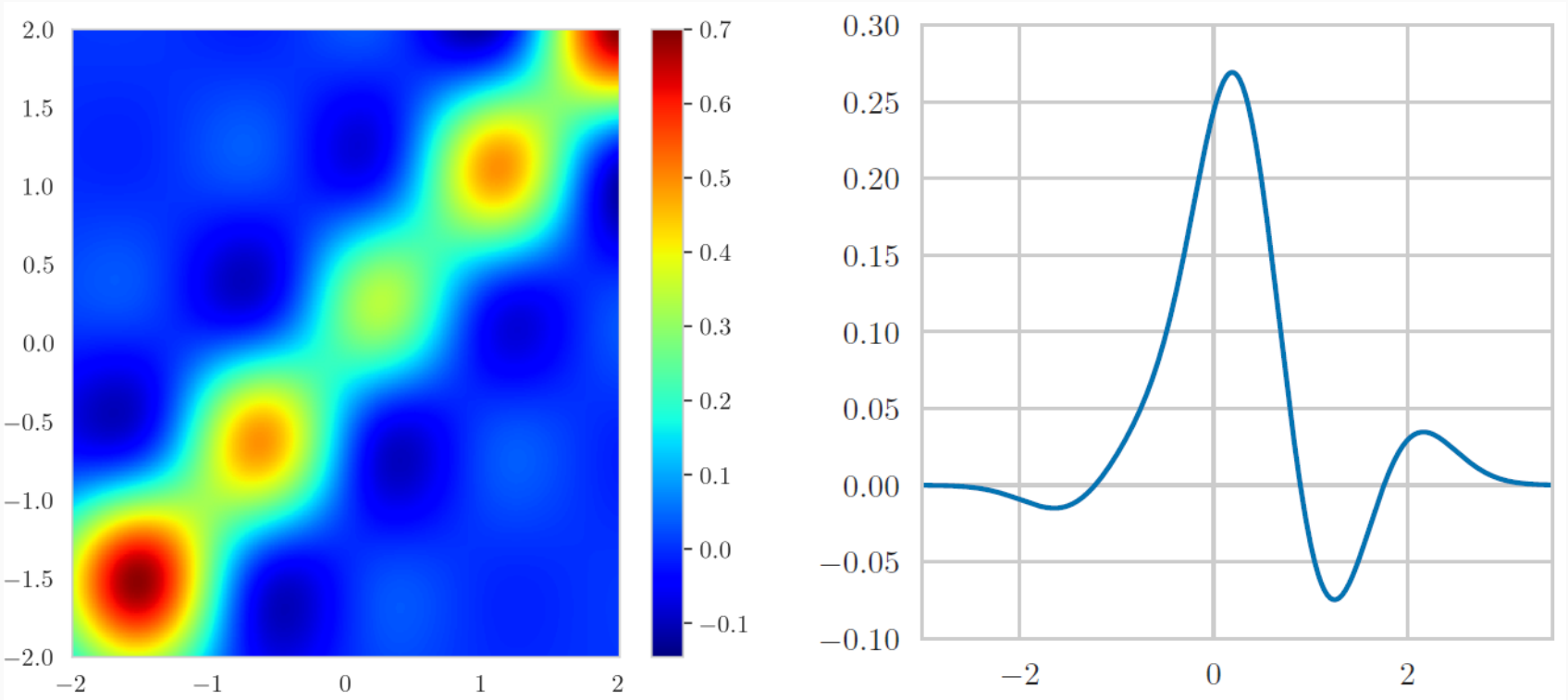
$$y(\mathbf{x}, \mu_N) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n.$$

- Т.е. мы предсказываем следующую точку как линейную комбинацию значений в известных точках.
- Функция $k(\mathbf{x}, \mathbf{x}') = \beta\phi(\mathbf{x})^\top \Sigma_N \phi(\mathbf{x}')$ называется *эквивалентным ядром* (equivalent kernel).

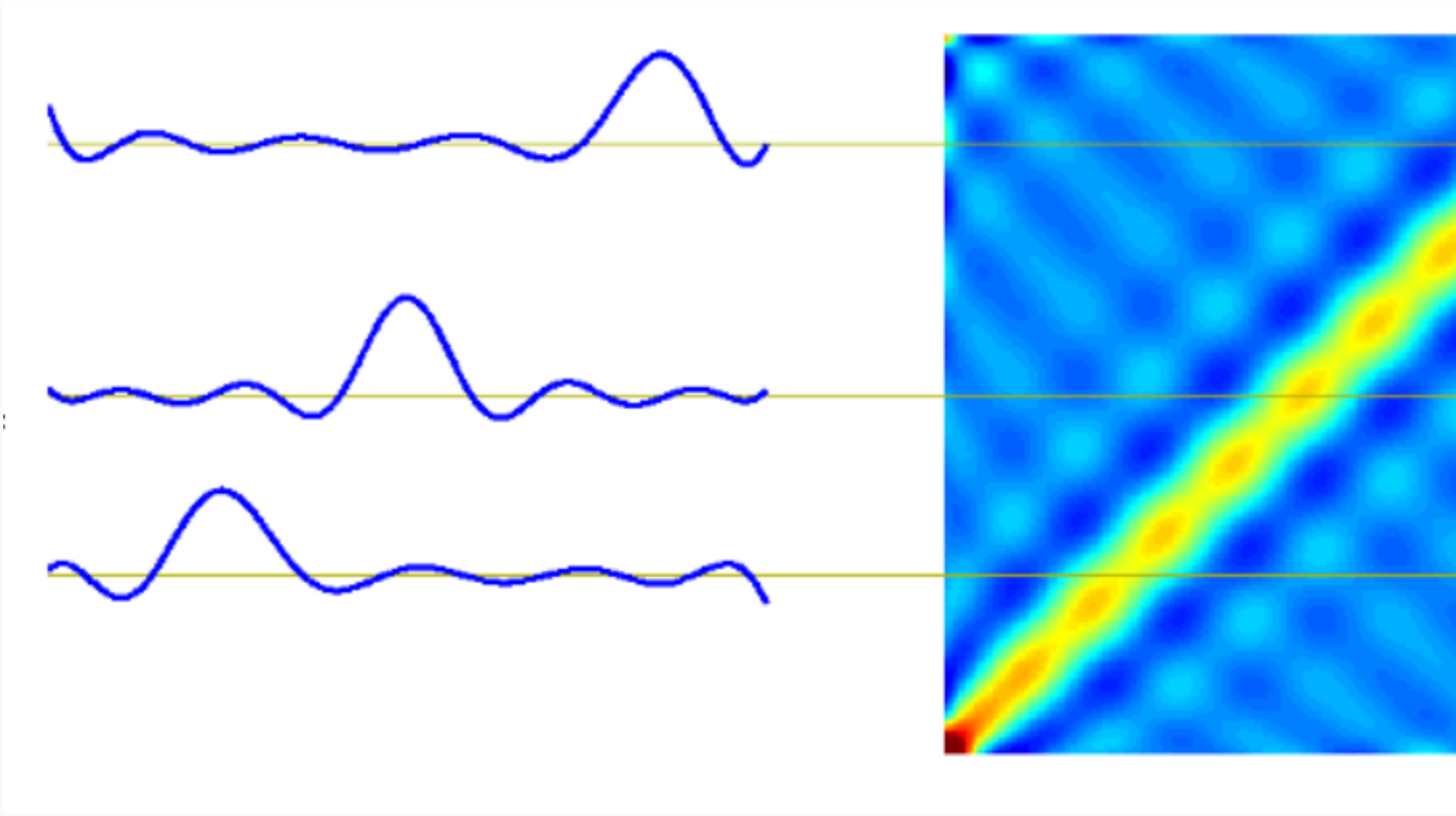
Эквивалентное ядро



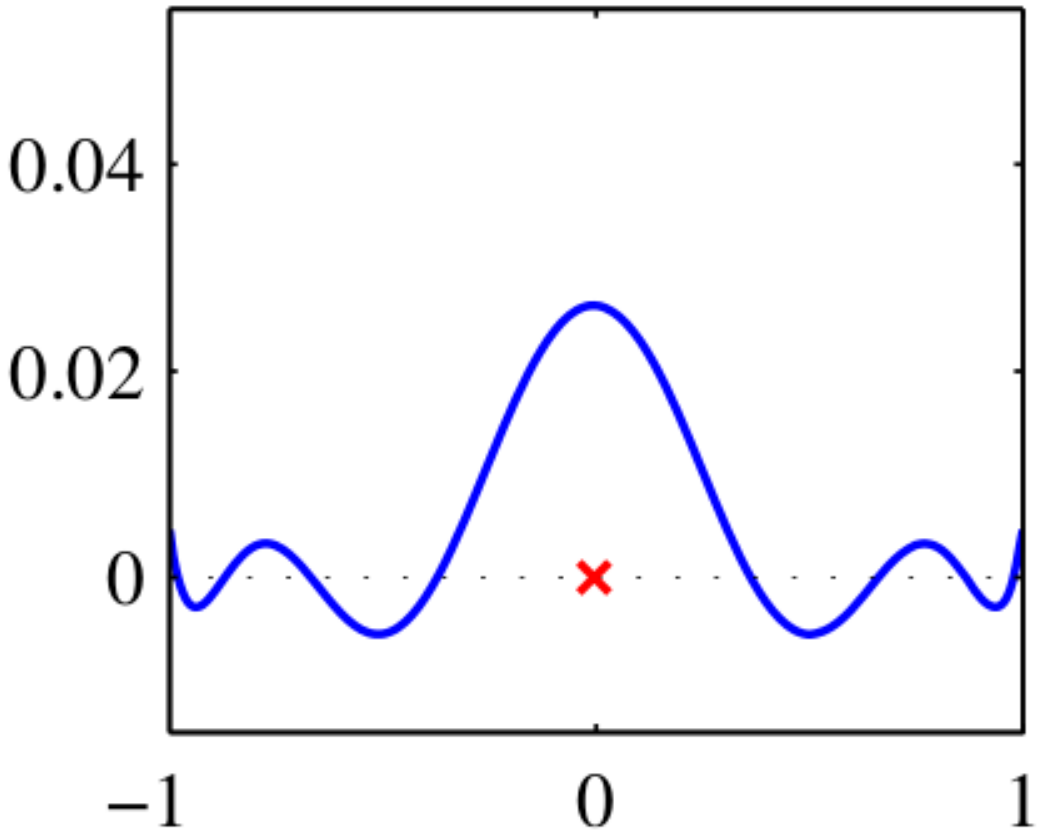
Эквивалентное ядро

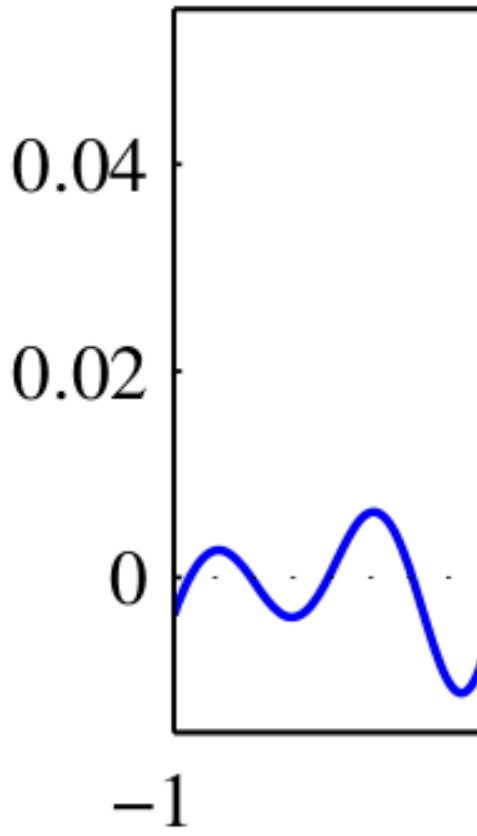


Эквивалентное ядро



Эквивалентное ядро





Выводы про эквивалентное ядро

- Эквивалентное ядро $k(\mathbf{x}, \mathbf{x}')$ локализовано вокруг $\mathbf{x}$ как функция $\mathbf{x}'$, т.е. каждая точка оказывает наибольшее влияние около себя и затухает потом.
- Можно было бы с самого начала просто определить ядро и предсказывать через него, безо всяких базисных функций ϕ – такой подход мы ещё будем рассматривать.

Упражнение. Докажите, что $\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) = 1$.

Ядерные методы

- То, что у нас получилось с линейной регрессией, — это частный случай очень общего подхода, другого взгляда на машинное обучение: *ядерных методов* (kernel methods).
- Предположим, что мы решаем задачу обучения с учителем для $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)\}$, где $x_n \in X$ и $y_n \in Y$; так выглядят и задачи регрессии, и задачи классификации.
- Цель машинного обучения — научиться предсказывать $y \in Y$ для данного $x \in X$ так, чтобы это было связано с имеющимися данными, то есть так, чтобы (и здесь появляется новый взгляд) пара (x, y) была *похожа* на пары (x_n, y_n), которые уже есть в D.

Ядерные методы

- Похожесть в пространстве Y обычно задать легко; например, в случае бинарной классификации Y и вовсе состоит из двух значений, например $Y = \{\pm 1\}$.
- Интересная часть здесь состоит в том, чтобы задать похожесть в пространстве X; для этого нам нужна некоторая функция похожести

$$k : X \times X \to \mathbb{R}, \quad (x, x') \mapsto k(x, x').$$

- Эта функция похожести и называется *ядром* (kernel).

Ядерные методы

- Главное техническое требование к функции k состоит в том, что она должна задавать скалярное произведение в некотором пространстве H, которое иногда называют *пространством признаков* (feature space). Иначе говоря,

$$k(x, x') = \Phi(x)^\top \Phi(x')$$

  для некоторого выделяющего признаки отображения $\Phi : X \to H$.

- Если это так, мы можем начать строить алгоритмы в пространстве признаков H, используя при этом функцию $k(x, x')$ вместо того, чтобы реально переходить в это пространство, а оно, скорее всего, будет иметь очень, очень высокую размерность.

Ядерные методы

- Например, предположим, что мы хотим построить простейший бинарный классификатор в пространстве H: найти центроиды обоих классов в данных и провести серединный перпендикуляр между ними. Это значит, что мы ищем

$$\mathbf{c}_1 = \frac{1}{N_1} \sum_{n:y_n=1} \Phi(x_n), \qquad \mathbf{c}_2 = \frac{1}{N_2} \sum_{n:y_n=-1} \Phi(x_n),$$

где $N_1 = \#\{n : y_n = 1\}$, $N_2 = \#\{n : y_n = -1\}$, а потом строим предсказания вида

$$y(x) = \operatorname{sign}\left(\mathbf{c}_1^\top \Phi(x) - \mathbf{c}_2^\top \Phi(x) - b\right), \quad \text{где} \quad b = \frac{\|\mathbf{c}_1\|^2 - \|\mathbf{c}_2\|^2}{2};$$

мы обсуждали такой подход, когда вели разговор о геометрической сути классификации.

- А теперь главное: давайте подставим $\mathbf{c}_1$ и $\mathbf{c}_2$ и распишем скалярное произведение по линейности:

$$y(x) = \text{sign}\left(\frac{1}{N_1}\sum_{n:y_n=1}\Phi(x_n)^\top\Phi(x) - \frac{1}{N_1}\sum_{n:y_n=1}\Phi(x_n)^\top\Phi(x) - b\right) =$$

$$= \text{sign}\left(\frac{1}{N_1}\sum_{n:y_n=1}k(x_n,x) - \frac{1}{N_1}\sum_{n:y_n=1}k(x_n,x) - b\right), \quad \text{где}$$

$$b = \frac{1}{2}\left(\frac{1}{N_1^2}\sum_{m,n:y_n=y_m=1}k(x_m,x_n) - \frac{1}{N_2^2}\sum_{m,n:y_n=y_m=-1}k(x_m,x_n)\right).$$

Ядерные методы

- Получилось, что мы обучили классификатор в пространстве H и выразили полученное правило классификации при помощи ядра $k(x, x')$, вообще *ни разу не записывая* ни одного вектора из H — все операции проводятся над элементами исходного пространства X.
- Это значит, что пространство H может иметь гигантскую размерность, а может быть и вовсе бесконечномерным!
- Этот эффект называется *ядерным трюком* (kernel trick)

Спасибо!

Спасибо за внимание!