

# Линейная регрессия

---

Сергей Николенко

Академия MADE — Mail.Ru

20 апреля 2020 г.

---

## *Random facts:*

- 20 апреля 1535 г. в Стокгольме наблюдали паргелий (гало в виде ложных солнц), запечатлённый на знаменитой картине *Vädersolstavlan*
- 20 апреля 1841 г. вышел в свет первый детективный рассказ — «Убийство на улице Морг» Эдгара По
- 20 апреля 1902 г. Мария и Пьер Кюри получили чистый хлорид радия
- 20 апреля 1918 г. Манфред фон Рихтгофен сбил два вражеских самолёта; на счету Красного барона они стали 79-м и 80-м и последними: на следующий день сбили уже его самого
- 20 апреля 2010 г. в Мексиканском заливе взорвалась буровая платформа *Deepwater Horizon*, что привело к крупнейшему разливу нефти в истории США и смерти по крайней мере 609 морских черепах и 100 дельфинов

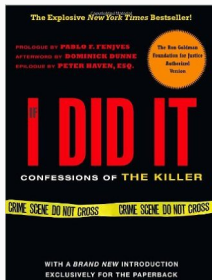
## Еще немного про вероятности

---

# Зачем нужны вероятностные модели

- Зачем нужны вероятностные модели? Апостериорные вероятности помогают:
  - добавить опцию «я не знаю»;
  - минимизировать риск, учесть разные веса ошибок;
  - перебалансировать классы или по-другому добавить априорные вероятности;
  - комбинировать модели (например, наивным Байесом)...
- Понимание смысла помогает:
  - понимать границы применимости, предположения, которые делают модели;
  - обобщать и переносить идеи моделей на другие задачи;
  - содержательно интерпретировать происходящее.

- (1) Прокурор указал, что O.J. Simpson уже бил жену в прошлом. Адвокат ответил: «Убивают только одну из 2500 женщин, подвергавшихся семейному насилию, так что это вообще нерелевантно». Суд согласился с адвокатом; верно ли это рассуждение?
- (2) У Салли Кларк погибли два младенца; прокурор указал, что вероятность двух случаев SIDS в одной семье, которую он получил из статистики одиночных случаев, — около 1 из 73 миллионов; в чём он не прав?



# Линейная регрессия

---

# Метод наименьших квадратов

- Линейная регрессия: рассмотрим линейную функцию

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^p x_j w_j = \mathbf{x}^\top \mathbf{w}, \quad \mathbf{x} = (1, x_1, \dots, x_p).$$

- Таким образом, по вектору входов  $\mathbf{x}^\top = (x_1, \dots, x_p)$  мы будем предсказывать выход  $y$  как

$$\hat{y}(\mathbf{x}) = \hat{w}_0 + \sum_{j=1}^p x_j \hat{w}_j = \mathbf{x}^\top \hat{\mathbf{w}}.$$

# Метод наименьших квадратов

- Как найти оптимальные параметры  $\hat{\mathbf{w}}$  по тренировочным данным вида  $(\mathbf{x}_i, y_i)_{i=1}^N$ ?
- Метод наименьших квадратов: будем минимизировать

$$\text{RSS}(\mathbf{w}) = \sum_{i=1}^N (y_i - \mathbf{x}_i^\top \mathbf{w})^2.$$

- Как минимизировать?

# Метод наименьших квадратов

- Можно на самом деле решить задачу точно – записать как

$$\text{RSS}(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}),$$

где  $\mathbf{X}$  – матрица  $N \times p$ , продифференцировать по  $\mathbf{w}$ , получится

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

если матрица  $\mathbf{X}^\top \mathbf{X}$  невырожденная.

- Замечание:  $(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$  называется *псевдообратной матрицей Мура–Пенроуза* (Moore–Penrose pseudo-inverse) матрицы  $\mathbf{X}$ ; это обобщение понятия обратной матрицы на неквадратные матрицы.



# Байесовская регрессия

- Теперь давайте поговорим о линейной регрессии по-байесовски.
- Основное наше предположение – в том, что шум (ошибка в данных) распределён нормально, т.е. переменная  $t$ , которую мы наблюдаем, получается как

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Иными словами,

$$p(t \mid \mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}(t \mid y(\mathbf{x}, \mathbf{w}), \sigma^2).$$

- Здесь пока  $y$  – любая функция.

- Чтобы не повторять совсем уж то же самое, мы рассмотрим не в точности линейную регрессию, а её естественное обобщение – линейную модель с базисными функциями:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})$$

( $M$  параметров,  $M - 1$  базисная функция,  $\phi_0(\mathbf{x}) = 1$ ).

- Базисные функции  $\phi_i$  – это, например:
  - результат feature extraction;
  - расширение линейной модели на нелинейные зависимости (например,  $\phi_j(x) = x^j$ );
  - локальные функции, которые существенно не равны нулю только в небольшой области (например, гауссовские базисные функции  $\phi_j(\mathbf{x}) = e^{-\frac{(\mathbf{x}-\mu_j)^2}{2s^2}}$ );
  - ...

# Байесовская регрессия

- Рассмотрим набор данных  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  со значениями  $\mathbf{t} = \{t_1, \dots, t_N\}$ .
- Будем предполагать, что данные взяты независимо по одному и тому же распределению:

$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^N \mathcal{N}(t_n \mid \mathbf{w}^\top \phi(\mathbf{x}_n), \sigma^2) .$$

- Прологарифмируем (опустим  $\mathbf{X}$ , т.к. по нему всегда условная вероятность будет):

$$\ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2 .$$

# Байесовская регрессия

- Прологарифмируем (опустим  $\mathbf{X}$ , т.к. по нему всегда условная вероятность будет):

$$\ln p(\mathbf{t} | \mathbf{w}, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2.$$

- И вот мы получили, что для максимизации правдоподобия по  $\mathbf{w}$  нам нужно как раз минимизировать среднеквадратичную ошибку!

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t} | \mathbf{w}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n)) \phi(\mathbf{x}_n).$$

- Решая систему уравнений  $\nabla \ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = 0$ , получаем то же самое, что и раньше:

$$\mathbf{w}_{ML} = \left( \Phi^\top \Phi \right)^{-1} \Phi^\top \mathbf{t}.$$

- Здесь  $\Phi = (\phi_j(\mathbf{x}_i))_{i,j}$ .

- Теперь можно и относительно  $\sigma^2$  максимизировать правдоподобие; получим

$$\sigma_{ML}^2 = \frac{1}{N} \sum_{n=1}^N (t_n - \mathbf{w}_{ML}^T \phi(\mathbf{x}_n))^2,$$

т.е. как раз выборочная дисперсия имеющихся данных вокруг предсказанного значения.

# Оверфиттинг в линейной регрессии

---



# Полиномиальная аппроксимация

- Мы говорили о регрессии с базисными функциями:

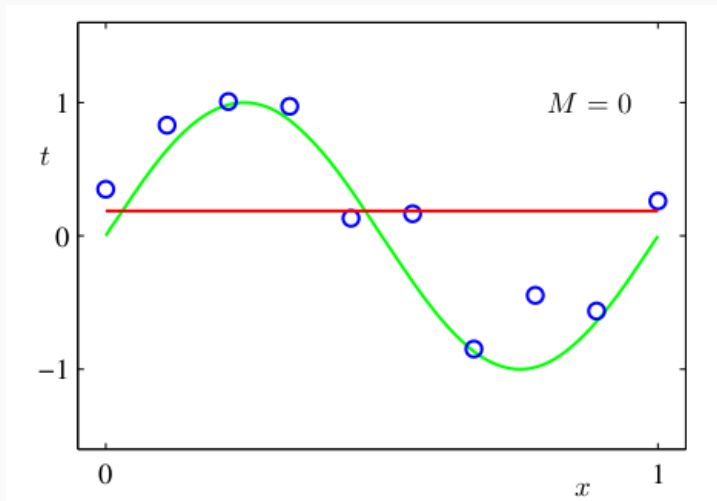
$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^M w_j \phi_j(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}).$$

- Давайте для примера рассмотрим такую регрессию для  $\phi_j(x) = x^j$ , т.е.

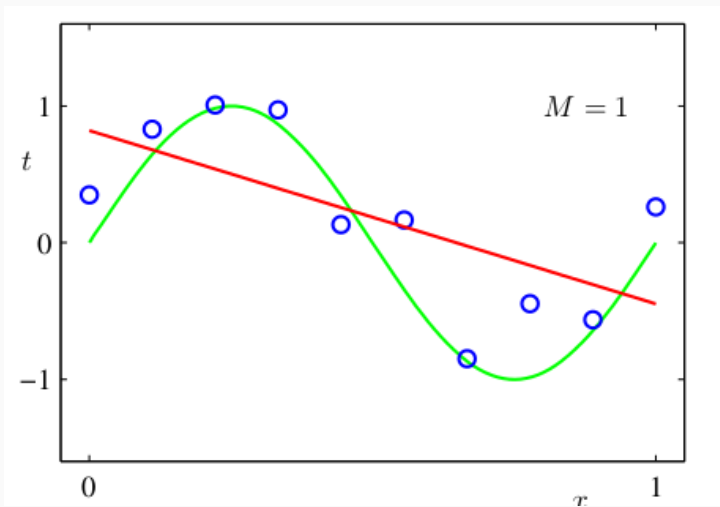
$$f(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M.$$

- И будем, как раньше, минимизировать квадратичную ошибку.
- Пример с кодом.

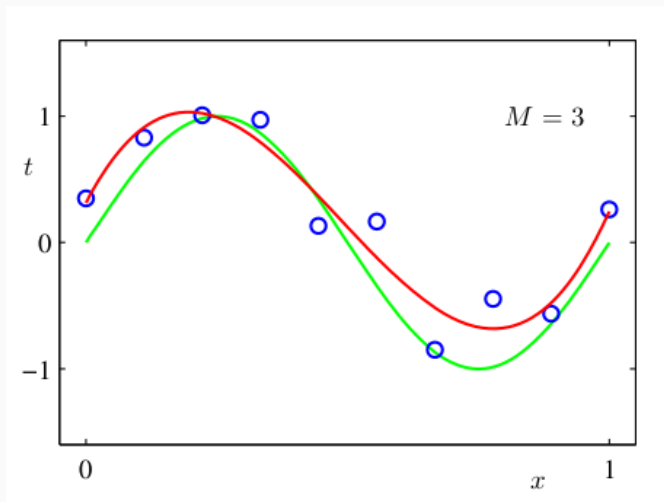
# Полиномиальная аппроксимация



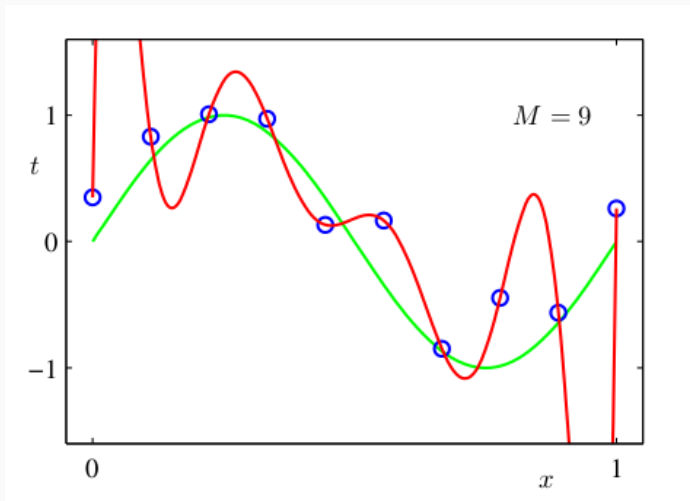
# Полиномиальная аппроксимация



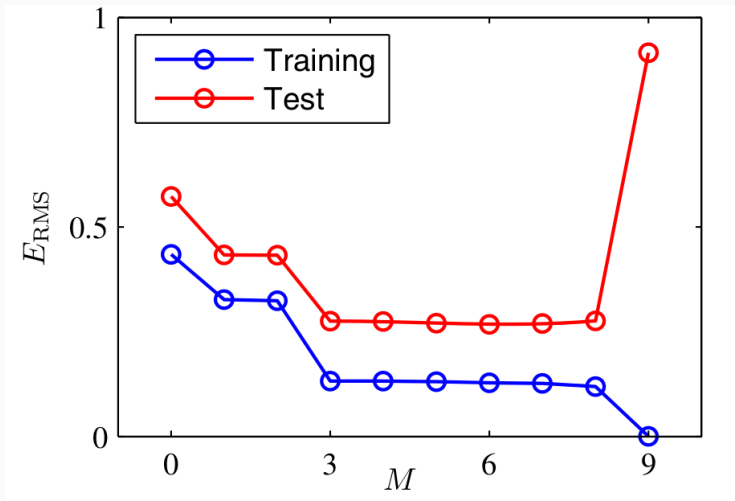
# Полиномиальная аппроксимация



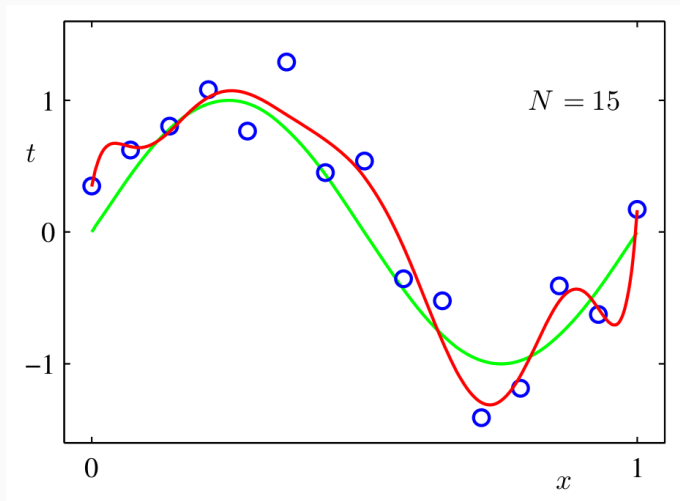
# Полиномиальная аппроксимация



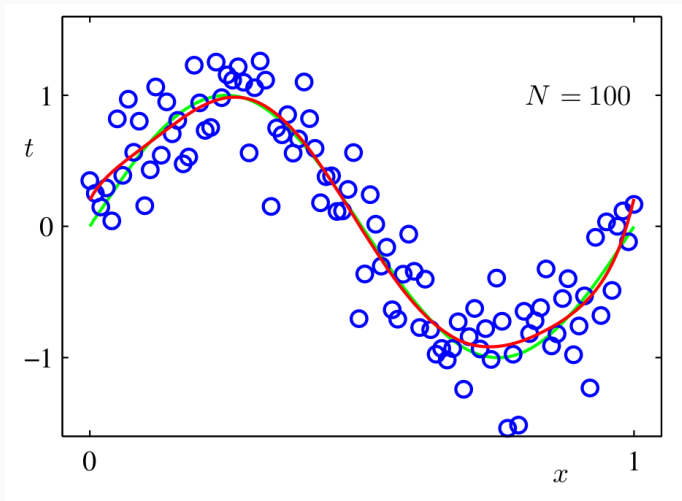
## Значения RMS



Можно собрать больше данных...



Можно собрать больше данных...





## Значения коэффициентов

	$M = 0$	$M = 1$	$M = 6$	$M = 9$
$w_0^*$	0.19	0.82	0.31	0.35
$w_1^*$		-1.27	7.99	232.37
$w_2^*$			-25.43	-5321.83
$w_3^*$			17.37	48568.31
$w_4^*$				-231639.30
$w_5^*$				640042.26
$w_6^*$				-1061800.52
$w_7^*$				1042400.18
$w_8^*$				-557682.99
$w_9^*$				125201.43

- Итак, мы увидели, что даже в линейной регрессии может наступить оверфиттинг.
- Что же делать?..

# Регуляризация в линейной регрессии

---

- Итак, получается, что у нас сильно растут коэффициенты.
- Давайте попробуем с этим бороться. Бороться будем прямолинейно и простодушно: возьмём и добавим размер коэффициентов в функцию ошибки.

- Было (для тестовых примеров  $\{(x_i, y_i)\}_{i=1}^N$ ):

$$\text{RSS}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N (f(x_i, \mathbf{w}) - y_i)^2.$$

- Стало:

$$\text{RSS}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N (f(x_i, \mathbf{w}) - y_i)^2 + \frac{\alpha}{2} \|\mathbf{w}\|^2,$$

где  $\alpha$  – коэффициент регуляризации (его надо будет как-нибудь выбрать).

- Как оптимизировать эту функцию ошибки?

- Да точно так же – запишем как

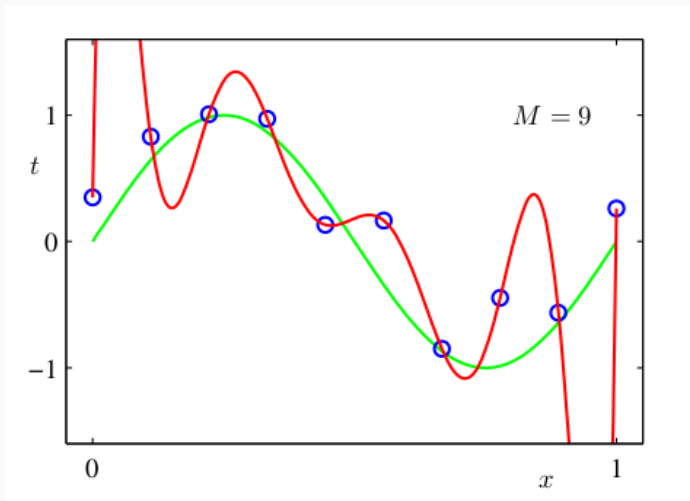
$$\text{RSS}(\mathbf{w}) = \frac{1}{2} (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}) + \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w}$$

и возьмём производную; получится

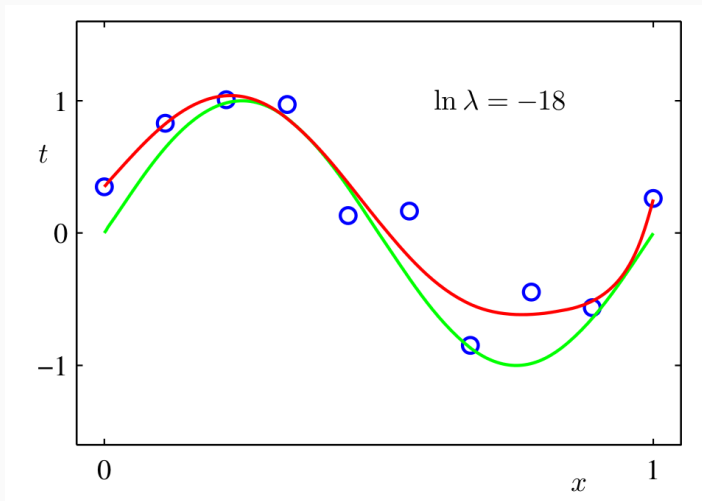
$$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X} + \alpha \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y}.$$

- Это *гребневая регрессия* (ridge regression); кстати, добавление  $\alpha \mathbf{I}$  к матрице неполного ранга делает её обратимой; это и есть *регуляризация*, и это и было исходной мотивацией для гребневой регрессии.

# Гребневая регрессия: $\ln \alpha = -\infty$

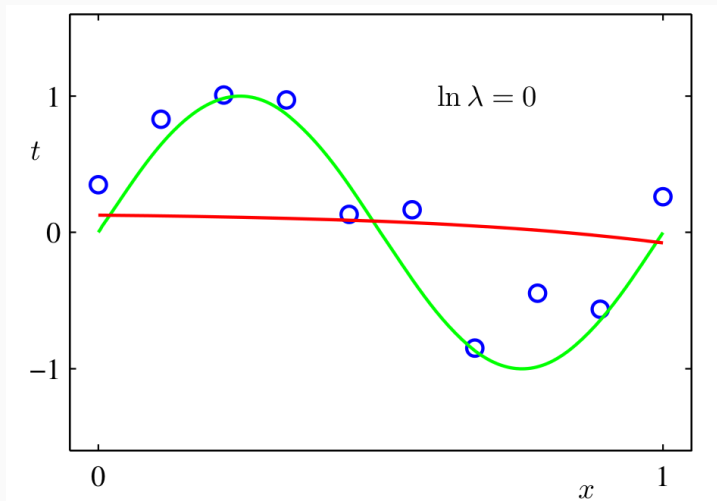


## Гребневая регрессия: $\ln \alpha = -18$





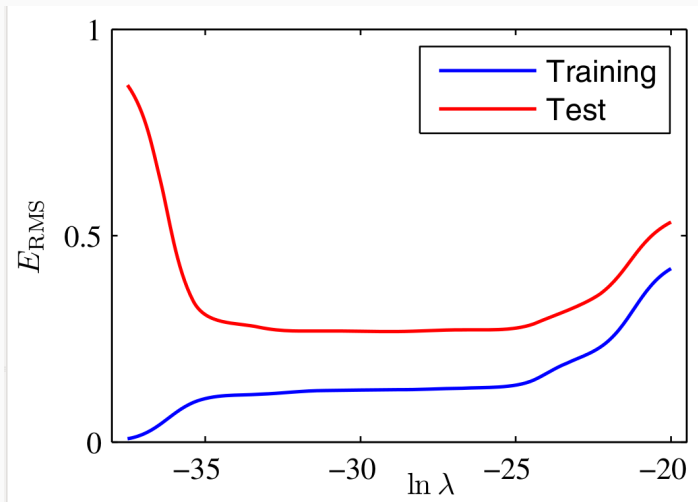
## Гребневая регрессия: $\ln \alpha = 0$



## Гребневая регрессия: коэффициенты

	$\ln \lambda = -\infty$	$\ln \lambda = -18$	$\ln \lambda = 0$
$w_0^*$	0.35	0.35	0.13
$w_1^*$	232.37	4.74	-0.05
$w_2^*$	-5321.83	-0.77	-0.06
$w_3^*$	48568.31	-31.97	-0.05
$w_4^*$	-231639.30	-3.89	-0.03
$w_5^*$	640042.26	55.28	-0.02
$w_6^*$	-1061800.52	41.32	-0.01
$w_7^*$	1042400.18	-45.95	-0.00
$w_8^*$	-557682.99	-91.53	0.00
$w_9^*$	125201.43	72.68	0.01

## Гребневая регрессия: RMS



- Почему именно так? Почему именно  $\frac{\alpha}{2} \|\mathbf{w}\|^2$ ?
- Мы сейчас ответим на этот вопрос, но, вообще говоря, это не обязательно.
- Лассо-регрессия (lasso regression) регуляризует  $L_1$ -нормой, а не  $L_2$ :

$$\text{RSS}(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N (f(x_i, \mathbf{w}) - y_i)^2 + \alpha \sum_{j=0}^M |w_j|.$$

- Есть и другие типы; об этом будем говорить позже.

# Регрессия по-байесовски

---

- А теперь давайте посмотрим на регрессию с совсем байесовской стороны.
- Напомним основу байесовского подхода:
  1. найти апостериорное распределение на гипотезах/параметрах:

$$p(\theta \mid D) \propto p(D|\theta)p(\theta)$$

(и/или найти максимальную апостериорную гипотезу  $\arg \max_{\theta} p(\theta \mid D)$ );

2. найти апостериорное распределение исходов дальнейших экспериментов:

$$p(x \mid D) \propto \int_{\theta \in \Theta} p(x \mid \theta) p(D|\theta) p(\theta) d\theta.$$

# Байесовская регуляризация

- В нашем рассмотрении пока не было никаких априорных распределений.
- Давайте какое-нибудь введём; например, нормальное (почему так – позже):

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0).$$

- Рассмотрим набор данных  $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  со значениями  $\mathbf{t} = \{t_1, \dots, t_N\}$ . В этой модели мы предполагаем, что данные независимы и одинаково распределены:

$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^N \mathcal{N}(t_n \mid \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \sigma^2).$$

- Тогда наша задача – посчитать

$$\begin{aligned} p(\mathbf{w} \mid \mathbf{t}) &\propto p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \sigma^2) p(\mathbf{w}) \\ &= \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0) \prod_{n=1}^N \mathcal{N}(t_n \mid \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \sigma^2). \end{aligned}$$

- Давайте подсчитаем.



- Получится

$$\begin{aligned}p(\mathbf{w} \mid \mathbf{t}) &= \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N), \\ \boldsymbol{\mu}_N &= \boldsymbol{\Sigma}_N \left( \boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \frac{1}{\sigma^2} \boldsymbol{\Phi}^\top \mathbf{t} \right), \\ \boldsymbol{\Sigma}_N &= \left( \boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \right)^{-1}.\end{aligned}$$

- Теперь давайте подсчитаем логарифм правдоподобия.

- Если мы возьмём априорное распределение около нуля:

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \mathbf{0}, \frac{1}{\alpha} I),$$

то логарифм правдоподобия получится

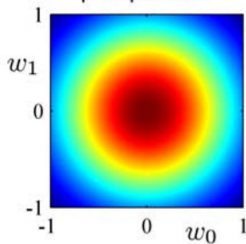
$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2 - \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} + \text{const},$$

то есть в точности гребневая регрессия.

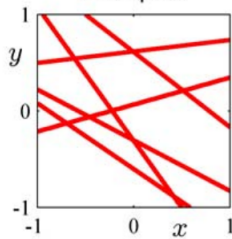
# Пример

likelihood

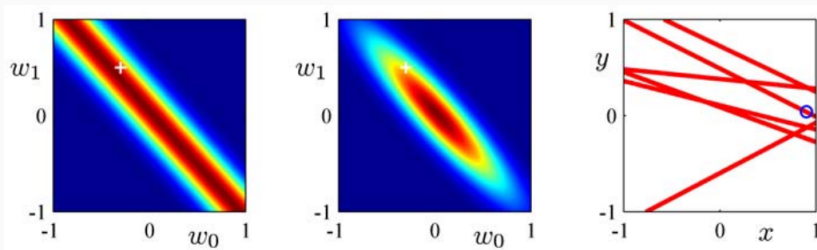
prior/posterior



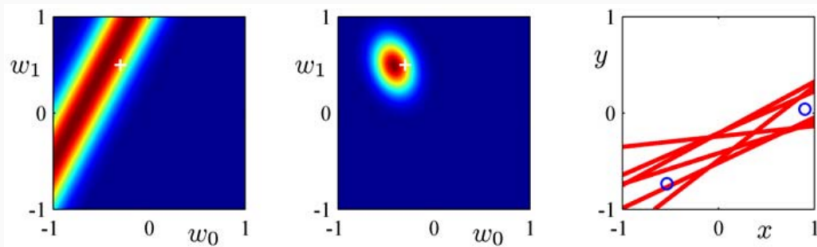
data space



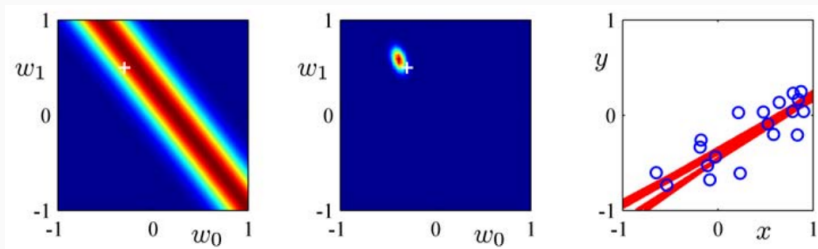
# Пример



# Пример



# Пример



- Теперь давайте рассмотрим лассо-регрессию:

$$L(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N (f(x_i, \mathbf{w}) - y_i)^2 + \lambda \sum_{j=0}^p |w_j|.$$

- Главное отличие – теперь форма ограничений (т.е. форма априорного распределения) такова, что весьма вероятно получить строго нулевые  $w_j$ .
- Кстати, что значит «форма ограничений»?

- Мы можем переписать регрессию с регуляризатором по-другому:

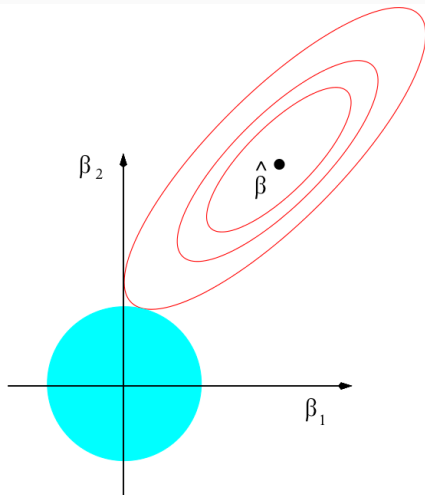
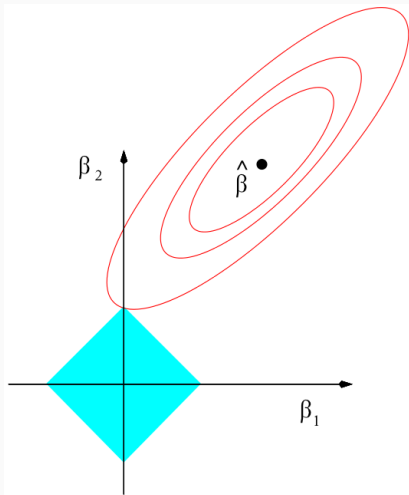
$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \left\{ \frac{1}{2} \sum_{i=1}^N (f(x_i, \mathbf{w}) - y_i)^2 + \lambda \sum_{j=0}^p |w_j| \right\},$$

эквивалентно

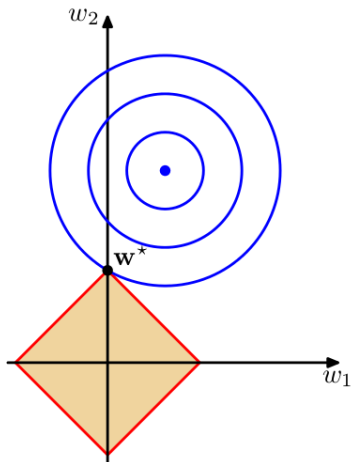
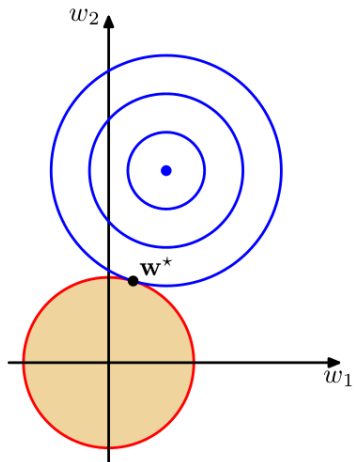
$$\mathbf{w}^* = \arg \min_{\mathbf{w}} \left\{ \frac{1}{2} \sum_{i=1}^N (f(x_i, \mathbf{w}) - y_i)^2 \right\} \text{ при } \sum_{j=0}^p |w_j| \leq t.$$

**Упражнение.** Докажите это.





# Гребень и лассо

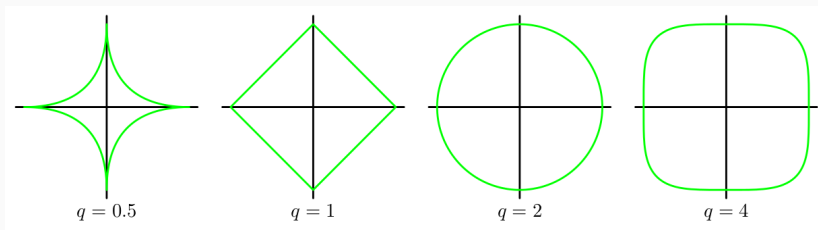
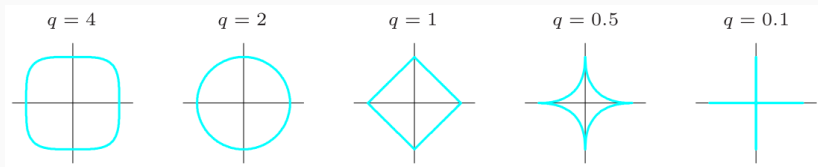


- Можно рассмотреть обобщение гребневой и лассо-регрессии:

$$L(\mathbf{w}) = \frac{1}{2} \sum_{i=1}^N (f(x_i, \mathbf{w}) - y_i)^2 + \lambda \sum_{j=0}^p (|w_j|)^q.$$

**Упражнение.** Какому априорному распределению на параметры  $\mathbf{w}$  соответствует эта задача?

# Разные $q$



Спасибо!

Спасибо за внимание!