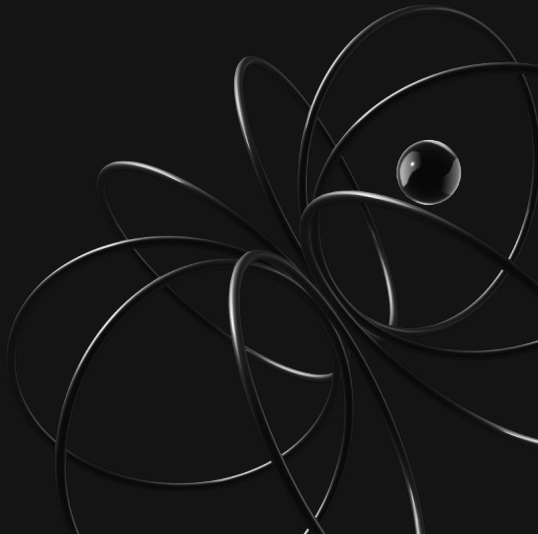


Линейная регрессия. Байесовский вывод в линейной регрессии



Машинное обучение
Основы байесовского вывода

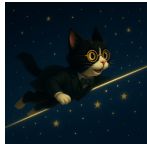
03



Линейная регрессия

Random facts:

- 18 сентября 1066 г. Гаральд Гардерада вместе с Тостигом Годвинсоном высадились в устье реки Хамбер; через два дня он разгромил английские войска в битве у Фулфорда, но через пять дней потерпел сокрушительное поражение при Стэмфорд-Бридж.
- 18 сентября 1698 г. в Бастилию был переведён таинственный узник под номером 64489001, известный как «Железная маска»; после его смерти в 1703 г. всю мебель и одежду специально уничтожили, закрасив стены и расплавив все металлические вещи.
- 18 сентября 1830 г. лошадь победила в гонке на 9 миль от Райлиз Таверн до Балтимора между лошастью и первым американским паровозом, а 18 сентября 1893 г. было завершено строительство Великой Северной железной дороги между Миссисипи и Тихим океаном.
- 18 сентября 1870 г. Генри Уошберн открыл гейзер Old Faithful, который впоследствии стал источником одного из самых известных датасетов для кластеризации.
- 18 сентября 1886 г. поэт Жан Мореас опубликовал в газете «Фигаро» статью, ставшую манифестом литераторов-символистов; сами термины «символист» и «символизм» были предложены для замены слова «декадент», которую раньше применяли критики для Бодлера и Малларме.



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

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

- Как минимизировать?

Метод наименьших квадратов

- Можно на самом деле решить задачу точно – записать как

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

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.
- Много ли нужно точек, чтобы обучить такую модель?

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

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

Иными словами,

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

- Здесь пока y – любая функция.

Байесовская регрессия

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, $\phi_0(\mathbf{x}) = 1$).

Байесовская регрессия

Базисные функции ϕ_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 $\phi_j(\mathbf{x}) = e^{-\frac{(x-\mu_j)^2}{2s^2}}$);
- ...

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:

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

- 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!

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

Байесовская регрессия

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

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

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

Байесовская регрессия

Теперь можно и относительно σ^2 максимизировать правдоподобие; получим

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

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

Полиномиальная аппроксимация

- Мы говорили о регрессии с базисными функциями:

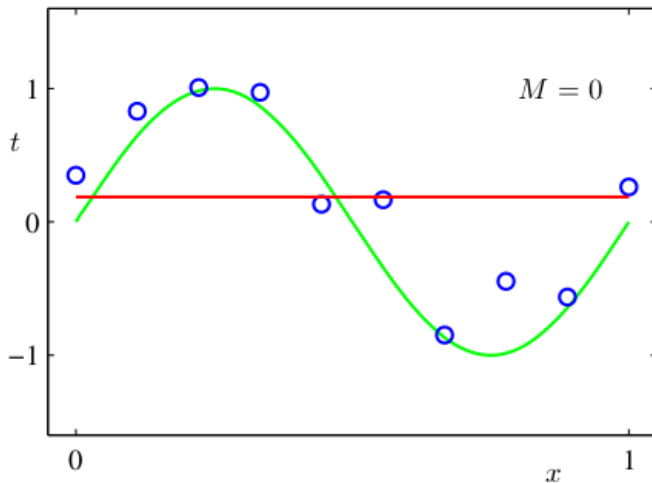
$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^M w_j \phi_j(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}).$$

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

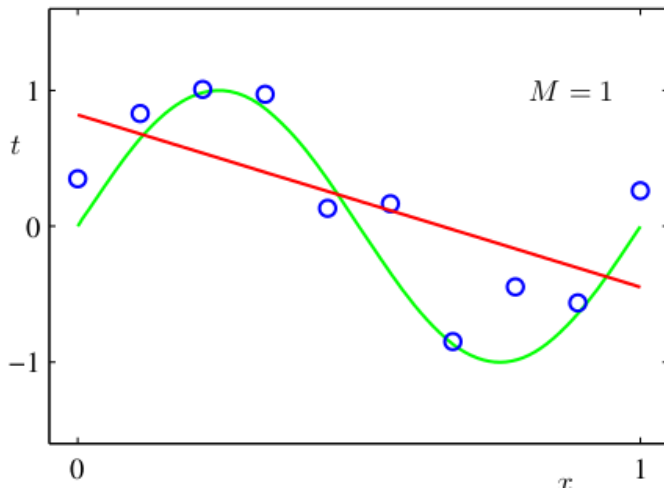
$$f(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M.$$

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

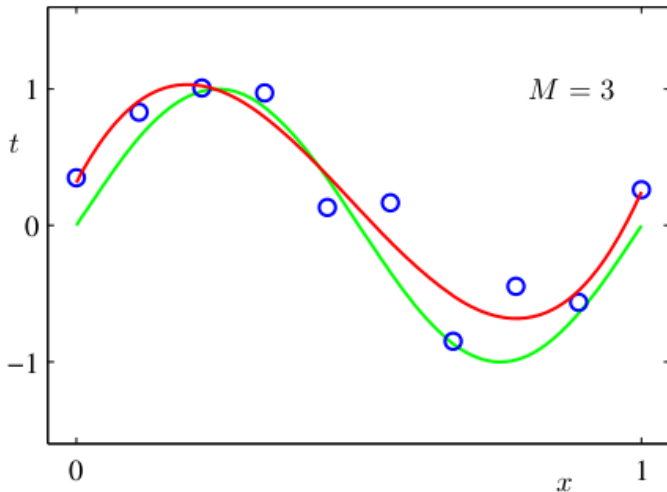
Полиномиальная аппроксимация



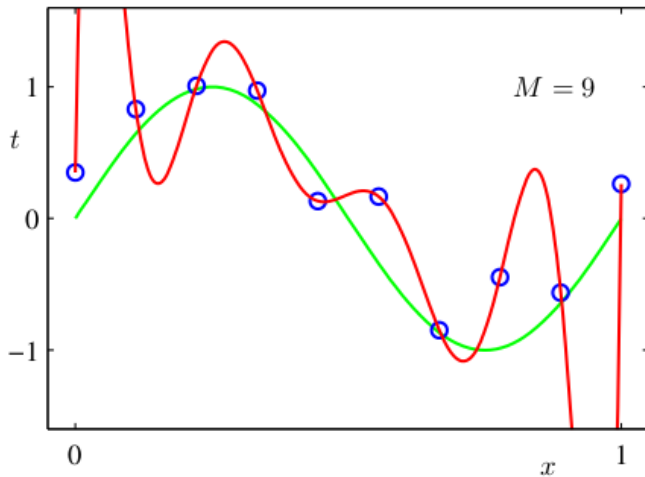
Полиномиальная аппроксимация



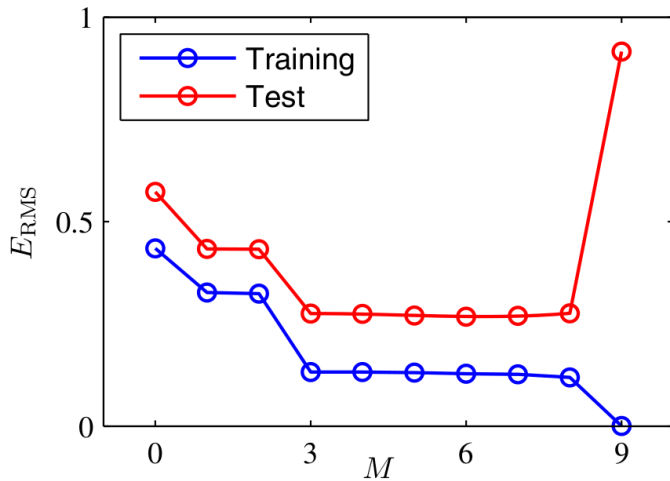
Полиномиальная аппроксимация



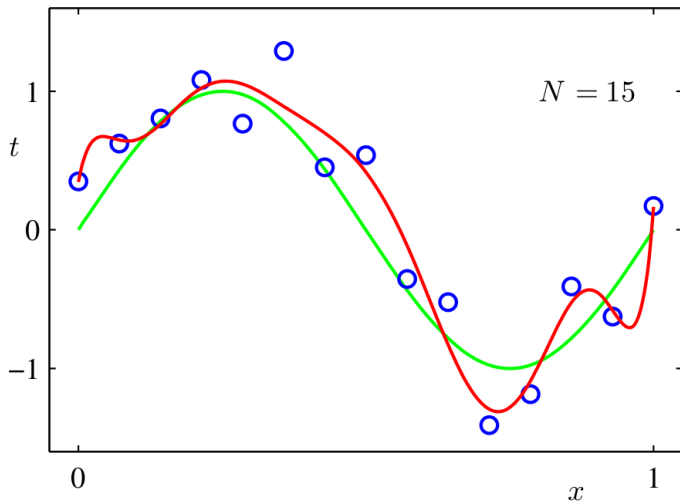
Полиномиальная аппроксимация



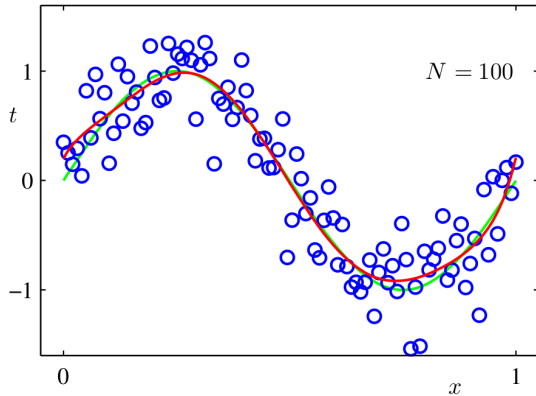
Значения RMS



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

Регуляризация по-байесовски

Random facts:

- 25 сентября 1066 г. Гаральд Гардерада вместе с Тостигом Годвинсоном потерпел сокрушительное поражение при Стэмфорд-Бридж, и попытка норвежского завоевания Англии провалилась
- 25 сентября 1303 г. произошло одно из самых смертоносных землетрясений в истории Земли: в Хундуне, в провинции Шаньси; погибли более 200 тысяч человек, землетрясение сровняло с землёй горы и холмы
- 25 сентября 1493 г. Христофор Колумб отправился в своё второе путешествие к Америке, а 25 сентября 1513 г. Васко де Бальбоа со своим отрядом пересёк Панамский перешеек и стал первым европейцем, достигшим Тихого океана



Регуляризация по-байесовски

- 25 сентября 1818 г. английский врач Джеймс Бландел впервые провёл операцию по переливанию крови от человека к человеку
- 25 сентября 1962 г. Фидель Кастро заявил, что СССР намерен создать на Кубе базу для своего флота; рыболовного, разумеется
- 25 сентября 1968 г. в первый и пока единственный раз британский хит-парад возглавила русская песня — романс «Дорогой длиною»; правда, называлась она «Those Were the Days» и исполнялась Мэри Хопкин со словами Джина Раскина

Байесовская регуляризация

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

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

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

- (2) найти апостериорное распределение исходов дальнейших экспериментов:

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

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

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

Байесовская регуляризация

- Тогда наша задача — посчитать

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

- Давайте подсчитаем.

Байесовская регуляризация

- Получится

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

$$\boldsymbol{\mu}_N = \boldsymbol{\Sigma}_N \left(\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \frac{1}{\sigma^2} \boldsymbol{\Phi}^\top \mathbf{t} \right),$$

$$\boldsymbol{\Sigma}_N = \left(\boldsymbol{\Sigma}_0^{-1} + \frac{1}{\sigma^2} \boldsymbol{\Phi}^\top \boldsymbol{\Phi} \right)^{-1}.$$

- Теперь давайте подсчитаем логарифм правдоподобия.

Байесовская регуляризация

- Если мы возьмём априорное распределение около нуля:

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

то логарифм правдоподобия получится

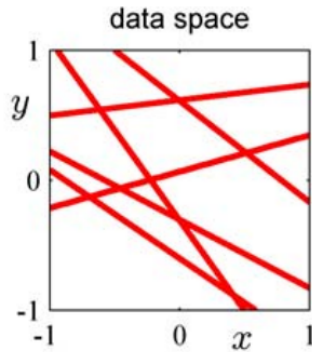
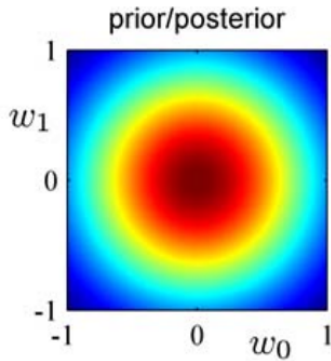
$$\ln p(\mathbf{w} \mid \mathbf{t}) = -\frac{1}{2\sigma^2} \sum_{n=1}^N (t_n - \mathbf{w}^\top \phi(\mathbf{x}_n))^2 - \frac{\alpha}{2} \mathbf{w}^\top \mathbf{w} + \text{const},$$

то есть в точности гребневая регрессия.

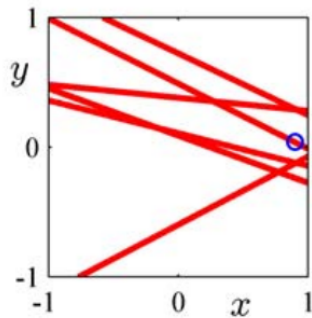
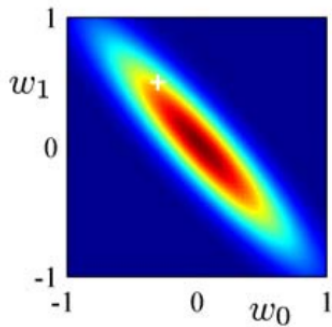
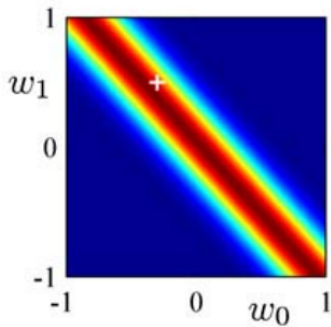
Пример



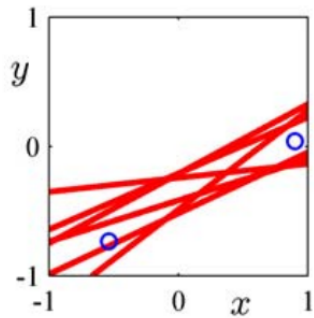
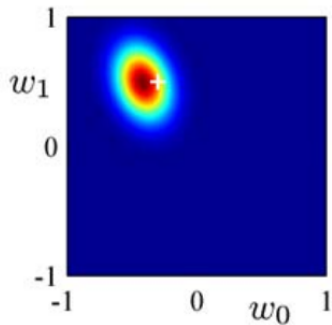
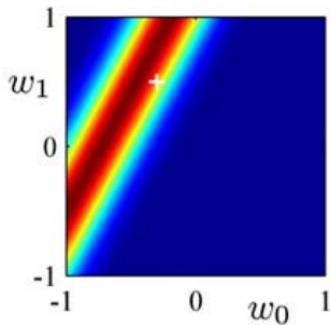
likelihood



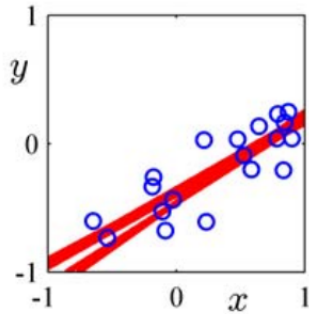
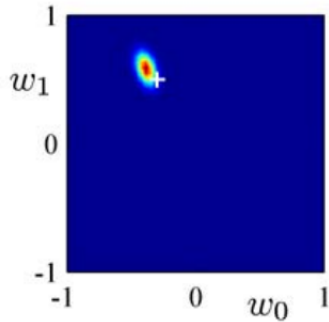
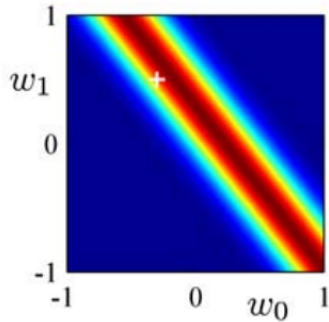
Пример



Пример



Пример



Обобщение

- Можно слегка обобщить — рассмотреть априорное распределение более общего вида

$$p(\mathbf{w} \mid \alpha) = \left[\frac{q}{2} \left(\frac{\alpha}{2} \right)^{1/q} \frac{1}{\Gamma(1/q)} \right]^M e^{-\frac{\alpha}{2} \sum_{j=1}^M |w_j|^q}.$$



Домашнее задание. Подсчитайте логарифм правдоподобия.

Лассо

- 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»?

Лассо

- Мы можем переписать регрессию с регуляризатором по-другому:

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

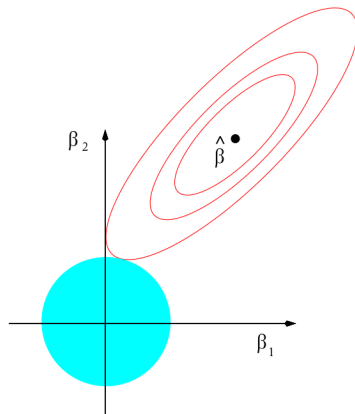
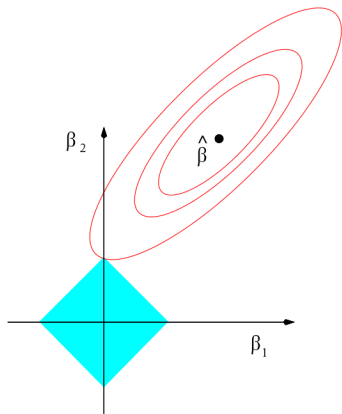
эквивалентно

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

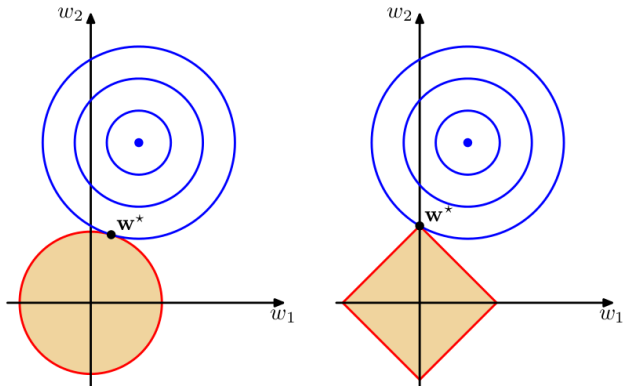


Домашнее задание. Докажите это.

Гребень и лассо



Гребень и лассо



Обобщение

- Можно рассмотреть обобщение гребневой и лассо-регрессии:

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

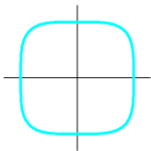


Домашнее задание. Какому априорному распределению на параметры $\mathbf{w}$ соответствует эта задача?

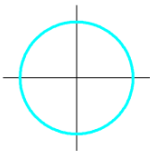
Разные q



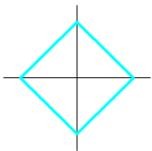
$q = 4$



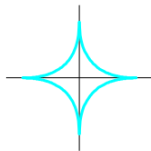
$q = 2$



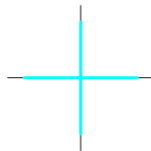
$q = 1$



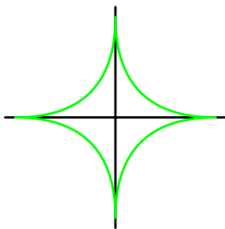
$q = 0.5$



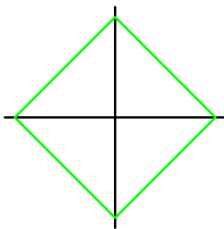
$q = 0.1$



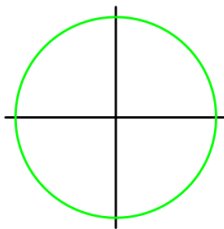
$q = 0.5$



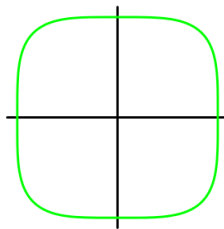
$q = 1$



$q = 2$



$q = 4$



Предсказание в линейной регрессии

- Теперь давайте вернёмся к байесовской постановке:

(1) найти апостериорное распределение на гипотезах/параметрах:

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

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

(2) найти апостериорное распределение исходов дальнейших экспериментов:

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

Предсказание в линейной регрессии

- В прошлый раз мы нашли апостериорное распределение: для гауссовского априорного

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

мы нашли

$$\begin{aligned} p(\mathbf{w} \mid \mathbf{t}, \alpha, \beta) &= \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N), \\ \boldsymbol{\mu}_N &= \boldsymbol{\Sigma}_N (\boldsymbol{\Sigma}_0^{-1} \boldsymbol{\mu}_0 + \beta \boldsymbol{\Phi}^\top \mathbf{t}), \\ \boldsymbol{\Sigma}_N &= (\boldsymbol{\Sigma}_0^{-1} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi})^{-1}, \end{aligned}$$

где $\beta = \frac{1}{\sigma^2}$ (precision нормального распределения).

Предсказание в линейной регрессии

- Теперь сделаем следующий шаг — найдём апостериорное распределение наших предсказаний

$$p(t \mid \mathbf{t}, \alpha, \beta) = \int p(t \mid \mathbf{w}, \beta) p(\mathbf{w} \mid \mathbf{t}, \alpha, \beta) d\mathbf{w}.$$

- Это свёртка двух гауссианов, и получается...

Предсказание в линейной регрессии

- ...тоже гауссиан:

$$p(t \mid \mathbf{t}, \alpha, \beta) = \mathcal{N}(t \mid \boldsymbol{\mu}_N^\top \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2),$$

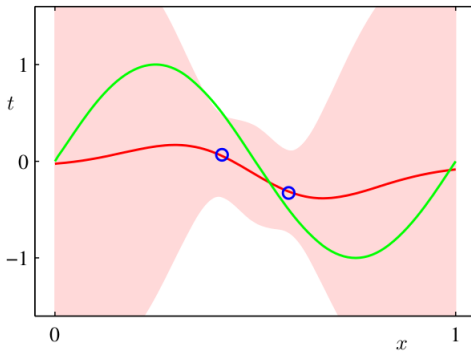
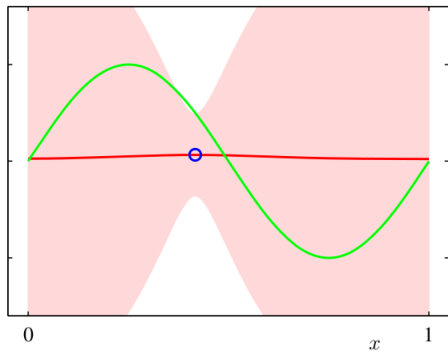
$$\text{где } \sigma_N^2 = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^\top \boldsymbol{\Sigma}_N \boldsymbol{\phi}(\mathbf{x}).$$

- Т.е. дисперсия складывается из шума в данных β и дисперсии параметров $\mathbf{w}$; гауссианы независимы, и их дисперсии просто складываются.

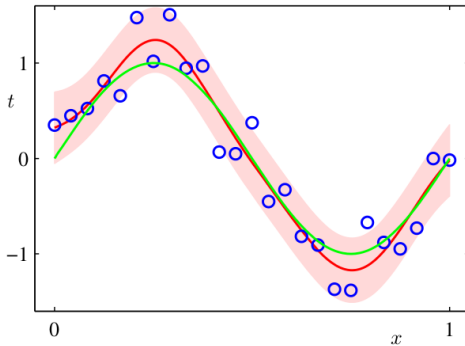
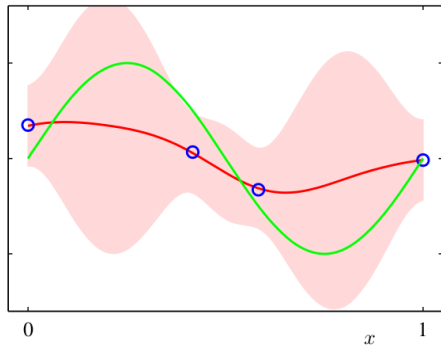


Домашнее задание. Оценка всё время уточняется: $\sigma_{N+1}^2 \leq \sigma_N^2$.

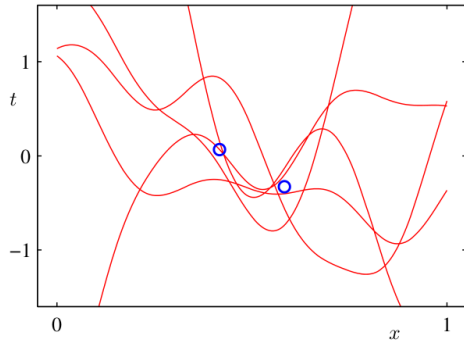
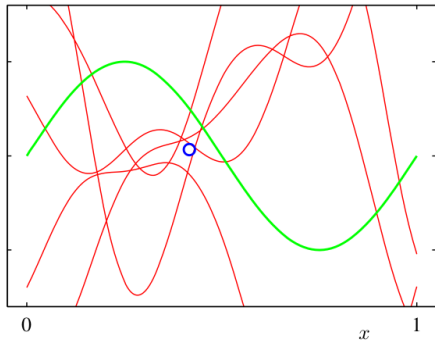
Предсказания



Предсказания



Предсказания



Предсказания

