

БАЙЕСОВСКИЙ ВЫВОД ДЛЯ МОНЕТКИ

Сергей Николенко

СПбГУ — Санкт-Петербург

19 сентября 2024 г.

Random facts:



- 19 сентября — Международный пиратский день, точнее, International Talk Like a Pirate Day, учреждённый в 1995 году и отмечающийся пастафарианцами; aaarrrrgh, maties!
- 19 сентября 1648 г. в Квебеке Жак Буадон открыл первую в Канаде таверну, а 19 сентября 1654 г. был зарегистрирован первый в Канаде брак (невесте, кстати, было 11 лет)
- 19 сентября 1783 г. в присутствии короля Людовика XVI монгольфьер унёс из Версаля барана, петуха и утку на 4 километра
- 19 сентября 1888 в Спа (Бельгия) прошёл первый в истории конкурс красоты: 21 претендентка отбиралась по фотографиям, участницы конкурса жили в отдельном доме и приезжали на соревнования в закрытом экипаже; победила 18-летняя креолка из Гваделупы Берта Сукаре
- 19 сентября 1990 г. Ассоциация пользователей UNIX зарегистрировала домен верхнего уровня .su для применения на территории СССР
- 19 сентября 1991 г. два немецких туриста в седловине Хауслабьох в Тироле нашли Этци, ледяную мумию человека, жившего около 5300 лет назад; Этци было около 45 лет (много!), при нём были медный топор, каменный нож и 182-сантиметровый тисовый лук

БАЙЕСОВСКИЙ ВЫВОД ДЛЯ МОНЕТКИ

- Мы остановились на том, что в статистике обычно ищут *гипотезу максимального правдоподобия* (maximum likelihood):

$$\theta_{ML} = \arg \max_{\theta} p(D | \theta).$$

- В байесовском подходе ищут *апостериорное распределение* (posterior)

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

и, возможно, *максимальную апостериорную гипотезу* (maximum a posteriori):

$$\theta_{MAP} = \arg \max_{\theta} p(\theta | D) = \arg \max_{\theta} p(D | \theta)p(\theta).$$

ПОСТАНОВКА ЗАДАЧИ

- Простая задача вывода: дана нечестная монетка, она подброшена N раз, имеется последовательность результатов падения монетки. Надо определить её «нечестность» и предсказать, чем она выпадет в следующий раз.

ПОСТАНОВКА ЗАДАЧИ

- Если у нас есть вероятность $p_h = \theta$ того, что монетка выпадет орлом (вероятность решки $p_t = 1 - \theta$), то вероятность того, что выпадет последовательность s , которая содержит n орлов и m решек, равна

$$p(s|p_h = \theta) = \theta^n(1 - \theta)^m.$$

- Сделаем предположение: будем считать, что вероятность монетки до экспериментов для нас распределена равномерно, т.е. у нас нет априорных предпочтений насчёт θ .
- Теперь нужно использовать теорему Байеса и вычислить скрытые параметры.

ПРИМЕР ПРИМЕНЕНИЯ ТЕОРЕМЫ БАЙЕСА

- Правдоподобие: $p(\theta|s) = \frac{p(s|\theta)p(\theta)}{p(s)}$.
- Здесь $p(\theta)$ следует понимать как непрерывную случайную величину, сосредоточенную на интервале $[0, 1]$, коей она и является. Наше предположение о равномерном распределении в данном случае значит, что априорная вероятность $p(\theta) = 1, \theta \in [0, 1]$ (т.е. априори мы не знаем, насколько нечестна монетка, и предполагаем это равновероятным). А $p(s|\theta)$ мы уже знаем.
- Итого получается:

$$p(\theta|s) = \frac{\theta^n(1-\theta)^m}{p(s)}.$$

- Итого получается:

$$p(\theta|s) = \frac{\theta^n(1-\theta)^m}{p(s)}.$$

- $p(s)$ можно подсчитать как

$$\begin{aligned} p(s) &= \int_0^1 \theta^n(1-\theta)^m d\theta = \\ &= \frac{\Gamma(n+1)\Gamma(m+1)}{\Gamma(n+m+2)} = \frac{n!m!}{(n+m+1)!}, \end{aligned}$$

но найти $\arg \max_{\theta} p(\theta | s) = \frac{n}{n+m}$ можно и без этого.

ПРИМЕР ПРИМЕНЕНИЯ ТЕОРЕМЫ БАЙЕСА

- Итого получается:

$$p(\theta|s) = \frac{\theta^n(1-\theta)^m}{p(s)}.$$

- Но это ещё не всё. Чтобы предсказать следующий исход, надо найти $p(\text{heads}|s)$:

$$\begin{aligned} p(\text{heads}|s) &= \int_0^1 p(\text{heads}|\theta)p(\theta|s)d\theta = \\ &= \int_0^1 \frac{\theta^{n+1}(1-\theta)^m}{p(s)} dp_h = \\ &= \frac{(n+1)!m!}{(n+m+2)!} \cdot \frac{(n+m+1)!}{n!m!} = \frac{n+1}{n+m+2}. \end{aligned}$$

- Получили правило Лапласа.

ПРИМЕР ПРИМЕНЕНИЯ ТЕОРЕМЫ БАЙЕСА

- Итого получается:

$$p(\theta|s) = \frac{\theta^n(1-\theta)^m}{p(s)}.$$

- Это была иллюстрация двух основных задач байесовского вывода:
 - (1) найти апостериорное распределение (posterior distribution) на гипотезах/параметрах:

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

и/или найти максимальную апостериорную гипотезу (maximum a posteriori) $\theta_{\text{MAP}} = \arg \max_{\theta} p(\theta | D)$;

- (2) найти предсказательное распределение (predictive distribution) исходов дальнейших экспериментов:

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

СОПРЯЖЁННЫЕ
РАСПРЕДЕЛЕНИЯ

АПРИОРНЫЕ



- Напоминаю, что основная наша задача – как обучить параметры распределения и/или предсказать следующие его точки по имеющимся данным.
- В байесовском выводе участвуют:
 - $p(x | \theta)$ – правдоподобие данных;
 - $p(\theta)$ – априорное распределение;
 - $p(x) = \int_{\Theta} p(x | \theta)p(\theta)d\theta$ – маргинальное правдоподобие;
 - $p(\theta | x) = \frac{p(x|\theta)p(\theta)}{p(x)}$ – апостериорное распределение;
 - $p(x' | x) = \int_{\Theta} p(x' | \theta)p(\theta | x)d\theta$ – предсказание нового x' .
- Задача обычно в том, чтобы найти $p(\theta | x)$ и/или $p(x' | x)$.

- Когда мы проводим байесовский вывод, у нас, кроме правдоподобия, должно быть ещё *априорное распределение* (prior distribution) по всем возможным значениям параметров.
- Мы раньше к ним специально не присматривались, но они очень важны.
- Задача байесовского вывода – как подсчитать $p(\theta | x)$ и/или $p(x' | x)$.
- Но чтобы это сделать, сначала надо выбрать $p(\theta)$. Как выбирать априорные распределения?

- Разумная цель: давайте будем выбирать распределения так, чтобы они оставались такими же и *a posteriori*.
- До начала вывода есть априорное распределение $p(\theta)$.
- После него есть какое-то новое апостериорное распределение $p(\theta \mid x)$.
- Я хочу, чтобы $p(\theta \mid x)$ тоже имело тот же вид, что и $p(\theta)$, просто с другими параметрами.

- Не слишком формальное определение: семейство распределений $p(\theta \mid \alpha)$ называется семейством *сопряжённых априорных распределений* для семейства правдоподобий $p(x \mid \theta)$, если после умножения на правдоподобие апостериорное распределение $p(\theta \mid x, \alpha)$ остаётся в том же семействе: $p(\theta \mid x, \alpha) = p(\theta \mid \alpha')$.
- α называются *гиперпараметрами* (hyperparameters), это «параметры распределения параметров».
- Тривиальный пример: семейство всех распределений будет сопряжённым чему угодно, но это не очень интересно.

- Разумеется, вид хорошего априорного распределения зависит от вида распределения собственно данных, $p(x \mid \theta)$.
- Сопряжённые априорные распределения подсчитаны для многих распределений, мы приведём несколько примеров.

- Каким будет сопряжённое априорное распределение для бросания нечестной монетки (испытаний Бернулли)?
- Ответ: это будет бета-распределение; плотность распределения нечестности монетки θ

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

- Плотность распределения нечестности монетки θ

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

- Тогда, если мы посэмплируем монетку, получив s орлов и f решек, получится

$$p(s, f \mid \theta) = \binom{s+f}{s} \theta^s (1-\theta)^f, \text{ и}$$

$$\begin{aligned} p(\theta \mid s, f) &= \frac{\binom{s+f}{s} \theta^{s+\alpha-1} (1-\theta)^{f+\beta-1} / B(\alpha, \beta)}{\int_0^1 \binom{s+f}{s} x^{s+\alpha-1} (1-x)^{f+\beta-1} / B(\alpha, \beta) dx} = \\ &= \frac{\theta^{s+\alpha-1} (1-\theta)^{f+\beta-1}}{B(s+\alpha, f+\beta)}. \end{aligned}$$

- Итого получается, что сопряжённое априорное распределение для параметра нечестной монетки θ – это

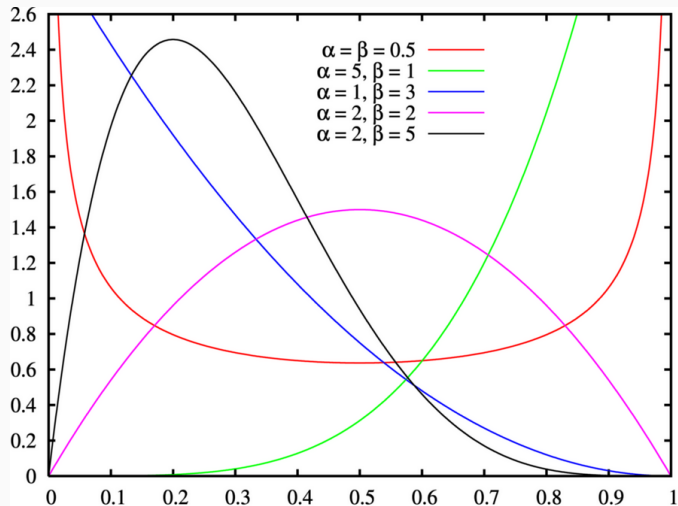
$$p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

- После получения новых данных с s орлами и f решками гиперпараметры меняются на

$$p(\theta \mid s + \alpha, f + \beta) \propto \theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}.$$

- На этом этапе можно забыть про сложные формулы и выводы, получилось очень простое правило обучения (под обучением теперь понимается изменение гиперпараметров).

БЕТА--РАСПРЕДЕЛЕНИЕ



МУЛЬТИНОМИАЛЬНОЕ РАСПРЕДЕЛЕНИЕ

- Простое обобщение: рассмотрим мультиномиальное распределение с n испытаниями, k категориями и по x_i экспериментов дали категорию i .
- Параметры θ_i показывают вероятность попасть в категорию i :

$$p(x \mid \theta) = \binom{n}{x_1, \dots, x_n} \theta_1^{x_1} \theta_2^{x_2} \dots \theta_k^{x_k}.$$

- Сопряжённым априорным распределением будет распределение Дирихле:

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \dots \theta_k^{\alpha_k-1}.$$

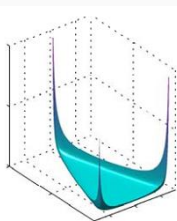
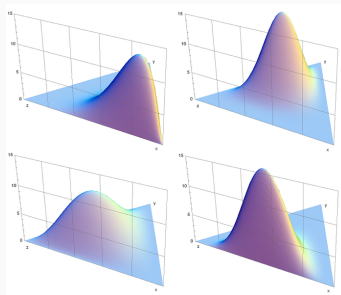
- Сопряжённым априорным распределением будет распределение Дирихле:

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1-1} \theta_2^{\alpha_2-1} \dots \theta_k^{\alpha_k-1}.$$

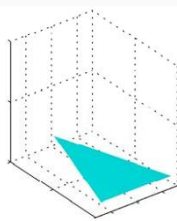
Упражнение. Докажите, что при получении данных $x_1, \dots, x_k$ гиперпараметры изменятся на

$$p(\theta \mid x, \alpha) = p(\theta \mid x + \alpha) \propto \theta_1^{x_1+\alpha_1-1} \theta_2^{x_2+\alpha_2-1} \dots \theta_k^{x_k+\alpha_k-1}.$$

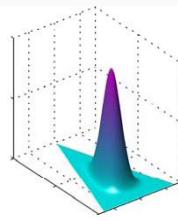
РАСПРЕДЕЛЕНИЕ ДИРИХЛЕ



$$\{\alpha_k\} = 0.1$$



$$\{\alpha_k\} = 1$$



$$\{\alpha_k\} = 10$$

ЕЩЁ НЕМНОГО О ВЕРОЯТНОСТЯХ

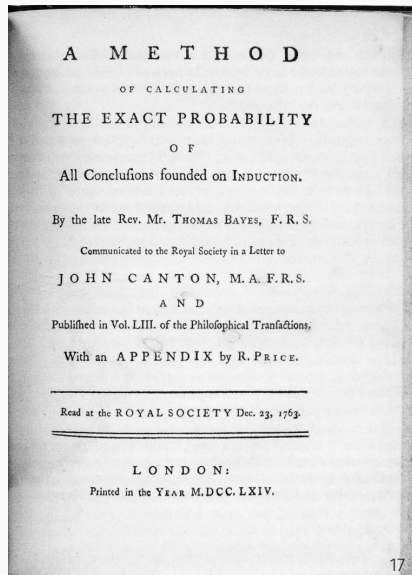
- Приведём классический пример из классической области применения статистики — медицины.
- Пусть некий тест на какую-нибудь болезнь имеет вероятность успеха 95% (т.е. 5% — вероятность как позитивной, так и негативной ошибки).
- Всего болезнь имеется у 1% респондентов (отложим на время то, что они разного возраста и профессий).
- Пусть некий человек получил позитивный результат теста (тест говорит, что он болен). С какой вероятностью он действительно болен?

- Приведём классический пример из классической области применения статистики — медицины.
- Пусть некий тест на какую-нибудь болезнь имеет вероятность успеха 95% (т.е. 5% — вероятность как позитивной, так и негативной ошибки).
- Всего болезнь имеется у 1% респондентов (отложим на время то, что они разного возраста и профессий).
- Пусть некий человек получил позитивный результат теста (тест говорит, что он болен). С какой вероятностью он действительно болен?
- Ответ: 16%.

- Обозначим через t результат теста, через d — наличие болезни.
- $p(t = 1) = p(t = 1|d = 1)p(d = 1) + p(t = 1|d = 0)p(d = 0)$.
- Используем теорему Байеса:

$$\begin{aligned} p(d = 1|t = 1) &= \\ &= \frac{p(t = 1|d = 1)p(d = 1)}{p(t = 1|d = 1)p(d = 1) + p(t = 1|d = 0)p(d = 0)} = \\ &= \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} = 0.16. \end{aligned}$$

- Вот такие задачи составляют суть вероятностного вывода (probabilistic inference).
- Поскольку они обычно основаны на теореме Байеса, вывод часто называют байесовским (Bayesian inference).
- Но не только поэтому.

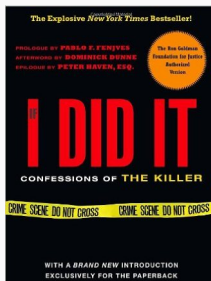


1. У моего знакомого два ребёнка. Будем предполагать, что пол ребёнка выбирается независимо и равновероятно, с вероятностью $\frac{1}{2}$. Две постановки вопроса:
 - (1) я спросил, есть ли у него мальчики, и он ответил «да»; какова вероятность того, что один из детей – девочка?
 - (2) я встретил одного из его детей, и это мальчик; какова вероятность того, что второй ребёнок – девочка?

2. Произошло убийство. На месте убийства найдена кровь, которая явно принадлежит убийце. Кровь принадлежит редкой группе, которая присутствует у 1% населения, в том числе у подсудимого.
- (1) Прокурор говорит: «Шанс, что у подсудимого была бы именно такая группа крови, если бы он был невиновен – всего 1%; значит, с вероятностью 99% он виновен». В чём не прав прокурор?
- (2) Адвокат говорит: «В городе живёт миллион человек, то есть у 10000 из них такая группа крови. Значит, всё, что говорит нам эта кровь – это что подсудимый совершил убийство с вероятностью 0.01%; никакое это не доказательство». В чём не прав адвокат?

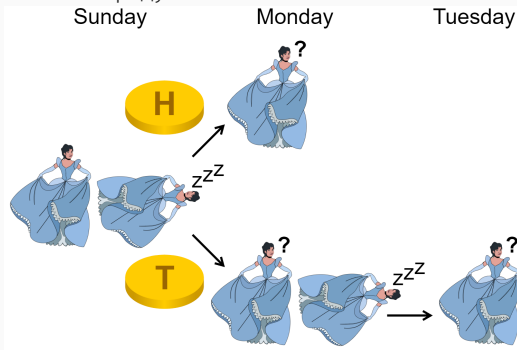
3. Реальные случаи.

- (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 Sally Clark 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?



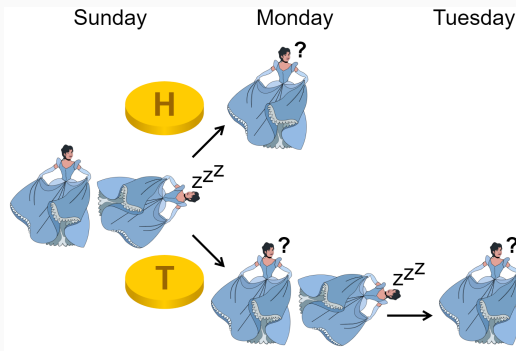
ЗАДАЧА О СПЯЩЕЙ КРАСАВИЦЕ

- А вот парадокс, который до сих пор вызывает споры (Zuboff, 1990; Elga, 2000). В воскресенье Спящую красавицу усыпляют и бросают монетку:
 - если выпадает орёл, её будят в понедельник и опять усыпляют до среды
 - если решка, то будят в понедельник, делают укол, от которого она забывает о пробуждении, и будят ещё раз во вторник, а потом так же в среду



ЗАДАЧА О СПЯЩЕЙ КРАСАВИЦЕ

- Спящая красавица знает всю процедуру эксперимента, но не может различить разные пробуждения и отличить понедельник от вторника, когда просыпается
- Представьте себя на месте Спящей красавицы; вас разбудили; какова вероятность того, что монета упала орлом?



ЛИНЕЙНАЯ РЕГРЕССИЯ

- 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

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

- 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?

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

- 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.

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

- 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):

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

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

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

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

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

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

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

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

- 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.

ПРИМЕР: ПОЛИНОМИАЛЬНАЯ
АППРОКСИМАЦИЯ

- Мы говорили о регрессии с базисными функциями:

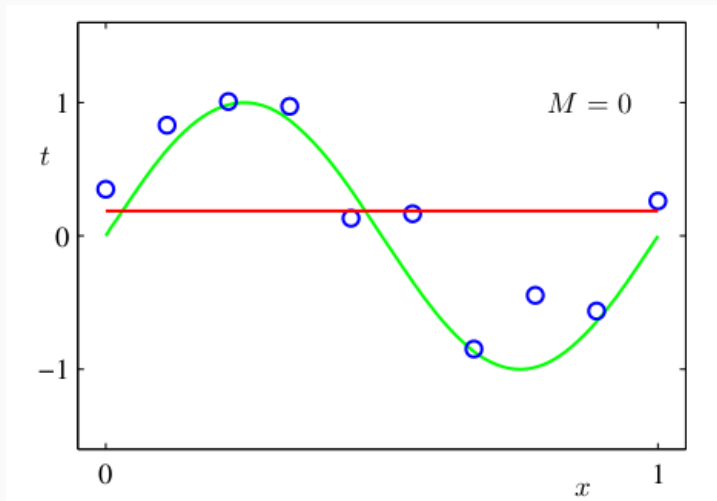
$$f(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^M w_j \phi_j(\mathbf{x}) = \mathbf{w}^\top \phi(\mathbf{x}).$$

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

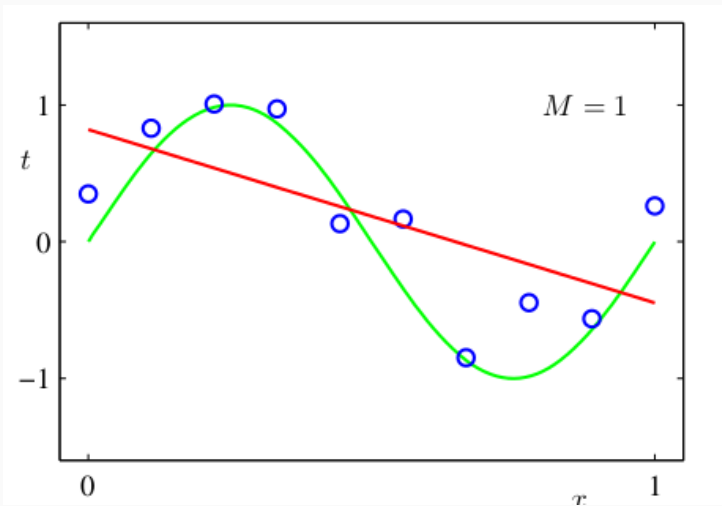
$$f(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \dots + w_M x^M.$$

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

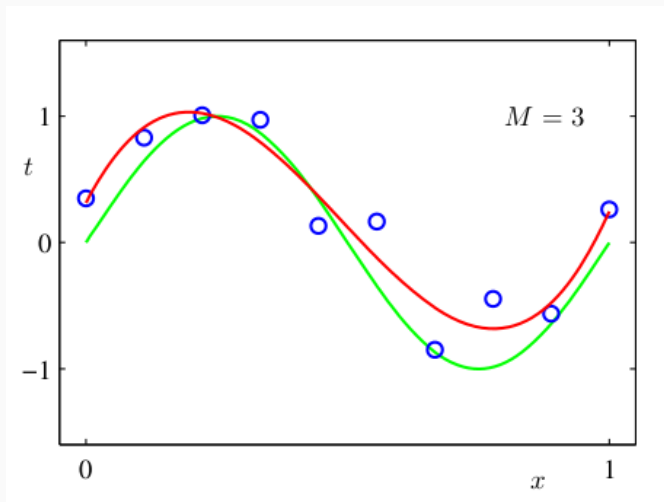
Полиномиальная аппроксимация



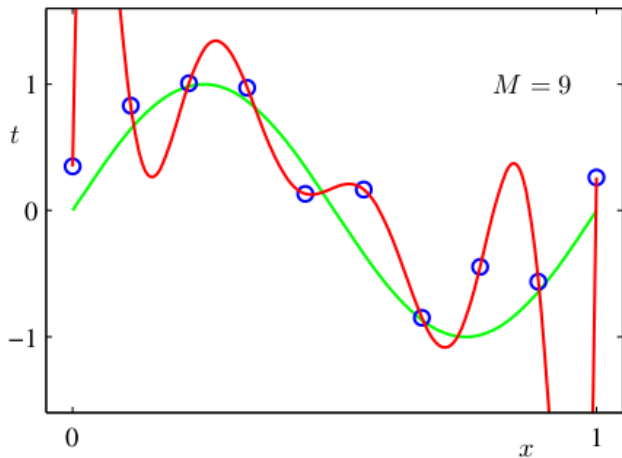
Полиномиальная аппроксимация



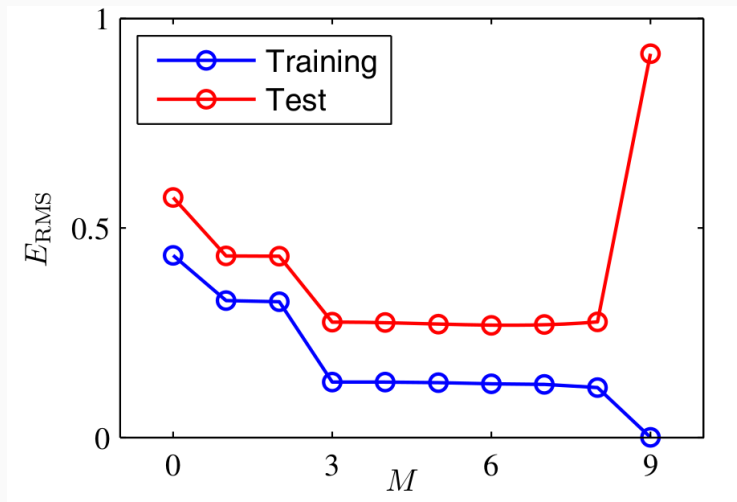
Полиномиальная аппроксимация



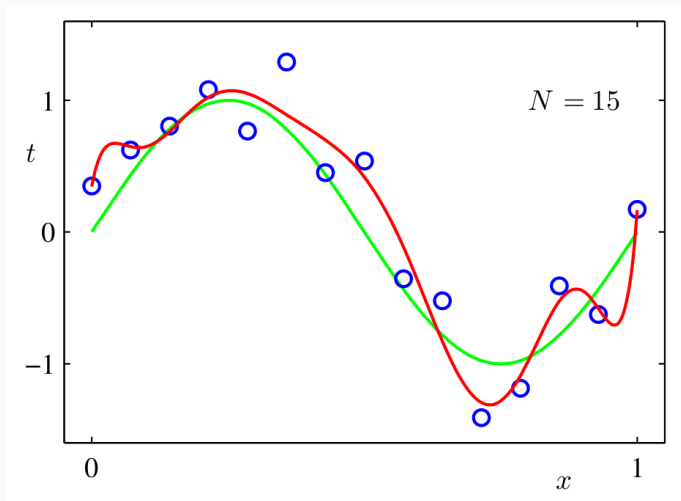
Полиномиальная аппроксимация



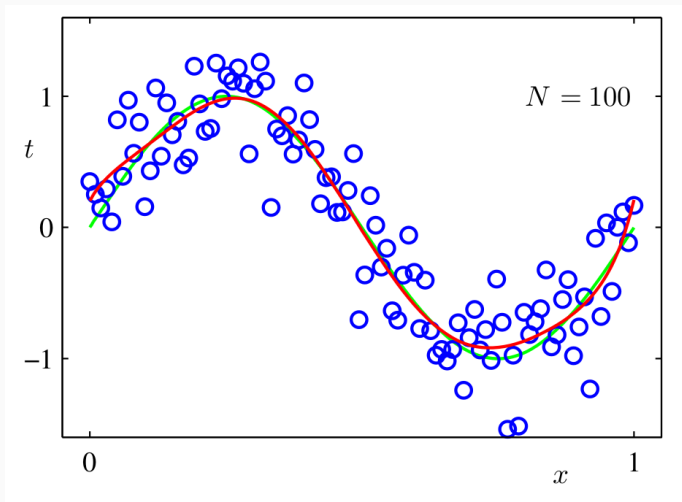
ЗНАЧЕНИЯ RMS



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

СПАСИБО!

Спасибо за внимание!

