

# БАЙЕСОВСКИЙ ВЫВОД ДЛЯ ГАУССИАНА

---

Сергей Николенко

НИУ ВШЭ — Санкт-Петербург

16 мая 2018 г.

---

## *Random facts:*

- 16 мая 1768 г. Екатерина II издала указ о постройке памятника Петру I (Медного всадника)
- 16 мая 1896 г. в Петербурге в саду Аквариум состоялся первый в России киносеанс
- 16 мая 1924 г. вышел первый номер журнала «Мурзилка»
- 16 мая 1929 г. состоялась первая церемония вручения (будущих) «Оскаров»
- 16 мая 1931 г. папа римский Пий XI заявил о несовместимости социализма с католицизмом

# БАЙЕСОВСКИЙ ВЫВОД ДЛЯ ГАУССИАНА

---

- На самом деле всё это — байесовский вывод для нормального распределения:

$$p(x_1, \dots, x_n \mid \mu, \sigma^2) \propto \frac{1}{\sigma^n} \exp \left( -\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2 \right).$$

- Хотим: найти сопряжённое априорное распределение, подсчитать правдоподобие, решить задачу предсказания.
- Для начала зафиксируем  $\sigma^2$  и будем в качестве параметра рассматривать только  $\mu$ .

- Сопряжённое априорное распределение для  $\mu$  при фиксированном  $\sigma^2$  тоже нормальное и выглядит как

$$p(\mu \mid \mu_0, \sigma_0^2) \propto \frac{1}{\sigma_0^n} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right).$$

- Обычно выбирают  $\mu_0 = 0$ ,  $\sigma_0^2 \rightarrow \infty$  (порой буквально).
- Давайте рассмотрим сначала случай ровно одного наблюдения  $x$  и найдём  $p(\mu \mid x)$ .

- При нашем априорном распределении у  $\mu$  и  $x$  совместное нормальное распределение:

$$x = \mu + \sigma\epsilon, \quad \mu = \mu_0 + \sigma_0\delta, \quad \epsilon, \delta \sim \mathcal{N}(0, 1).$$

**Упражнение.** Пусть  $(z_1, z_2)$  – случайные величины с совместным нормальным распределением. Докажите, что случайная величина  $z_1 | z_2$  распределена нормально с параметрами

$$E(z_1 | z_2) = E(z_1) + \frac{\text{Cov}(z_1, z_2)}{\text{Var}(z_2)} (z_2 - E(z_2)),$$

$$\text{Var}(z_1 | z_2) = \text{Var}(z_1) - \frac{\text{Cov}^2(z_1, z_2)}{\text{Var}(z_2)}$$

$$(\text{Var}(x) = E[(x - Ex)^2], \text{Cov}(x, y) = E[(x - Ex)(y - Ey)]).$$

- В нашем случае:

$$x = \mu + \sigma\epsilon, \quad \mu = \mu_0 + \sigma_0\delta, \quad \epsilon, \delta \sim \mathcal{N}(0, 1),$$

$$E(x) = \mu_0,$$

$$\text{Var}(x) = E(\text{Var}(x | \mu)) + \text{Var}(E(x | \mu)) = \sigma^2 + \sigma_0^2,$$

$$\text{Cov}(x, \mu) = E[(x - \mu_0)(\mu - \mu_0)] = \sigma_0^2.$$

- Применив упражнение, получаем:

$$E(\mu | x) = \mu_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}(x - \mu_0) = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}x + \frac{\sigma^2}{\sigma_0^2 + \sigma^2}\mu_0,$$

$$\text{Var}(\mu | x) = \frac{\sigma^2\sigma_0^2}{\sigma_0^2 + \sigma^2} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}.$$

- Итого:

$$p(\mu | x) \sim \mathcal{N} \left( \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2} x + \frac{\sigma^2}{\sigma_0^2 + \sigma^2} \mu_0, \left( \frac{1}{\sigma_0^2} + \frac{1}{\sigma^2} \right)^{-1} \right).$$

- Опять же, сложные вычисления можно забыть и пользоваться этими формулами.
- Замечание: часто используют  $\tau = \frac{1}{\sigma^2}$  как параметр нормального распределения (precision). Тогда

$$\tau_{\mu|x} = \tau_{\mu} + \tau.$$

- А что, если данных больше,  $x_1, \dots, x_n$ ?
- Тогда можно повторить всё то же самое, а можно заметить, что набор данных описывается своим средним.

**Упражнение.** Докажите, что если  $p(x_i | \mu) \sim \mathcal{N}(\mu, \sigma^2)$  и  $x_i$  независимы, то  $p(\bar{x} | \mu) \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ .



- Для апостериорной вероятности будет

$$p(\mu | x_1, \dots, x_n) \propto p(x_1, \dots, x_n | \mu)p(\mu) \propto p(\bar{x} | \mu)p(\mu) \propto p(\mu | \bar{x}).$$

- Подставляя в наш предыдущий результат, получим:

$$p(\mu | x_1, \dots, x_n) \sim \mathcal{N} \left( \frac{\sigma_0^2}{\sigma_0^2 + \frac{\sigma^2}{n}}x + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0, \left( \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1} \right).$$

- Если зафиксировать  $\mu$  и менять  $\sigma^2$ , то сопряжённым априорным распределением будет обратное гамма-распределение:

$$p(\sigma^2 \mid \alpha, \beta) \propto IG(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} \exp\left(\frac{-\beta}{z}\right).$$

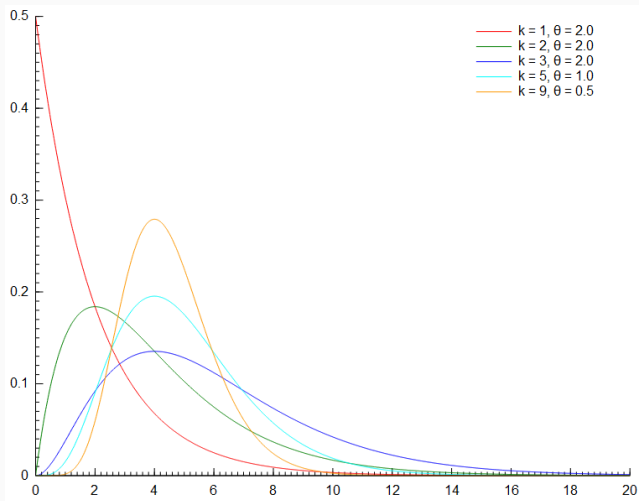
- Тогда в апостериорном распределении будет

$$p(\sigma^2 \mid x_1, \dots, x_n, \alpha, \beta) \propto IG\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum (x_i - \mu)\right).$$

- А в терминах  $\tau = \frac{1}{\sigma^2}$  будет обычное гамма-распределение:

$$p(\tau \mid x_1, \dots, x_n, \alpha, \beta) \propto \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum (x_i - \mu)\right).$$

# ГАММА--РАСПРЕДЕЛЕНИЕ



## КОГДА И $\mu$ , И $\sigma^2$ МЕНЯЮТСЯ

- Что делать, когда и  $\mu$ , и  $\sigma^2$  меняются?
- Можно было бы предположить, что  $\mu$  и  $\sigma^2$  независимы; тогда просто априорное распределение будет

$$p(\mu, \sigma \mid \mu_0, \sigma_0, \alpha, \beta) \propto \mathcal{N}(\mu_0, \sigma_0^2) \cdot IG(\alpha, \beta).$$

- К сожалению, это распределение не будет сопряжённым к нормальному. Почему?

## КОГДА И $\mu$ , И $\sigma^2$ МЕНЯЮТСЯ

- Что делать, когда и  $\mu$ , и  $\sigma^2$  меняются?
- Можно было бы предположить, что  $\mu$  и  $\sigma^2$  независимы; тогда просто априорное распределение будет

$$p(\mu, \sigma \mid \mu_0, \sigma_0, \alpha, \beta) \propto \mathcal{N}(\mu_0, \sigma_0^2) \cdot IG(\alpha, \beta).$$

- К сожалению, это распределение не будет сопряжённым к нормальному. Почему?
- Потому что  $\mu$  и  $\sigma^2$  зависимы. :) Новая точка  $x$  вводит зависимость между ними.

- В настоящем сопряжённом априорном распределении будут:

$$\begin{aligned}x \mid \mu, \tau &\sim \mathcal{N}(\mu, \tau), \\ \mu \mid \tau &\sim \mathcal{N}(\mu_0, n_0\tau), \\ \tau &\sim G(\alpha, \beta).\end{aligned}$$

- Давайте выясним, как изменятся параметры, и заодно докажем.

- Самое простое – это, по уже известным результатам,

$$\mu \mid x, \tau \sim \mathcal{N} \left( \frac{n\tau}{n\tau + n_0\tau} \bar{x} + \frac{n_0\tau}{n\tau + n_0\tau} \mu_0, n\tau + n_0\tau \right).$$

- Затем давайте разберёмся с  $\tau \mid x$ :

$$p(\tau, \mu \mid x) \propto p(\tau) \cdot p(\mu \mid \tau) \cdot p(x \mid \tau, \mu),$$

и мы хотим это распределение маргинализовать по  $\mu$ ...

- Подсчитаем:

$$\begin{aligned} p(\tau, \mu | x) &\propto p(\tau) \cdot p(\mu | \tau) \cdot p(x | \tau, \mu) \\ &\propto \tau^{\alpha-1} e^{-\tau\beta} \cdot \tau^{\frac{1}{2}} e^{-\frac{n_0\tau}{2}(\mu-\mu_0)^2} \cdot \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\sum(x_i-\mu)^2} \\ &\propto \tau^{\alpha+\frac{n}{2}-\frac{1}{2}} e^{-\tau(\beta+\frac{1}{2}\sum(x_i-\bar{x})^2)} e^{-\frac{\tau}{2}(n_0(\mu-\mu_0)^2+n(\bar{x}-\mu)^2)} \end{aligned}$$

(простой трюк:  $x_i - \mu = x_i - \bar{x} + \bar{x} - \mu$ ).



- Теперь надо проинтегрировать

$$\int_{\mu} e^{-\frac{\tau}{2}(n_0(\mu-\mu_0)^2+n(\bar{x}-\mu)^2)} d\mu.$$

**Упражнение.** Проинтегрируйте. :) Должна получиться нормировочная константа

$$\tau^{-\frac{1}{2}} e^{\frac{-nn_0\tau}{2(n+n_0)}(\bar{x}-\mu_0)^2}.$$

- Таким образом, получается апостериорное распределение

$$p(\tau | x) \propto \tau^{\alpha + \frac{n}{2} - 1} e^{-\tau \left( \beta + \frac{1}{2} \sum (x_i - \bar{x})^2 + \frac{nn_0}{2(n+n_0)} (\bar{x} - \mu_0)^2 \right)}.$$

- Итого результаты такие:

$$\begin{aligned} \mu | \tau, x &\sim \mathcal{N} \left( \frac{n\tau}{n\tau + n_0\tau} \bar{x} + \frac{n_0\tau}{n\tau + n_0\tau} \mu_0, n\tau + n_0\tau \right), \\ \tau | x &\sim G \left( \alpha + \frac{n}{2}, \beta + \frac{1}{2} \sum (x_i - \bar{x})^2 + \frac{nn_0}{2(n+n_0)} (\bar{x} - \mu_0)^2 \right). \end{aligned}$$

- Теперь предсказание нового  $x_{\text{new}}$ :

$$\begin{aligned} p(x_{\text{new}} | x) &= \int \int \underbrace{\text{Gamma}}_{\tau|x} \cdot \underbrace{\text{Gaussian}}_{\mu|\tau,x} \cdot \underbrace{\text{Gaussian}}_{x_{\text{new}}|\tau,\mu} d\tau d\mu = \\ &= \int \underbrace{\text{Gamma}}_{\tau|x} \int \underbrace{\text{Gaussian}}_{\mu|\tau,x} \cdot \underbrace{\text{Gaussian}}_{x_{\text{new}}|\tau,\mu} d\tau d\mu = \\ &= \int \underbrace{\text{Gamma}}_{\tau|x} \cdot \underbrace{\text{Gaussian}}_{x_{\text{new}}|\tau,x} d\tau = \dots \end{aligned}$$

- В результате получится распределение Стьюдента.  
**Упражнение.** Проведите эти вычисления.

- Вообще говоря, всё, о чём мы говорили – частные случаи экспоненциального семейства распределений:

$$p(\mathbf{x} | \eta) = h(\mathbf{x})g(\eta)e^{\eta^T \mathbf{u}(\mathbf{x})}.$$

- $\eta$  называются *естественными параметрами* (natural parameters).

- Например, распределение Бернулли:

$$\begin{aligned} p(x | \mu) &= \mu^x (1 - \mu)^{1-x} = e^{x \ln \mu + (1-x) \ln(1-\mu)} = \\ &= (1 - \mu) e^{\ln\left(\frac{\mu}{1-\mu}\right)x}, \end{aligned}$$

и естественный параметр получился  $\eta = \ln\left(\frac{\mu}{1-\mu}\right)$ :

$$p(x | \eta) = \sigma(-\eta) e^{-\eta x},$$

где  $\sigma(y) = \frac{1}{1+e^{-y}}$  – сигмоид-функция.

- Для мультиномиального распределения с параметрами  $\mu_1, \dots, \mu_{M-1}$  получаются

$$\eta_k = \ln \left( \frac{\mu_k}{1 - \sum_j \mu_j} \right) \text{ и}$$

$$p(\mathbf{x} | \eta) = \left( 1 + \sum_{k=1}^{M-1} e^{\eta_k} \right)^{-1} e^{\eta^\top \mathbf{x}}.$$

Упражнение. Проверьте!

- Так вот, для распределений из экспоненциального семейства

$$p(\mathbf{x} | \eta) = h(\mathbf{x})g(\eta)e^{\eta^T \mathbf{u}(\mathbf{x})}$$

можно сразу оптом найти сопряжённые априорные распределения:

$$p(\eta | \chi, \nu) = f(\chi, \nu)g(\eta)^\nu e^{\nu \eta^T \chi},$$

где  $\chi$  – гиперпараметры, а  $g$  то же самое, что в исходном распределении.

**Упражнение.** Проверьте это и получите вышеописанные примеры как частные случаи.

## ПРОКЛЯТИЕ РАЗМЕРНОСТИ

---



- Прежде чем узнать, что делать, общее замечание: модели бывают параметрические и непараметрические.
- Мы в основном будем заниматься моделями с фиксированным числом параметров, которые делают сильные предположения.
- Но есть класс непараметрических моделей, которые не делают предположений почти никаких (это не совсем правда), а основаны непосредственно на данных; они в некоторых ситуациях очень хороши, но плохо обобщаются на высокие размерности и большие датасеты.

- Пример непараметрической модели: метод ближайших соседей.
- Давайте на примере задачи классификации.
- Не будем строить вообще никакой модели, а будем классифицировать новые примеры как

$$\hat{y}(\mathbf{x}) = \frac{1}{k} \sum_{\mathbf{x}_i \in N_k(\mathbf{x})} y_i,$$

где  $N_k(\mathbf{x})$  – множество  $k$  ближайших соседей точки  $\mathbf{x}$  среди имеющихся данных  $(\mathbf{x}_i, y_i)_{i=1}^N$ .

- Единственный «параметр» – это  $k$ , но от него многое зависит.
- Для разумно большого  $k$  у нас в нашем примере стало меньше ошибок.
- Но это не предел – для  $k = 1$  на тестовых данных вообще никаких ошибок нету!
- Что это значит? В чём недостаток метода ближайших соседей при  $k = 1$ ?
- Как выбрать  $k$ ? Можно ли просто подсчитать ошибку классификации и минимизировать её?

- В прошлый раз  $k$ -NN давали гораздо более разумные результаты, чем линейная модель, особенно если хорошо выбрать  $k$ .
- Может быть, нам в этой жизни больше ничего и не нужно?
- Давайте посмотрим, как  $k$ -NN будет вести себя в более высокой размерности (что очень реалистично).

- Давайте поищем ближайших соседей у точки в единичном гиперкубе. Предположим, что наше исходное распределение равномерное.
- Чтобы покрыть долю  $\alpha$  тестовых примеров, нужно (ожидаемо) покрыть долю  $\alpha$  объёма, и ожидаемая длина ребра гиперкуба-окрестности в размерности  $p$  будет  $e_p(\alpha) = \alpha^{1/p}$ .
- Например, в размерности 10  $e_{10}(0.1) = 0.8$ ,  $e_{10}(0.01) = 0.63$ , т.е. чтобы покрыть 1% объёма, нужно взять окрестность длиной больше половины носителя по каждой координате!
- Это скажется и на  $k$ -NN: трудно отвергнуть по малому числу координат, быстрые алгоритмы хуже работают.

- Второе проявление the curse of dimensionality: пусть  $N$  точек равномерно распределены в единичном шаре размерности  $p$ . Тогда среднее расстояние от нуля до точки равно

$$d(p, N) = \left(1 - \frac{1}{2^{1/N}}\right)^{1/p},$$

т.е., например, в размерности 10 для  $N = 500$   $d \approx 0.52$ , т.е. больше половины.

- Большинство точек в результате ближе к границе носителя, чем к другим точкам, а это для ближайших соседей проблема – придётся не интерполировать внутри существующих точек, а экстраполировать наружу.

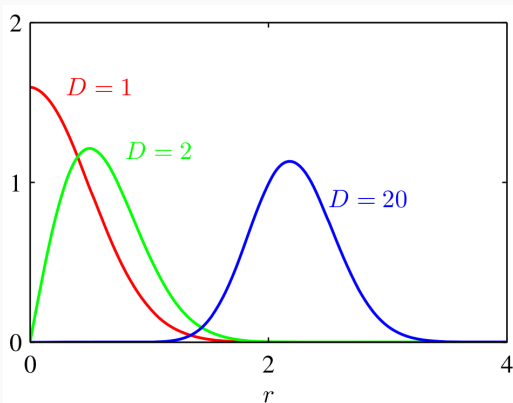
- Третье проявление: проблемы в оптимизации, которые и имел в виду Беллман.
- Если нужно примерно оптимизировать функцию от  $d$  переменных, на решётке с шагом  $\epsilon$  понадобится примерно  $(\frac{1}{\epsilon})^d$  вычислений функции.
- В численном интегрировании – чтобы интегрировать функцию с точностью  $\epsilon$ , нужно тоже примерно  $(\frac{1}{\epsilon})^d$  вычислений.

- Плотные множества становятся очень разреженными. Например, чтобы получить плотность, создаваемую в размерности 1 при помощи  $N = 100$  точек, в размерности 10 нужно будет  $100^{10}$  точек.
- Поведение функций тоже усложняется с ростом размерности – чтобы строить регрессии в высокой размерности с той же точностью, может потребоваться экспоненциально больше точек, чем в низкой размерности.
- А у линейной модели ничего такого не наблюдается, она не подвержена проклятию размерности.



# ПРОКЛЯТИЕ РАЗМЕРНОСТИ

- Ещё пример: нормально распределённая величина будет сосредоточена в тонкой оболочке.



**Упражнение.** Переведите плотность нормального распределения в полярные координаты и проверьте это утверждение.

СТАТИСТИЧЕСКАЯ  
ТЕОРИЯ ПРИНЯТИЯ РЕШЕНИЙ

---

- Сейчас мы попытаемся понять, что же на самом деле происходит в этих методах.
- Начнём с обычной регрессии – непрерывный вещественный вход  $\mathbf{x} \in \mathbb{R}^p$ , непрерывный вещественный выход  $y \in \mathbb{R}$ ; у них есть некоторое совместное распределение  $p(\mathbf{x}, y)$ .
- Мы хотим найти функцию  $f(\mathbf{x})$ , которая лучше всего предсказывает  $y$ .

- Введём *функцию потерь* (loss function)  $L(y, f(\mathbf{x}))$ , которая наказывает за ошибки; естественно взять квадратичную функцию потерь

$$L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2.$$

- Тогда каждому  $f$  можно сопоставить *ожидаемую ошибку предсказания* (expected prediction error):

$$\text{EPE}(f) = \mathbb{E}(y - f(\mathbf{x}))^2 = \int \int (y - f(\mathbf{x}))^2 p(\mathbf{x}, y) d\mathbf{x} dy.$$

- И теперь самая хорошая функция предсказания  $\hat{f}$  – это та, которая минимизирует  $\text{EPE}(f)$ .

- Это можно переписать как

$$\text{EPE}(f) = E_{\mathbf{x}} E_{y|\mathbf{x}} [(y - f(\mathbf{x}))^2 | \mathbf{x}],$$

и, значит, можно теперь минимизировать EPE поточечно:

$$\hat{f}(\mathbf{x}) = \arg \min_c E_{y|\mathbf{x}'} [(y - c)^2 | \mathbf{x}' = \mathbf{x}],$$

а это можно решить и получить

$$\hat{f}(\mathbf{x}) = E_{y|\mathbf{x}'}(y | \mathbf{x}' = \mathbf{x}).$$

- Это решение называется *функцией регрессии* и является наилучшим предсказанием  $y$  в любой точке  $\mathbf{x}$ .

- Теперь мы можем понять, что такое  $k$ -NN.
- Давайте оценим это ожидание:

$$f(\mathbf{x}) = \mathbb{E}_{y|\mathbf{x}'}(y \mid \mathbf{x}' = \mathbf{x}).$$

- Оценка ожидания – это среднее всех  $y$  с данным  $\mathbf{x}$ . Конечно, у нас таких нету, поэтому мы приближаем это среднее как

$$\hat{f}(\mathbf{x}) = \text{Average}[y_i \mid \mathbf{x}_i \in N_k(\mathbf{x})].$$

- Это сразу два приближения: ожидание через среднее и среднее в точке через среднее в ближних точках.
- Иначе говоря,  $k$ -NN предполагает, что в окрестности  $\mathbf{x}$  функция  $y(\mathbf{x})$  не сильно меняется, а лучше всего – она кусочно-постоянна.

- А линейная регрессия – это модельный подход, мы предполагаем, что функция регрессии линейна от своих аргументов:

$$f(\mathbf{x}) \approx \mathbf{x}^T \mathbf{w}.$$

- Теперь мы не берём условие по  $\mathbf{x}$ , как в  $k$ -NN, а просто собираем много значений для разных  $\mathbf{x}$  и обучаем модель.

- То же самое можно и с задачей классификации сделать. Пусть у нас переменная  $g$  с  $K$  возможными значениями  $g_1, \dots, g_k$  предсказывается.
- Введём функцию потерь, равную 1 за каждый неверный ответ. Получим

$$\text{EPE} = \text{E} [L(g, \hat{g}(\mathbf{x}))].$$

- Перепишем как раньше:

$$\text{EPE} = \text{E}_{\mathbf{x}} \sum_{k=1}^K L(g_k, \hat{g}(\mathbf{x})) p(g_k | \mathbf{x}).$$

- Опять достаточно оптимизировать поточечно:

$$\hat{g}(\mathbf{x}) = \arg \min_g \sum_{k=1}^K L(g_k, \hat{g}(\mathbf{x})) p(g_k | \mathbf{x}).$$



- Опять достаточно оптимизировать поточечно:

$$\hat{g}(\mathbf{x}) = \arg \min_g \sum_{k=1}^K L(g_k, \hat{g}(\mathbf{x})) p(g_k | \mathbf{x}).$$

- Для 0-1 функции потери это упрощается до

$$\hat{g}(\mathbf{x}) = \arg \min_g [1 - p(g | \mathbf{x})], \text{ т.е.}$$

$$\hat{g}(\mathbf{x}) = g_k, \text{ если } p(g_k | \mathbf{x}) = \max_g p(g | \mathbf{x}).$$

- Это называется *оптимальным байесовским классификатором*; если модель известна, то его обычно можно построить.

- Рассмотрим совместное распределение  $p(y, \mathbf{x})$  и квадратичную функцию потерь  $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$ .
- Мы знаем, что тогда оптимальная оценка – это функция регрессии

$$\hat{f}(\mathbf{x}) = \mathbb{E}[y | \mathbf{x}] = \int yp(y | \mathbf{x})dx.$$

- Давайте подсчитаем ожидаемую ошибку и перепишем её в другой форме:

$$\begin{aligned} E[L] &= E[(y - f(\mathbf{x}))^2] = E[(y - E[y | \mathbf{x}] + E[y | \mathbf{x}] - f(\mathbf{x}))^2] = \\ &= \int (f(\mathbf{x}) - E[y | \mathbf{x}])^2 p(\mathbf{x}) d\mathbf{x} + \int (E[y | \mathbf{x}] - y)^2 p(\mathbf{x}, y) d\mathbf{x} dy, \end{aligned}$$

потому что

$$\int (f(\mathbf{x}) - E[y | \mathbf{x}]) (E[y | \mathbf{x}] - y) p(\mathbf{x}, y) d\mathbf{x} dy = 0.$$

- Эта форма записи – разложение на bias-variance и noise:

$$E[L] = \int (f(\mathbf{x}) - E[y | \mathbf{x}])^2 p(\mathbf{x}) d\mathbf{x} + \int (E[y | \mathbf{x}] - y)^2 p(\mathbf{x}, y) d\mathbf{x} dy,$$

- Отсюда, кстати, тоже сразу видно, что от  $f(\mathbf{x})$  зависит только первый член, и он минимизируется, когда

$$f(\mathbf{x}) = \hat{f}(\mathbf{x}) = E[y | \mathbf{x}].$$

- А noise,  $\int (E[y | \mathbf{x}] - y)^2 p(\mathbf{x}, y) d\mathbf{x} dy$ , – это просто свойство данных, дисперсия шума.

- Если бы у нас был всемогущий компьютер и неограниченный датасет, мы бы, конечно, на этом и закончили, посчитали бы  $\hat{f}(\mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$ , и всё.
- Однако жизнь – борьба, и у нас есть только ограниченный датасет из  $N$  точек. Предположим, что этот датасет берётся по распределению  $p(\mathbf{x}, y)$  – т.е. фактически рассмотрим много-много экспериментов такого вида:
  - взяли датасет  $D$  из  $N$  точек по распределению  $p(\mathbf{x}, y)$ ;
  - подсчитали нашу чудо-регрессию;
  - получили новую функцию предсказания  $f(\mathbf{x}; D)$ .
- Разные датасеты будут приводить к разным функциям предсказания...

- ...а потому давайте усредним теперь по датасетам.
- Наш первый член в ожидаемой ошибке выглядел как  $(f(\mathbf{x}) - \hat{f}(\mathbf{x}))^2$ , а теперь будет  $(f(\mathbf{x}; D) - \hat{f}(\mathbf{x}))^2$ , и его можно усреднить по  $D$ , применив такой же трюк:

$$\begin{aligned} & (f(\mathbf{x}; D) - \hat{f}(\mathbf{x}))^2 \\ &= (f(\mathbf{x}; D) - \mathbb{E}_D[f(\mathbf{x}; D)] + \mathbb{E}_D[f(\mathbf{x}; D)] - \hat{f}(\mathbf{x}))^2 \\ &= (f(\mathbf{x}; D) - \mathbb{E}_D[f(\mathbf{x}; D)])^2 + (\mathbb{E}_D[f(\mathbf{x}; D)] - \hat{f}(\mathbf{x}))^2 + 2(\dots)(\dots), \end{aligned}$$

и в ожидании получится...

- ...и в ожидании получится

$$\begin{aligned} \mathbb{E}_D \left[ (f(\mathbf{x}; D) - \hat{f}(\mathbf{x}))^2 \right] &= \\ &= \mathbb{E}_D \left[ (f(\mathbf{x}; D) - \mathbb{E}_D [f(\mathbf{x}; D)])^2 \right] + \left( \mathbb{E}_D [f(\mathbf{x}; D)] - \hat{f}(\mathbf{x}) \right)^2. \end{aligned}$$

- Разложили на дисперсию  $\mathbb{E}_D \left[ (f(\mathbf{x}; D) - \mathbb{E}_D [f(\mathbf{x}; D)])^2 \right]$  и квадрат систематической ошибки  $\left( \mathbb{E}_D [f(\mathbf{x}; D)] - \hat{f}(\mathbf{x}) \right)^2$ ; это и есть bias-variance decomposition.

Expected loss = (bias)<sup>2</sup> + variance + noise,

где

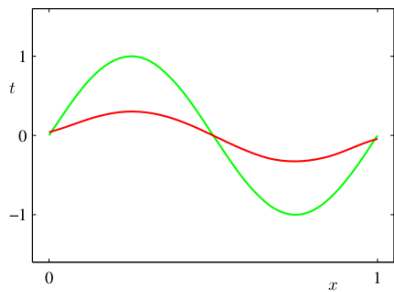
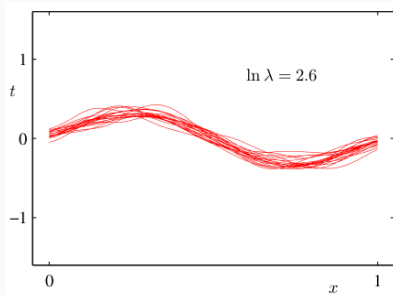
$$(\text{bias})^2 = \left( \mathbb{E}_D [f(\mathbf{x}; D)] - \hat{f}(\mathbf{x}) \right)^2,$$

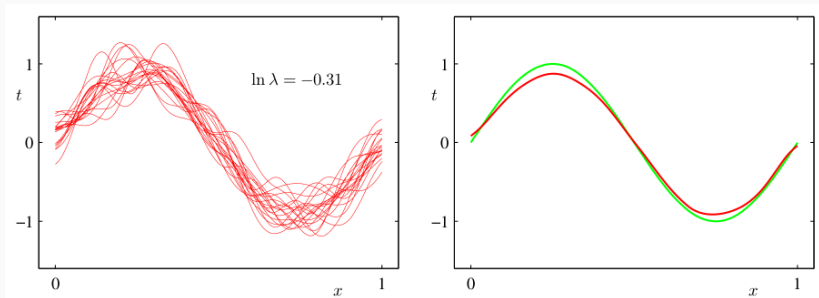
$$\text{variance} = \mathbb{E}_D \left[ (f(\mathbf{x}; D) - \mathbb{E}_D [f(\mathbf{x}; D)])^2 \right],$$

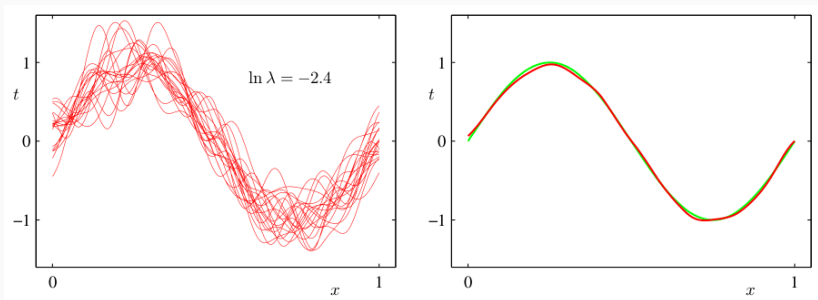
$$\text{noise} = \int (\mathbb{E}[y | \mathbf{x}] - y)^2 p(\mathbf{x}, y) d\mathbf{x} dy.$$

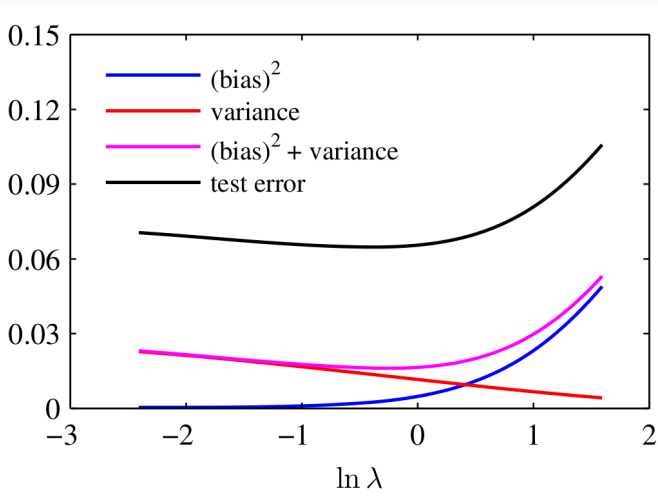


- Теперь давайте посмотрим на пример: опять та же синусоида, опять приближаем её линейной регрессией с полиномиальными признаками (максимальным их числом).
- И мы регуляризуем эту регрессию с параметром  $\alpha$ .
- Будем набрасывать много датасетов и смотреть, что меняется при этом.









Спасибо за внимание!