

БАЙЕСОВСКИЙ ВЫВОД ДЛЯ ГАУССИАНОВ

Сергей Николенко

НИУ ВШЭ — Санкт-Петербург

9 февраля 2019 г.

Random facts:

- 9 февраля --- Международный день стоматолога; дата выбрана как день почитания святой мученицы Аполлонии Александрийской, «дивной старушки-девственницы» (Евсевий Памфил), которой язычники во времена гонений императора Деция выбили все зубы
- 9 февраля 1849 г. по предложению Гарибальди была провозглашена Римская республика, просуществовавшая до начала июня того же года
- 9 февраля 1896 г. в Юсуповском саду Санкт-Петербурга прошёл первый чемпионат мира по фигурному катанию
- 9 февраля 1904 г. произошёл бой у Чемульпо (ныне Инчхон): броненосные крейсера «Асама» и «Чиода», бронепалубные крейсера «Нанива», «Такачихо», «Ниитака», «Акаси» и миноносцы «Хаябуса», «Тидори» и «Манадзуру» сумели обеспечить высадку десанта и нейтрализовать крейсер и канонерскую лодку противника

ПРЕДСКАЗАНИЯ В ЛИНЕЙНОЙ РЕГРЕССИИ

- Теперь давайте вернёмся к байесовской постановке:
 1. найти апостериорное распределение на гипотезах/параметрах:

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

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

2. найти апостериорное распределение исходов дальнейших экспериментов:

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

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

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

мы нашли

$$\begin{aligned} p(w | t, \alpha, \beta) &= N(w | \mu_N, \Sigma_N), \\ \mu_N &= \Sigma_N (\Sigma_0^{-1} \mu_0 + \beta \Phi^T t), \\ \Sigma_N &= (\Sigma_0^{-1} + \beta \Phi^T \Phi)^{-1}, \end{aligned}$$

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

- Теперь сделаем следующий шаг – найдём апостериорное распределение наших предсказаний

$$p(t | t, \alpha, \beta) = \int p(t | w, \beta)p(w | t, \alpha, \beta)dw.$$

- Это свёртка двух гауссианов, и получается...

- ...тоже гауссиан:

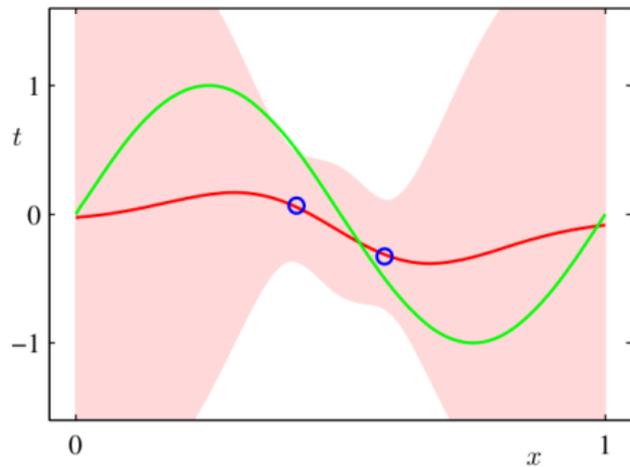
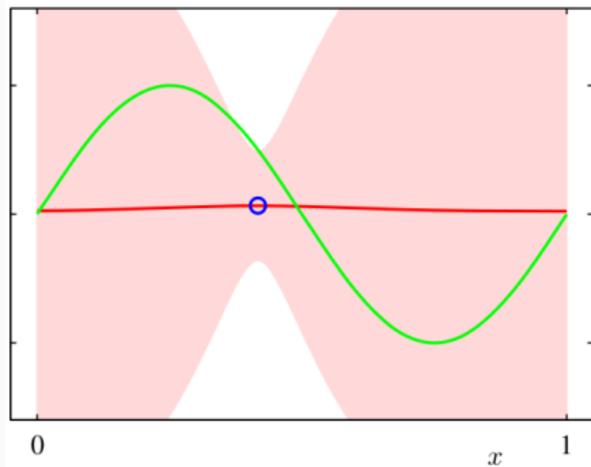
$$p(t | t, \alpha, \beta) = N(t | \mu_N^\top \phi(x), \sigma_N^2),$$

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

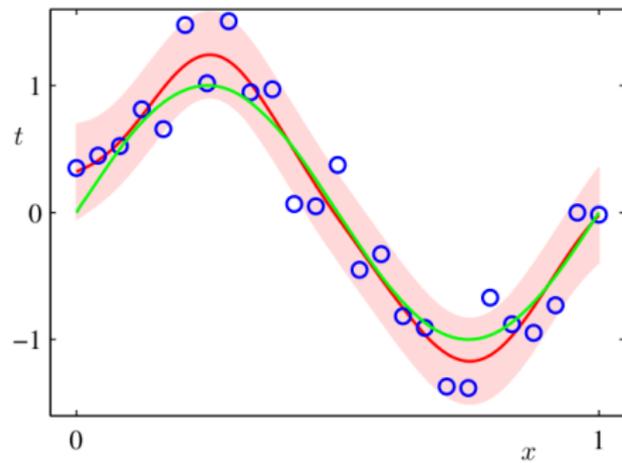
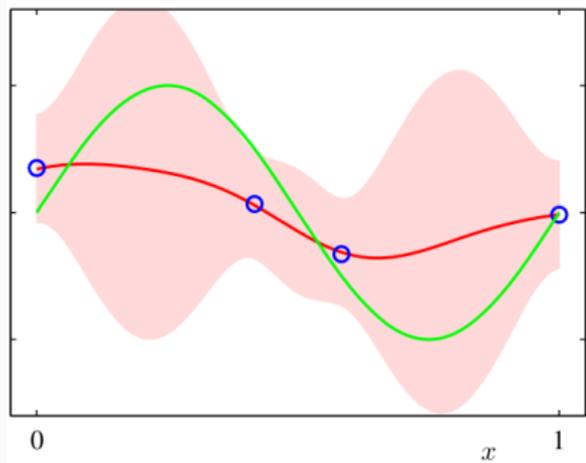
- Т.е. дисперсия складывается из шума в данных β и дисперсии параметров w ; гауссианы независимы, и их дисперсии просто складываются.

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

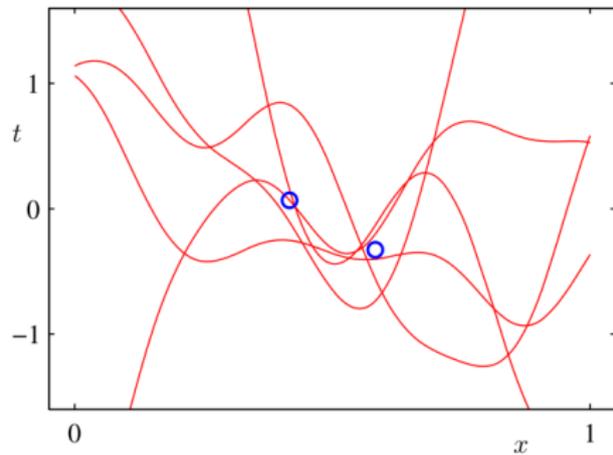
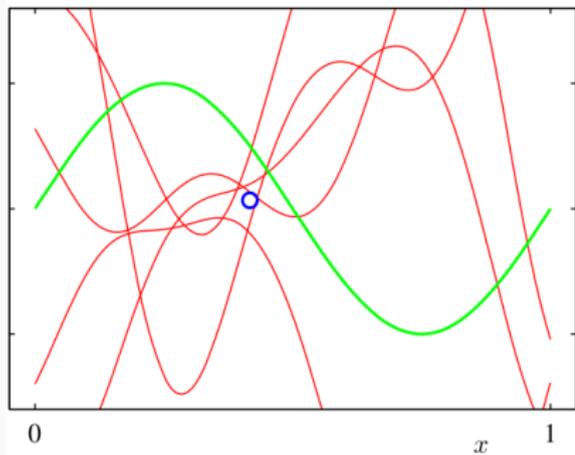
ПРЕДСКАЗАНИЯ



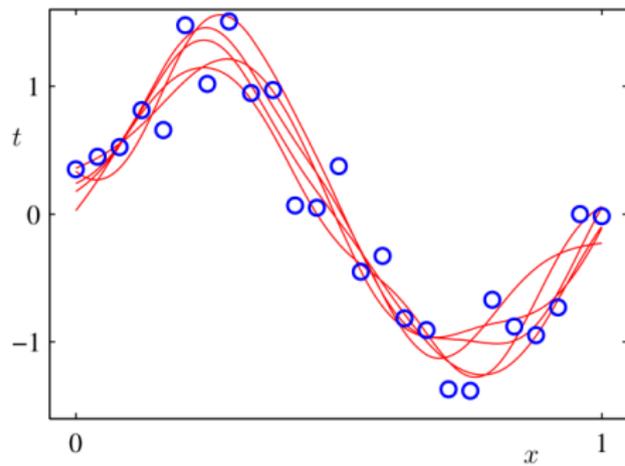
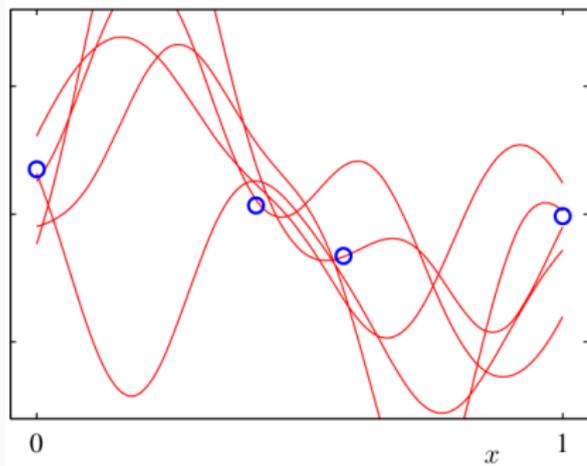
ПРЕДСКАЗАНИЯ



ПРЕДСКАЗАНИЯ



ПРЕДСКАЗАНИЯ



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 σ^2 и будем в качестве параметра рассматривать только μ .

- Сопряжённое априорное распределение для μ при фиксированном σ^2 тоже нормальное и выглядит как

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

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

- При нашем априорном распределении у μ 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

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

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

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

НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ: ФИКСИРУЕМ σ

- 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 μ и менять σ^2 , то сопряжённым априорным распределением будет обратное гамма-распределение:

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

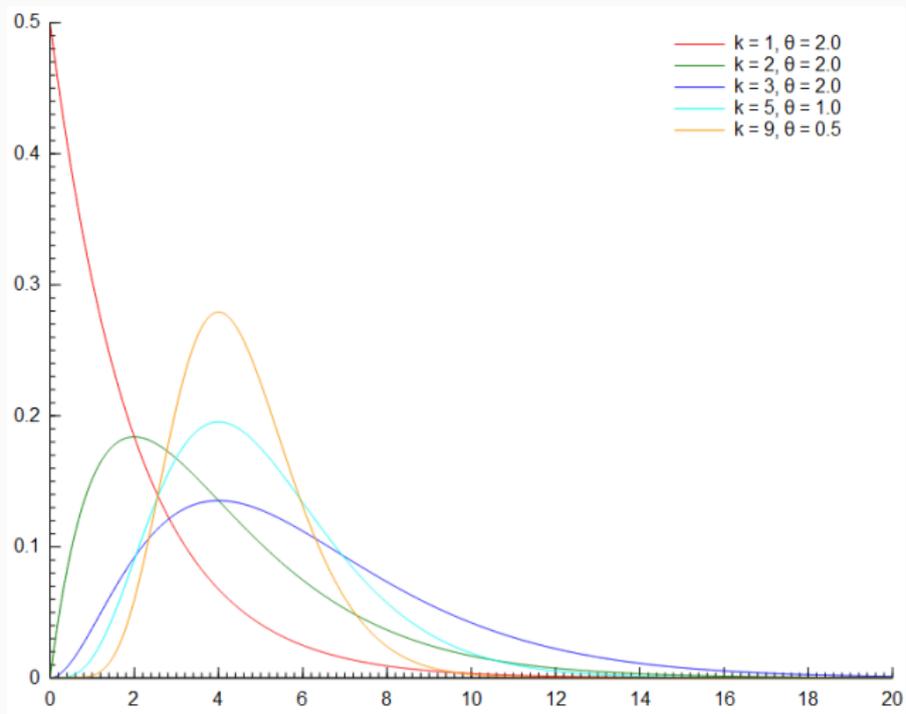
- Тогда в апостериорном распределении будет

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

- 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 μ , И σ^2 МЕНЯЮТСЯ

- Что делать, когда и μ , и σ^2 меняются?
- Можно было бы предположить, что μ и σ^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 μ , И σ^2 МЕНЯЮТСЯ

- Что делать, когда и μ , и σ^2 меняются?
- Можно было бы предположить, что μ и σ^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 μ и σ^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 μ ...

- 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_{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:

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

- η 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 μ_1, \dots, μ_{M-1} получаются

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

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

Упражнение. Проверьте!

- Так вот, для распределений из экспоненциального семейства

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

можно сразу оптом найти сопряжённые априорные распределения:

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

где χ – 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

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

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

- 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 α тестовых примеров, нужно (ожидаемо) покрыть долю α 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.

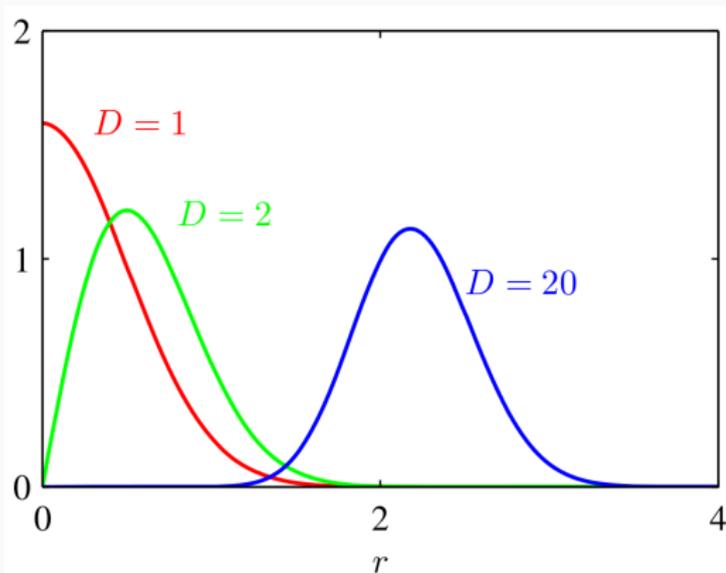
- Большинство точек в результате ближе к границе носителя, чем к другим точкам, а это для ближайших соседей проблема – придётся не интерполировать внутри существующих точек, а экстраполировать наружу.

- Третье проявление: проблемы в оптимизации, которые и имел в виду Беллман.
- Если нужно примерно оптимизировать функцию от d переменных, на решётке с шагом ϵ понадобится примерно $(\frac{1}{\epsilon})^d$ вычислений функции.
- В численном интегрировании – чтобы интегрировать функцию с точностью ϵ , 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 $x \in \mathbb{R}^p$, непрерывный вещественный выход $y \in \mathbb{R}$; у них есть некоторое совместное распределение $p(x, y)$.
- Мы хотим найти функцию $f(x)$, которая лучше всего предсказывает y .

- Введём функцию *потери* (loss function) $L(y, f(x))$, которая наказывает за ошибки; естественно взять квадратичную функцию потери

$$L(y, f(x)) = (y - f(x))^2.$$

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

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

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

- Это можно переписать как

$$\text{EPE}(f) = \mathbb{E}_x \mathbb{E}_{y|x} [(y - f(x))^2 | x],$$

и, значит, можно теперь минимизировать EPE поточечно:

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

а это можно решить и получить

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

- Это решение называется *функцией регрессии* и является наилучшим предсказанием y в любой точке x .

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

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

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

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

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

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

$$f(x) \approx x^\top w.$$

- Теперь мы не берём условие по x , как в k -NN, а просто собираем много значений для разных 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

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

- Перепишем как раньше:

$$\text{EFE} = \mathbb{E}_x \sum_{k=1}^K L(g_k, \hat{g}(x)) p(g_k | x).$$

- Опять достаточно оптимизировать поточечно:

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

- Опять достаточно оптимизировать поточечно:

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

- Для 0-1 функции потери это упрощается до

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

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

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

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

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

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

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

ПОТОМУ ЧТО

$$\int (f(x) - E[y | x]) (E[y | x] - y) p(x, y) dx dy = 0.$$

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

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

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

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

- А noise, $\int (E[y | x] - y)^2 p(x, y) dx dy$, – 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 $\hat{f}(x) = \mathbb{E}[y | x]$, и всё.
- Однако жизнь – борьба, и у нас есть только ограниченный датасет из N точек. Предположим, что этот датасет берётся по распределению $p(x, y)$ – т.е. фактически рассмотрим много-много экспериментов такого вида:
 - взяли датасет D из N точек по распределению $p(x, y)$;
 - подсчитали нашу чудо-регрессию;
 - получили новую функцию предсказания $f(x; D)$.
- Разные датасеты будут приводить к разным функциям предсказания...

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

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

и в ожидании получится...

- ...и в ожидании получится

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

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

Expected loss = (bias)² + variance + noise,

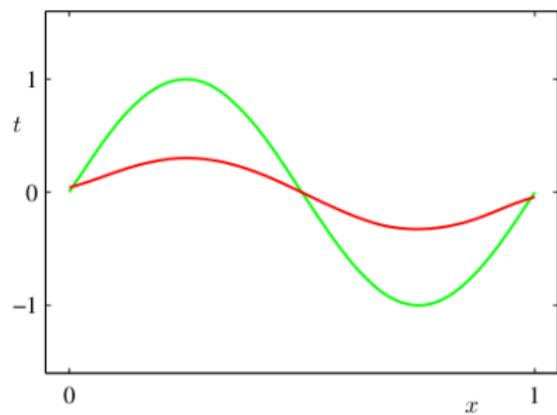
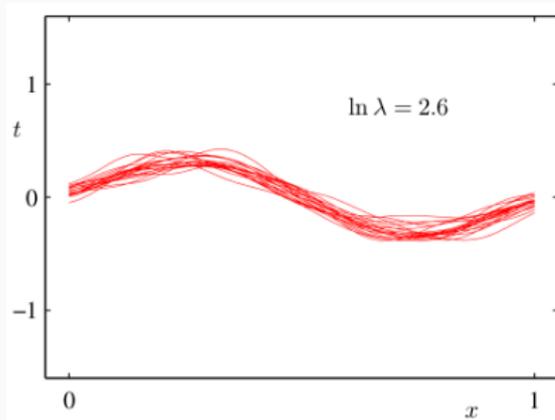
где

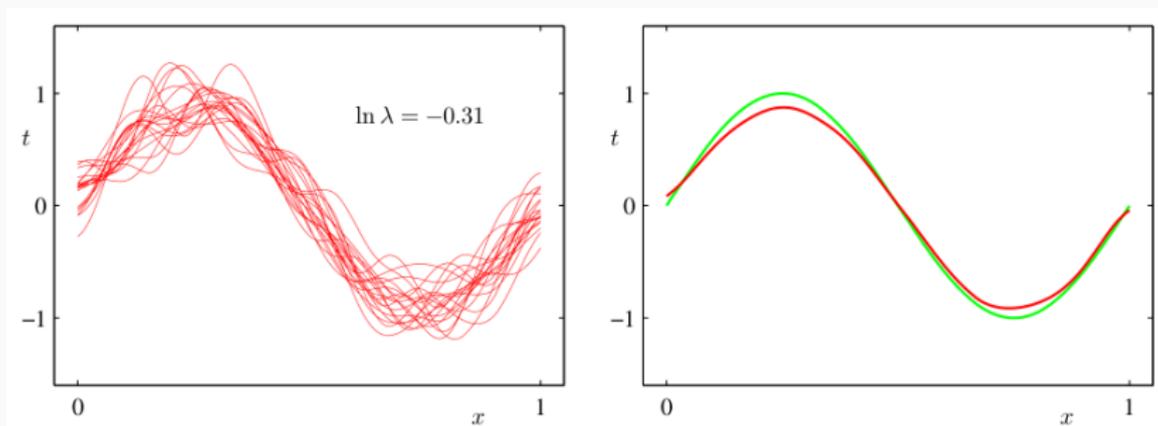
$$(\text{bias})^2 = \left(\mathbb{E}_D [f(x; D)] - \hat{f}(x) \right)^2,$$

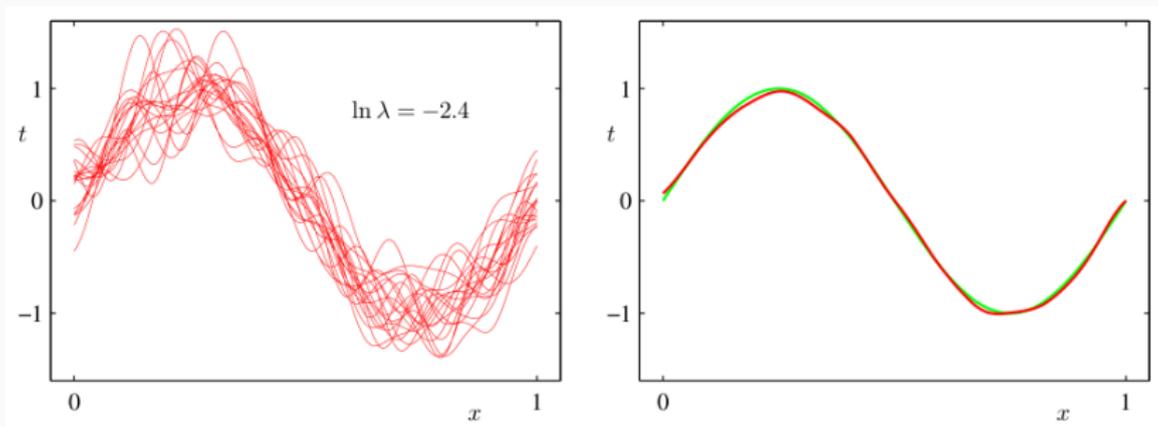
$$\text{variance} = \mathbb{E}_D \left[(f(x; D) - \mathbb{E}_D [f(x; D)])^2 \right],$$

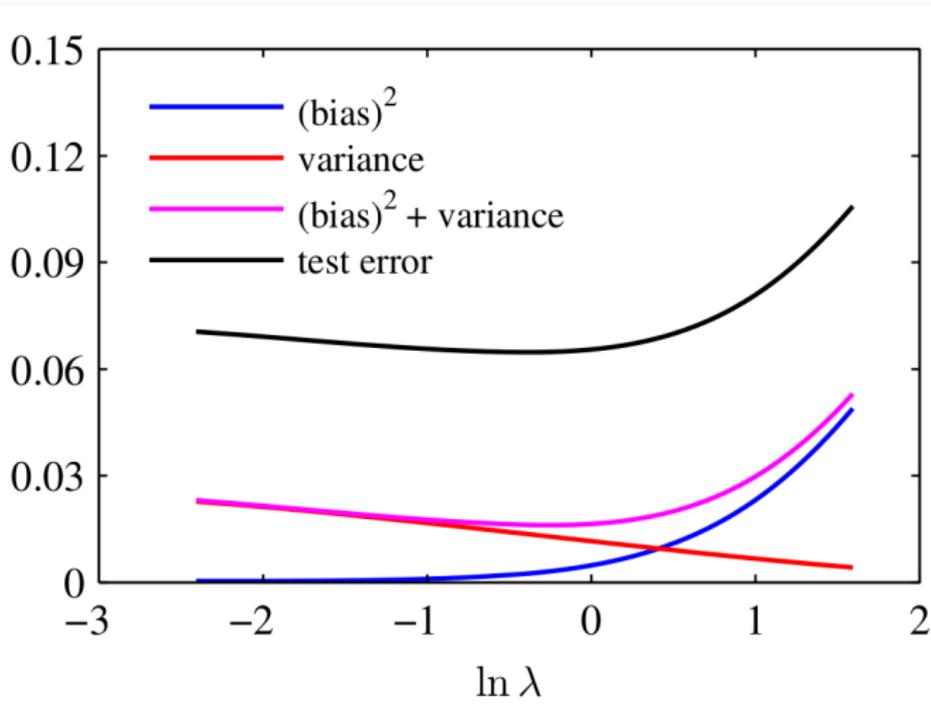
$$\text{noise} = \int (\mathbb{E}[y | x] - y)^2 p(x, y) dx dy.$$

- 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 α .
- Будем набрасывать много датасетов и смотреть, что меняется при этом.









Спасибо за внимание!