Гауссовские процессы

аппроксимация и оптимизация

Антон Лиознов

12 ноября 2018 г.

План

Задача оптимизации

Мотивация

Модель

Регрессия

Оптимизация

Ссылки

- Все материалы по лекции
  https://github.com/Lavton/gauss_process_lecture

- Источник, по которому сделана лекция – школа Deep|Bayes https://habr.com/post/337028/

План

Задача оптимизации

Мотивация

Модель

Регрессия

Оптимизация

Постановка задачи оптимизации

Предположим, у вас есть неизвестная функция $f(x)$, где

$$f(x) \in C \text{ or } C^1$$

$$x \in \mathbb{R}^n$$

Постановка задачи оптимизации

Предположим, у вас есть неизвестная функция $f(x)$, где

$$f(x) \in C \text{ or } C^1$$

$$x \in \mathbb{R}^n$$

Цель: найти такой x, что

$$f(x) \to \min$$

Оракул

Алгоритмы оптимизации работают с понятием **Оракул**.

Определение (Оракул $O(x)$)

это множество значений, которые мы можем получить от функции в точке $x \in \mathbb{R}^n$.

Оракул

Алгоритмы оптимизации работают с понятием **Оракул**.

Определение (Оракул $O(x)$)

это множество значений, которые мы можем получить от функции в точке $x \in \mathbb{R}^n$.

- Оракул нулевого порядка $\equiv \{f(x)\}$
- Оракул первого порядка $\equiv \{f(x), f'(x)\}$
- ...

Сущуствует и более экзотические оракулы: Разделяющий, Стохастический,..

Схема процесса оптимизации

$I_{-1} = \varnothing$. для некого начального x_0 и $k = 0, 1, 2, \ldots$

1. вызвать $O(x_k)$
2. $I_k := I_{k-1} \cup O(x_k)$
3. вычислить x_{k+1} по I_k
4. проверить критерий остановки

Схема процесса оптимизации

$I_{-1} = \varnothing$. для некого начального x_0 и $k = 0, 1, 2, \ldots$

1. вызвать $O(x_k)$
2. $I_k := I_{k-1} \cup O(x_k)$
3. вычислить x_{k+1} по I_k
4. проверить критерий остановки

Большинство методов предполагают у функции какое-то поведение, которое бы позволило быстрее найти минимум. Например, квадратичность $U_{x_*} f(x) \simeq a(x - x_*)^2$, линейность производной $U_{x_*} f'(x) \simeq a x_*$,..

Схема процесса оптимизации

$I_{-1} = \varnothing$. для некого начального x_0 и $k = 0, 1, 2, \ldots$

1. вызвать $O(x_k)$
2. $I_k := I_{k-1} \cup O(x_k)$
3. вычислить x_{k+1} по I_k
4. проверить критерий остановки

Большинство методов предполагают у функции какое-то поведение, которое бы позволило быстрее найти минимум. Например, квадратичность $U_{x_*} f(x) \simeq a(x - x_*)^2$, линейность производной $U_{x_*} f'(x) \simeq a x_*$,..**Или что функция порождена Гауссовским процессом.**

План

Задача оптимизации

Мотивация

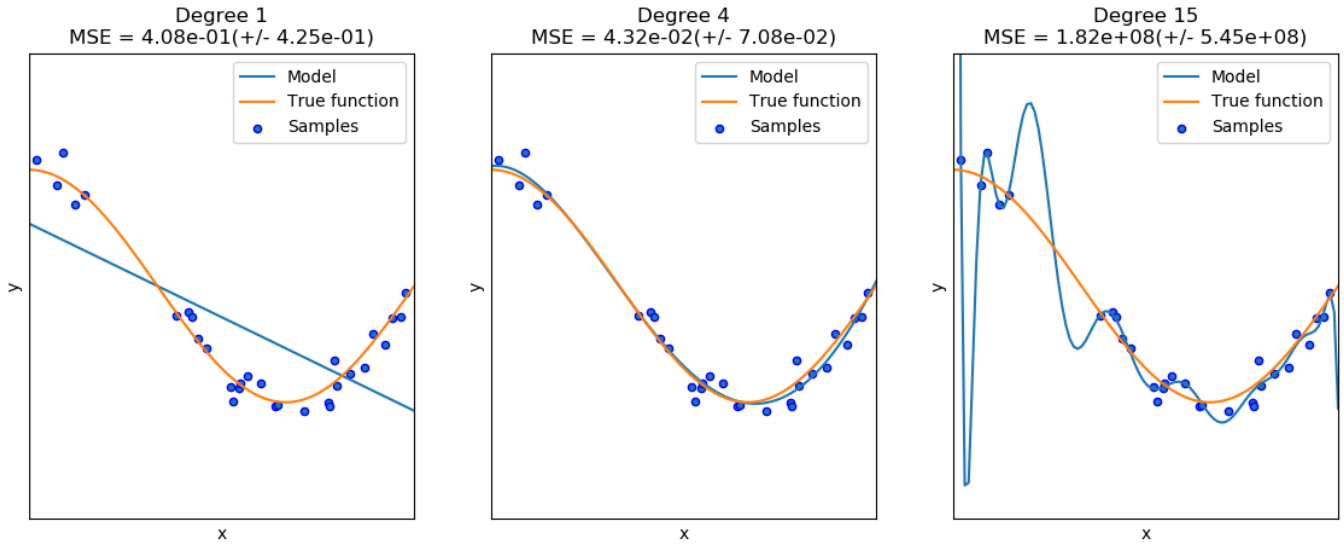
Модель

Регрессия

Оптимизация

Почему не подходит обычная регрессия

Переобучение



https://scikit-learn.org/stable/auto_examples/model_selection/plot_underfitting_overfitting.html

Почему не подходит обычная регрессия

Точность

Пусть вы исследуете функцию энергии реликтовых нейтрино от параметров первичного нуклеосинтеза. Аппроксимируя зависимость, вы получили

$$E_{\nu_e}(x)\big|_{x=x_n} = 10^{-4} eV$$

В чём проблема?

Почему не подходит обычная регрессия

Точность

Пусть вы исследуете функцию энергии реликтовых нейтрино от параметров первичного нуклеосинтеза. Аппроксимируя зависимость, вы получили

$$E_{\nu_e}(x)\big|_{x=x_n} = 10^{-4} eV$$

В чём проблема? Нет погрешности

План

Задача оптимизации

Мотивация

Модель

Регрессия

Оптимизация

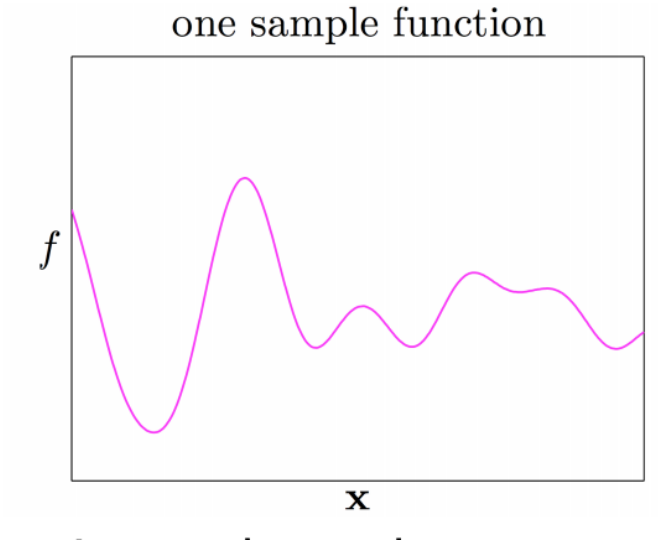
Что такое Гауссовский процесс

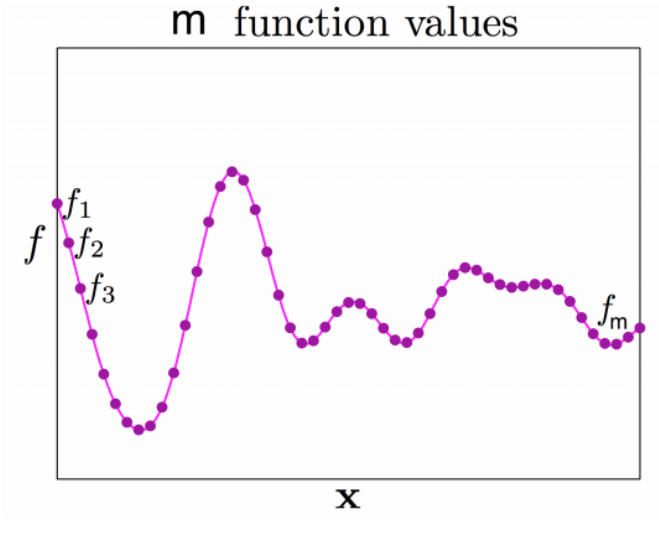
Определение (Гауссовский процесс)

стохастический процесс (совокупность случайных величин, индексированных некоторым параметром, чаще всего временем или координатами), такой что любой конечный набор этих случайных величин имеет **многомерное нормальное распределение**, то есть любая конечная линейная комбинация из них нормально распределена.

https://ru.wikipedia.org/wiki/Гауссовский_процесс

Что такое Гауссовский процесс





Пусть

https://habr.com/post/337028/

$$\mathbf{X} = \{x_1, x_2, \cdots, x_m\}$$ d-мерный набор точек

$$\mathbf{f} = \{f_i | f_i = f(x_i)\}$$ d-мерный набор значений функций

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mu, \mathbf{K})$$

физический смысл параметров

среднее значение $\mu(x)$

Среднее значение $\mu(x) = \{\mu(x_i)\}$

Как правило, мы узнаём среднее значение из физического понимания процесс – например, если мы знаем, что есть какой-то линейный тренд, мы можем его добавить в модель.

А если ничего не знаем, можем положить $\mu = 0$

физический смысл параметров

Ковариационная функция $\mathbf{K}(x_i, x_j)$

Ковариационная функция $\mathbf{K} = \{K(x_i, x_j)\}$

Если точки «достаточно близко» – между ними большая ковариация. Если «далеко» – ковариация низкая. Это задаёт гладкость и характерный масштаб изменений для функции.

Виды ковариационных функций

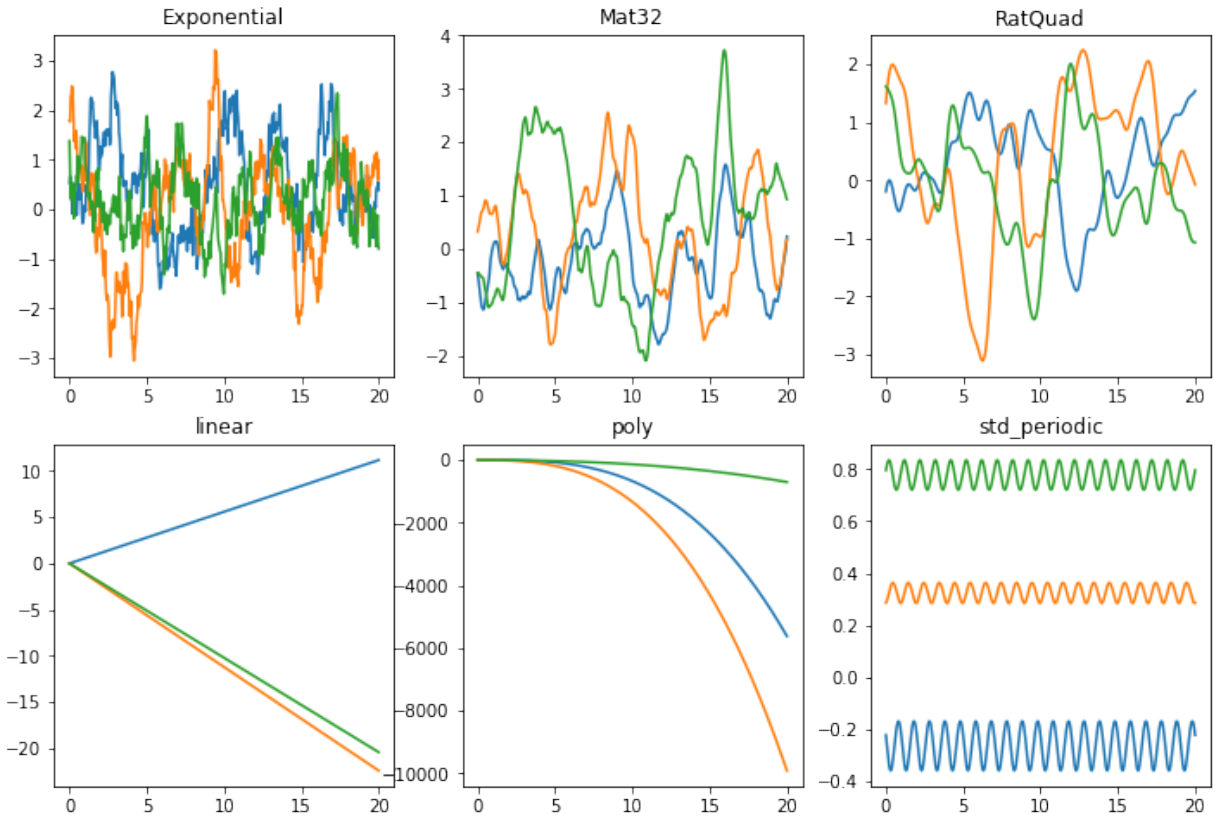
- Константа $K(x, x') = C$
- Линейная $K(x, x') = x^T x'$
- Гауссовский шум $K(x, x') = \sigma^2 \delta(x - x')$
- Матерна

$$K(x, x') = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}|x - x'|}{l} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}|x - x'|}{l} \right)$$

  $\Gamma(\nu)$ – Гамма-функция, K_ν – модифицированная функция Бесселя

- Периодическая

$$K(x, x') = \exp\left(-\frac{2 \sin^2 \frac{x - x'}{2}}{r^2} \right)$$


Exponential
Mat32
RatQuad
linear
poly
std_periodic

Виды ковариационных функций

Гауссовская

- Квадратичная экспоненциальная функция

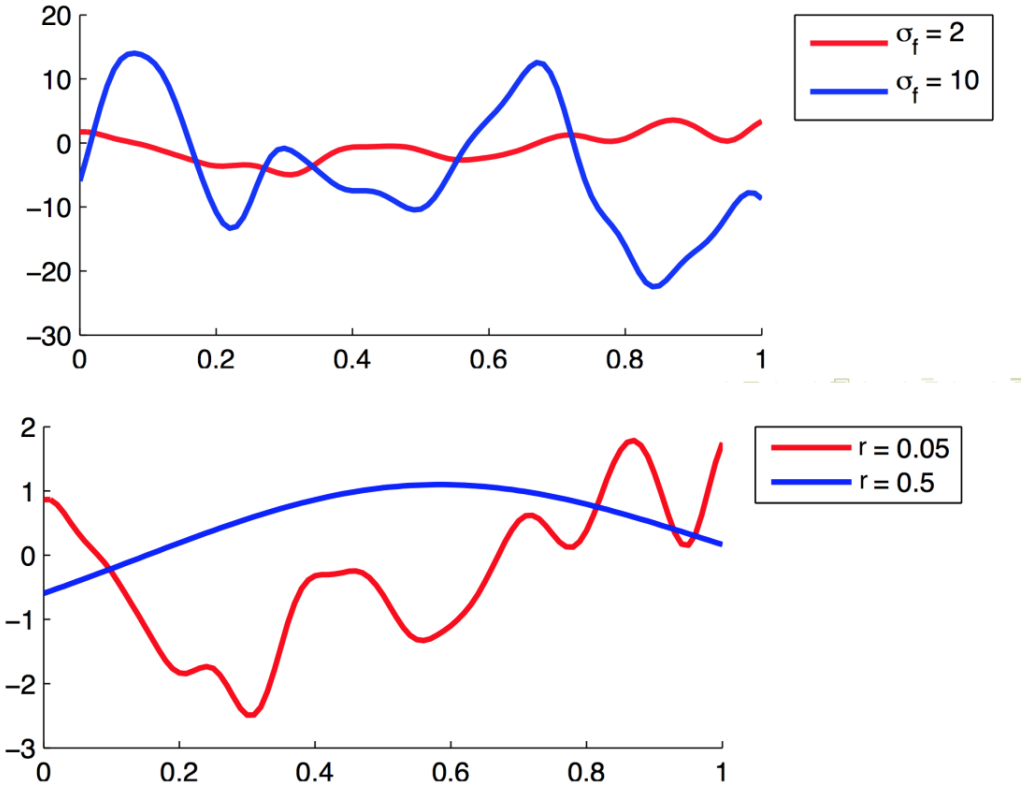
$$K(x, x') = \sigma_f^2 \exp\left\{ -\sum_{i=1}^{d} \frac{(x - x')^2}{2r^2} \right\}$$

- или более общий случай

$$K(x, x') = \sigma_f^2 \exp\left\{ -\sum_{i=1}^{d} \left(\frac{(x - x')}{r} \right)^{\alpha} \right\} + \sigma_1^2 + \sigma_2^2 \delta(x - x')$$

Виды ковариационных функций

Гауссовская



https://habr.com/post/337028/

Виды ковариационных функций

Создание новых

- Сумма $K(x, x') = K_1(x, x') + K_2(x, x')$
- Произведение $K(x, x') = K_1(x, x') \cdot K_2(x, x')$
- Свёртка $K(x, x') = \int dz\, dz'\, h(x, z) K(x, x') h(x', z')$

План

Задача оптимизации

Мотивация

Модель

Регрессия

Оптимизация

Задача регрессии

у нас есть m наблюдений с белым шумом

$$y_i = f(x_i) + \mathcal{N}(0, \sigma^2)$$

Мы предположили, что

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}|0, \mathbf{K})$$

(Условное математическое ожидание – это среднее значение случайной величины относительно условного распределения.)

И шум

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma^2 I_m)$$

Тогда функция правдоподобия

$$p(\mathbf{y}) = \mathcal{N}(0, \mathbf{K} + \sigma^2 I_m)$$

Предсказание

Мы хотим понять, каким будет значение функции f_* в точке x_*.
Построим совместное распределение

$$p(\mathbf{y}, f_*) = \mathcal{N}\left(\begin{vmatrix}0\\0\end{vmatrix}, \begin{vmatrix}\mathbf{K} + \sigma^2 I_m & \{K(x_*, x_i)\}_{i=1}^m \\ \{K(x_*, x_i)\}_{i=1}^m{}^T & K(x_*, x_*)\end{vmatrix}\right)$$

(Дальше обозначим $k_* \equiv \{K(x_*, x_i)\}_{i=1}^m$ и $K_{**} \equiv K(x_*, x_*)$)

Какую функцию мы хотим посчитать?

Предсказание

Мы хотим понять, каким будет значение функции f_* в точке x_*.
Построим совместное распределение

$$p(\mathbf{y}, f_*) = \mathcal{N}\left(\begin{vmatrix}0\\0\end{vmatrix}, \begin{vmatrix}\mathbf{K} + \sigma^2 I_m & \{K(x_*, x_i)\}_{i=1}^m \\ \{K(x_*, x_i)\}_{i=1}^m{}^T & K(x_*, x_*)\end{vmatrix}\right)$$

(Дальше обозначим $k_* \equiv \{K(x_*, x_i)\}_{i=1}^m$ и $K_{**} \equiv K(x_*, x_*)$)

Какую функцию мы хотим посчитать?

$$p(f_*|y)$$

Небольшой факт из теорвера

$$\begin{pmatrix}\mathbf{f_1}\\ \mathbf{f_2}\end{pmatrix} = \mathcal{N}\left(\begin{vmatrix}\mu_1\\ \mu_2\end{vmatrix}, \begin{vmatrix}\Sigma_{11} & \Sigma_{12}\\ \Sigma_{12}^T & \Sigma_{22}\end{vmatrix}\right)$$

$$p(f_1) = \mathcal{N}(f_1|\mu_1, \Sigma_{11})$$

$$p(f_1|f_2) = \mathcal{N}(f_1|\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(f_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T)$$

Предсказание

$$p(f_*|y) = \mathcal{N}(f_*|\mu_*, \sigma_*^2) \quad ,$$

где

$$\mu_* = k_*^T(\mathbf{K} + \sigma^2 I_m)^{-1}\mathbf{y}$$

$$\sigma_*^2 = K_{**} - k_*^T(\mathbf{K} + \sigma^2 I_m)^{-1}k_*$$

Обучение

Используем метод максимального правдоподобия и ищем минимум в зависимости от параметра θ у следующей функции

$$L = -\log p(\mathbf{y}|\theta) = \frac{1}{2}\log\det(\mathbf{K}+\sigma^2 I_m)+\frac{1}{2}y^T(\mathbf{K}+\sigma^2 I_m)^{-1}y+\frac{m}{2}\log 2\pi$$

Регрессия: итог

Плюсы

- Хорошо работает для небольшого числа точек
- Предсказывает не только значение, но и погрешность
- Подборка гиперпараметров происходит на исходной области

Регрессия: итог

Плюсы

- Хорошо работает для небольшого числа точек
- Предсказывает не только значение, но и погрешность
- Подборка гиперпараметров происходит на исходной области

Минусы

- Нужно обращение матрицы, а значит работа за $O(m^3)$

План

Задача оптимизации

Мотивация

Модель

Регрессия

Оптимизация

В каких задачах применяется?

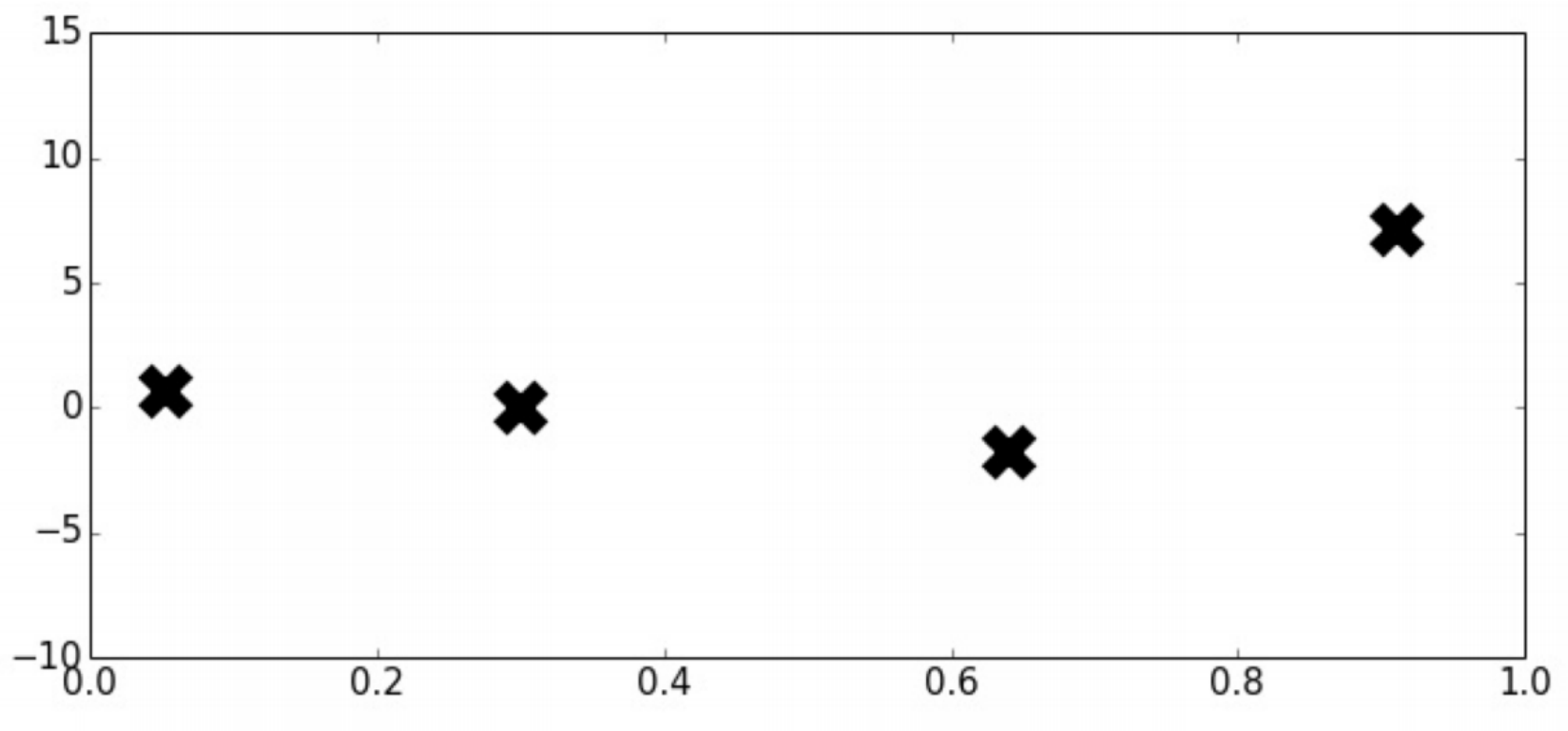
Оптимизация на основе Гауссовских процессов применяется когда:

- ваш оракул нулевого порядка – возвращает лишь значение функции в точке
- оракул считается «дорого»

Это может быть:

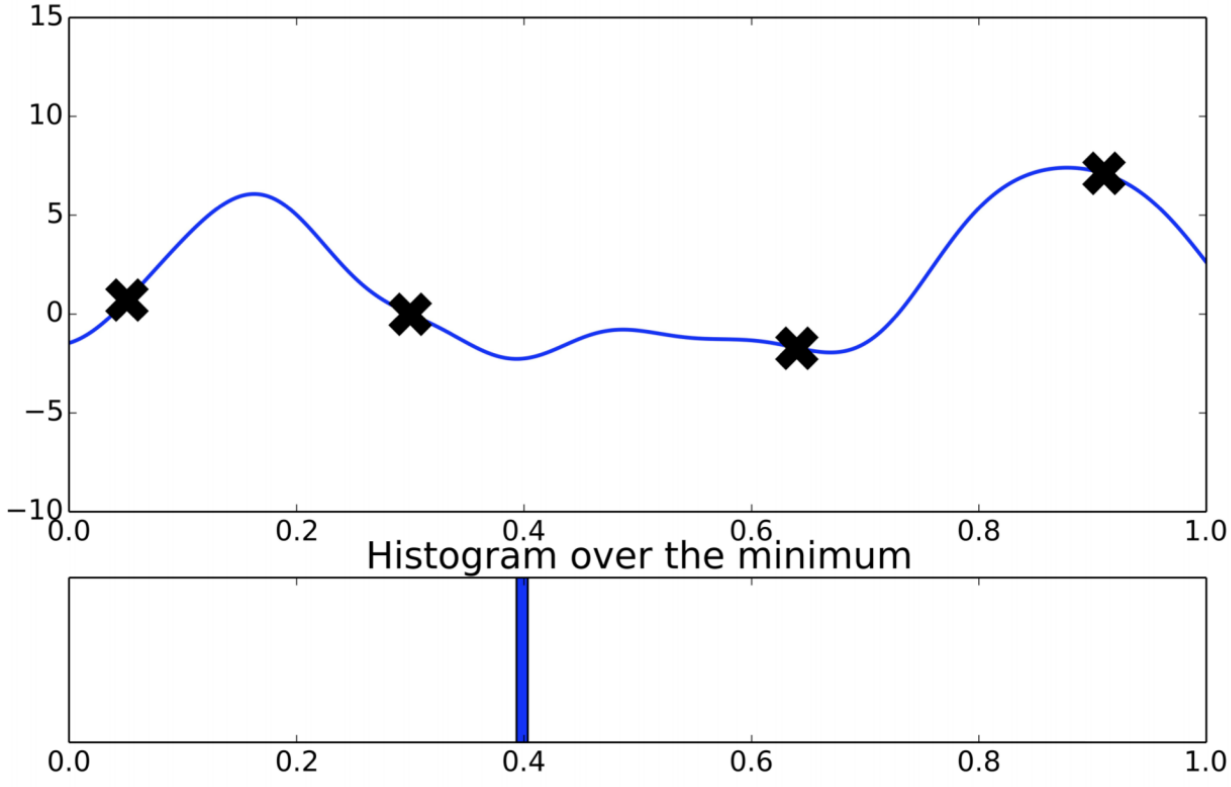
- Подборка гиперпараметров при моделировании
- Проведение сложных экспериментов

Типичная ситуация



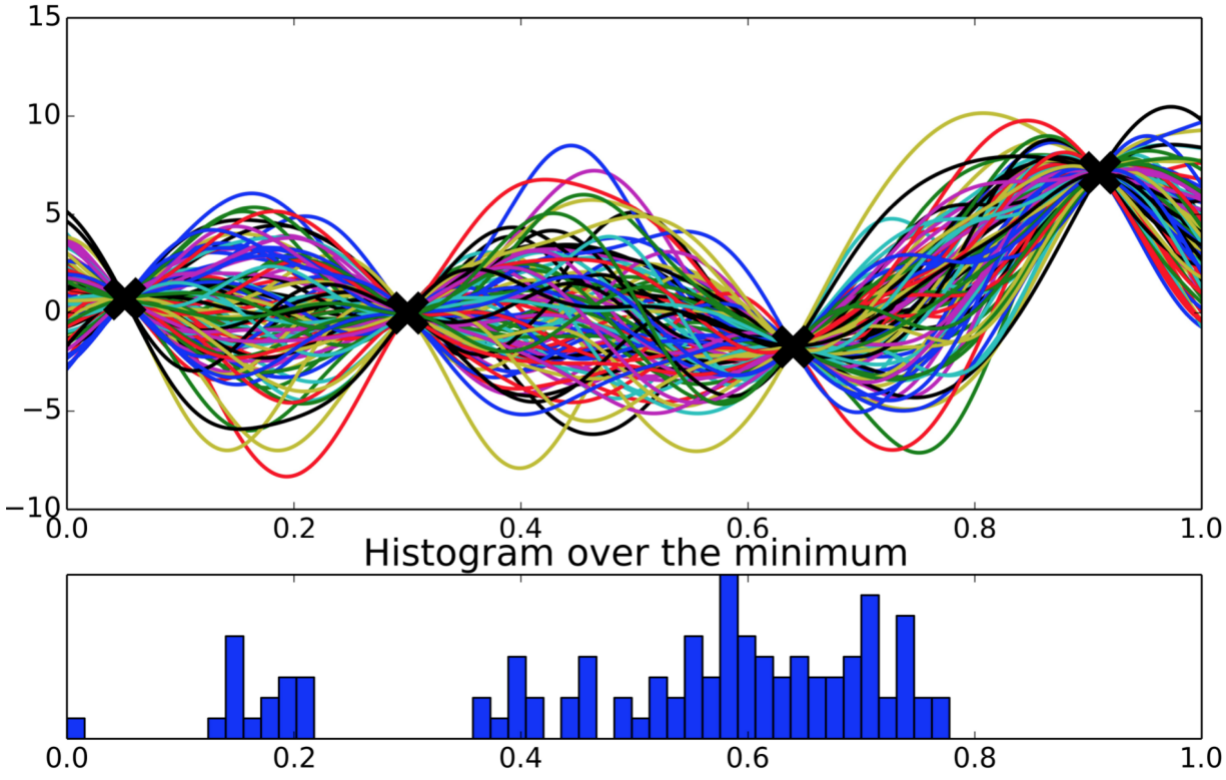
https://habr.com/post/337028/

Построим кривую



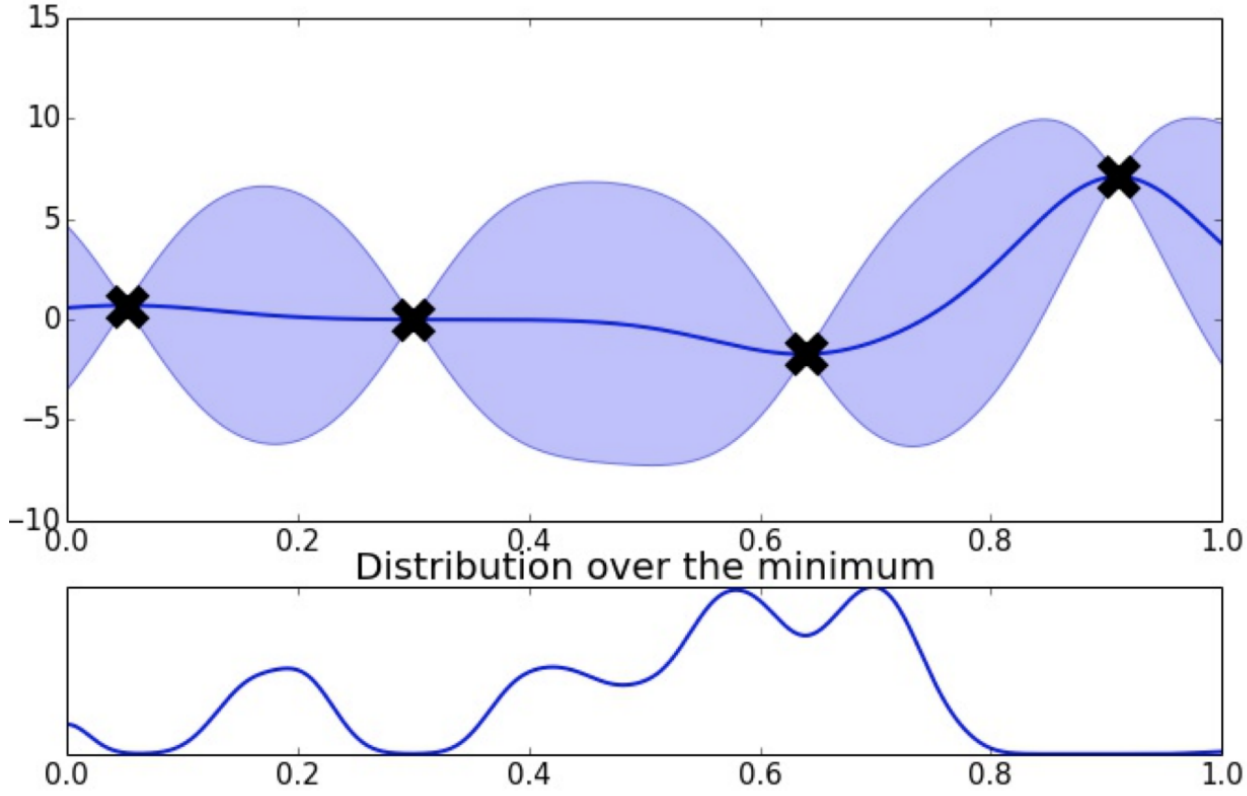
https://habr.com/post/337028/

Много



https://habr.com/post/337028/

Бесконечность



https://habr.com/post/337028/

Как выбрать следующую точку?

Как выбрать следующую точку?

Acquisition functions

функция «сбора данных» Гауссовский процесс в каждой точке имеет предсказанное μ_* и σ_*. Наша цель – сбалансировать предсказание минимума и нашу уверенность в том, что тут что-то есть.

- Exploration (исследование) – поиск точек, где дисперсия велика
- Exploitation (эксплуатация) – поиск точек, где среднее мало

Алгоритмы выбора

- Верхняя (нижняя) граница уверенности: $-\mu_*(x) + c\sigma_*(x)$
- Ожидаемое улучшение (Expected Improvement):
  $\int_y \max(0, y_{best} - y_*) p(y_*|x) dy_* =$
  $\max(0, \Delta(x)) - \sigma_*(x)\phi\left(\frac{\Delta(x)}{\sigma_*(x)}\right) + |\Delta(x)|\Phi\left(\frac{\Delta(x)}{\sigma_*(x)}\right)$, где
  $\Delta(x) = y_{best} - \mu_*(x)$
- Максимальная вероятность улучшения:
  $P(f(x) < y_{best}) = \Phi\left(\frac{\mu_*(x) - y_{best}}{\sigma(x)}\right)$

https://www.cse.wustl.edu/~garnett/cse515t/spring_2015/files/lecture_notes/12.pdf

Что произошло?

Мы заменили задачу, которую не умеем решать – поиск минимума исходной функции задачей, которую умеем решать – поиск минимума функции сбора данных

Практика