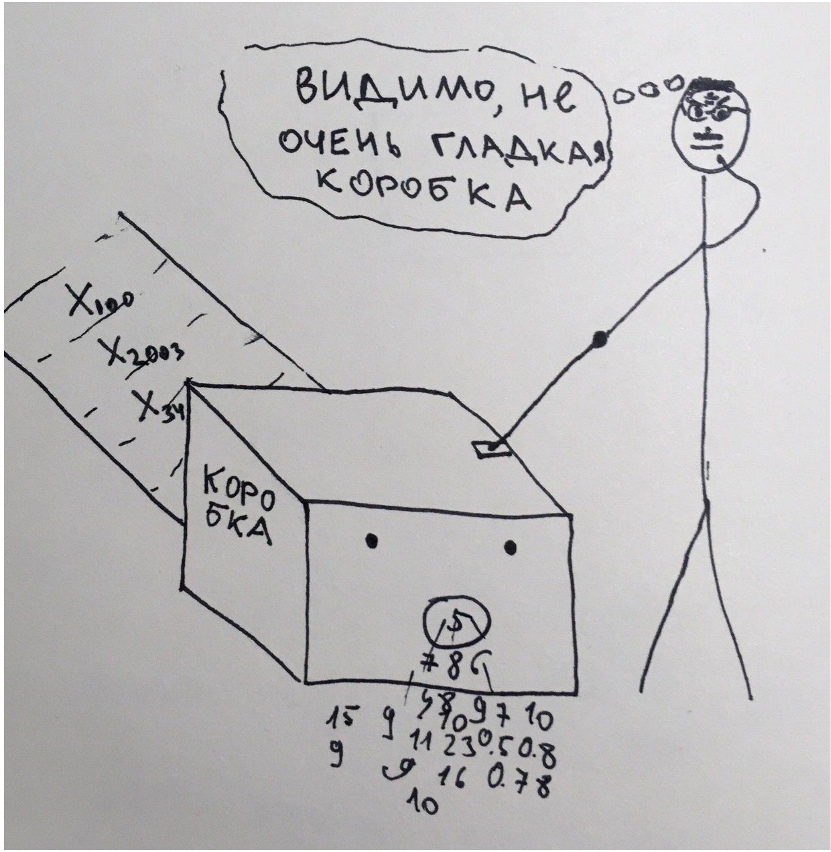
Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design



Максим Манайнен. ФУПМ.

Постановка задачи

$f : X \to \mathbb{R},\ X \subset \mathbb{R}^D$ найдём $x_M = arg\,\mathbf{min}_{x \in X} f(x)$

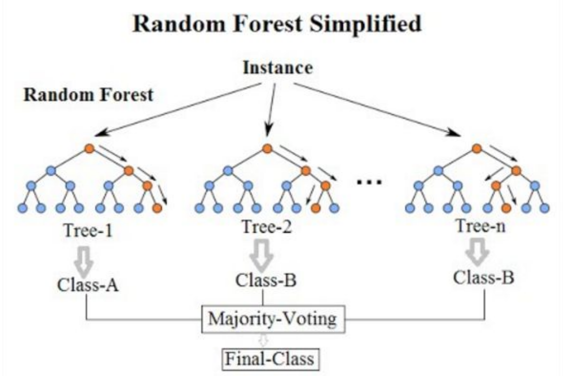
Причём f :

- не задана аналитически и мультимодальна
- шумная
- оценка в точке стоит дорого
- Липшец непрерывна

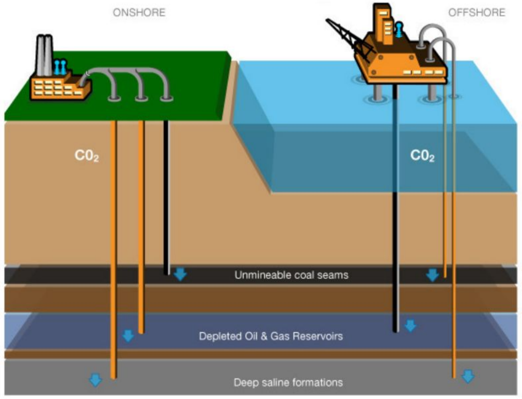
$$\exists L : |f(x) - f(y)| \leq L|x - y|$$

Примеры

- Оптимизация гиперпараметров ML моделей
- Выбор места для бурения новой нефтяной скважины
- Подбор цены на товары в интернет магазине
- Подбор рекламных объявлений для показа пользователю



Will Koehrsen



Richard Wilkinson

гауссовские процессы

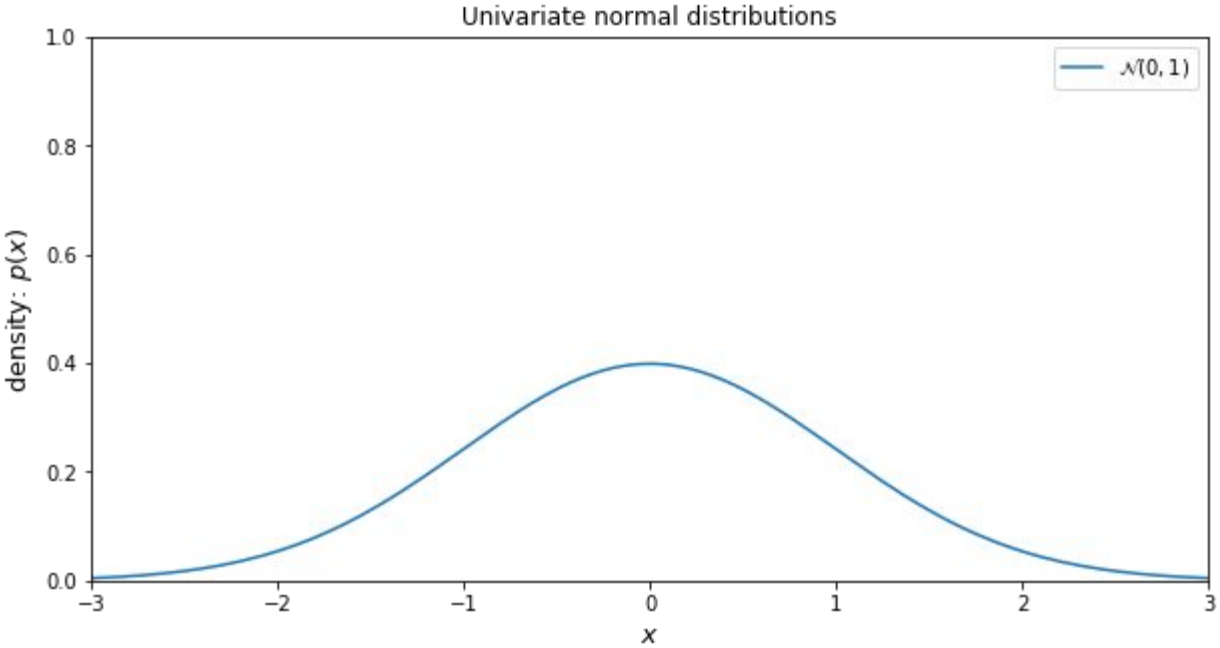
Одномерное гауссовское распределение

$$X \sim N(\mu, \sigma^2)$$

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\mu = \mathbb{E}(X)$$

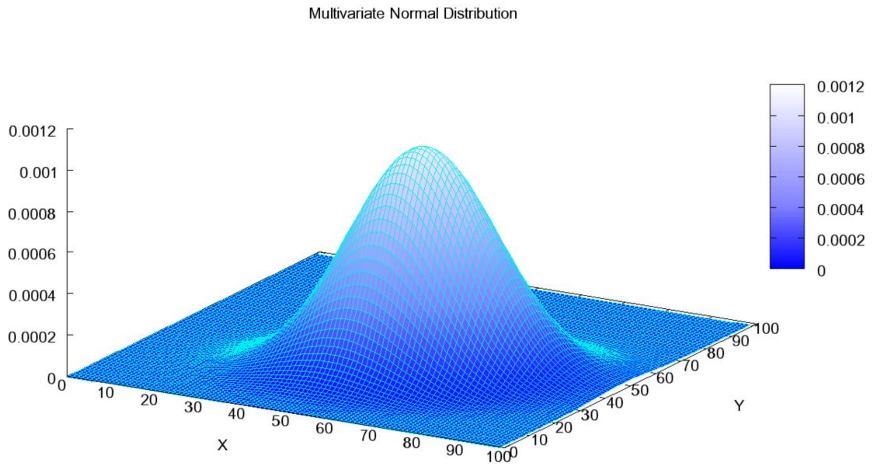
$$Var(x) = \mathbb{E}[\,(X-\mu)^2\,] = \sigma^2$$



Многомерное гауссовское распределение

$$X \in \mathbb{R}^n, \; X \sim N(\mu, \Sigma)$$

$$p(\mathbf{x} \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d \Sigma}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1} (\mathbf{x}-\mu)\right)$$



Wikipedia

Σ - матрица ковариаций

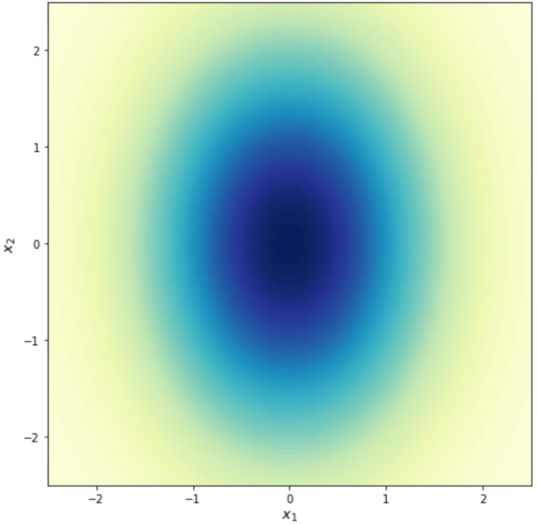
Матрица ковариаций

$$Cov(x_1, x_2) = \mathbb{E}[(x_1 - \mathbb{E}(x_1))(x_2 - \mathbb{E}(x_2))] = \mathbb{E}[x_1 x_2] - \mathbb{E}[x_1]\mathbb{E}[x_2]$$

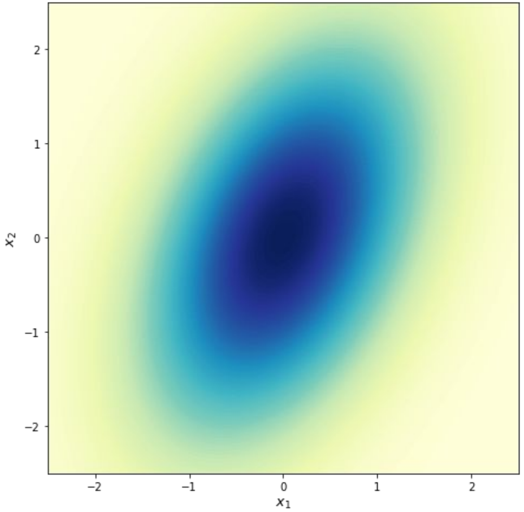
$$\Sigma = \begin{bmatrix} cov(x_1, x_1) & cov(x_1, x_2) & \dots \\ cov(x_2, x_1) & cov(x_2, x_2) & \dots \\ \dots & & \end{bmatrix}$$

- квадратная симметрическая, неотрицательноопределённая матрица

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$$



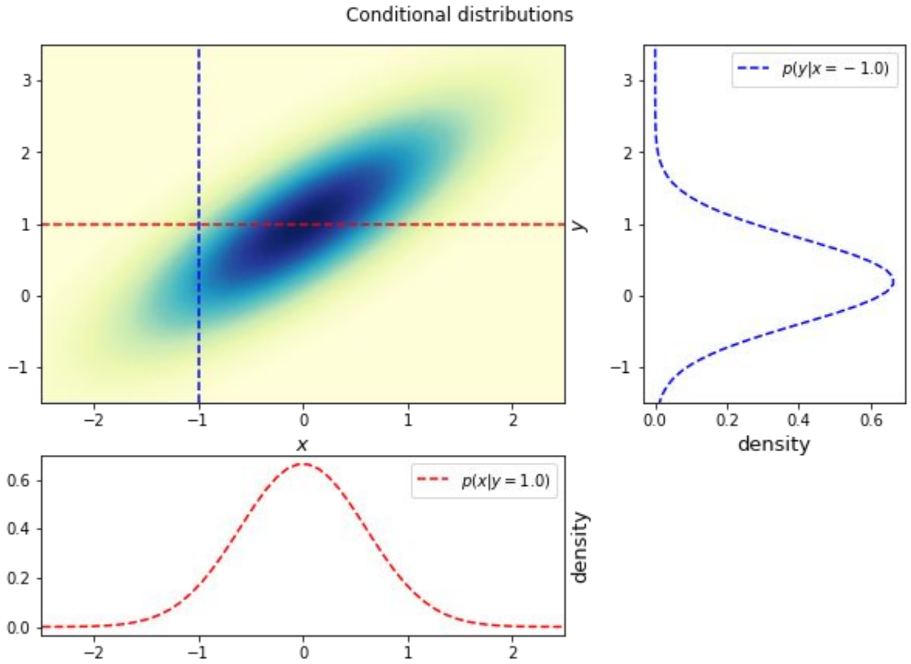
$$\Sigma = \begin{bmatrix} 1 & 0.6 \\ 0.6 & 2 \end{bmatrix}$$



Условная вероятность

$p(\mathbf{x}) = N(\mu, \Sigma)$ - совместное распределение

$p(x|\mathbf{y}) = N(\mu_{x|y}, \Sigma_{x|y})$ - условное распределение компоненты гауссовского случайного вектор - гауссовское



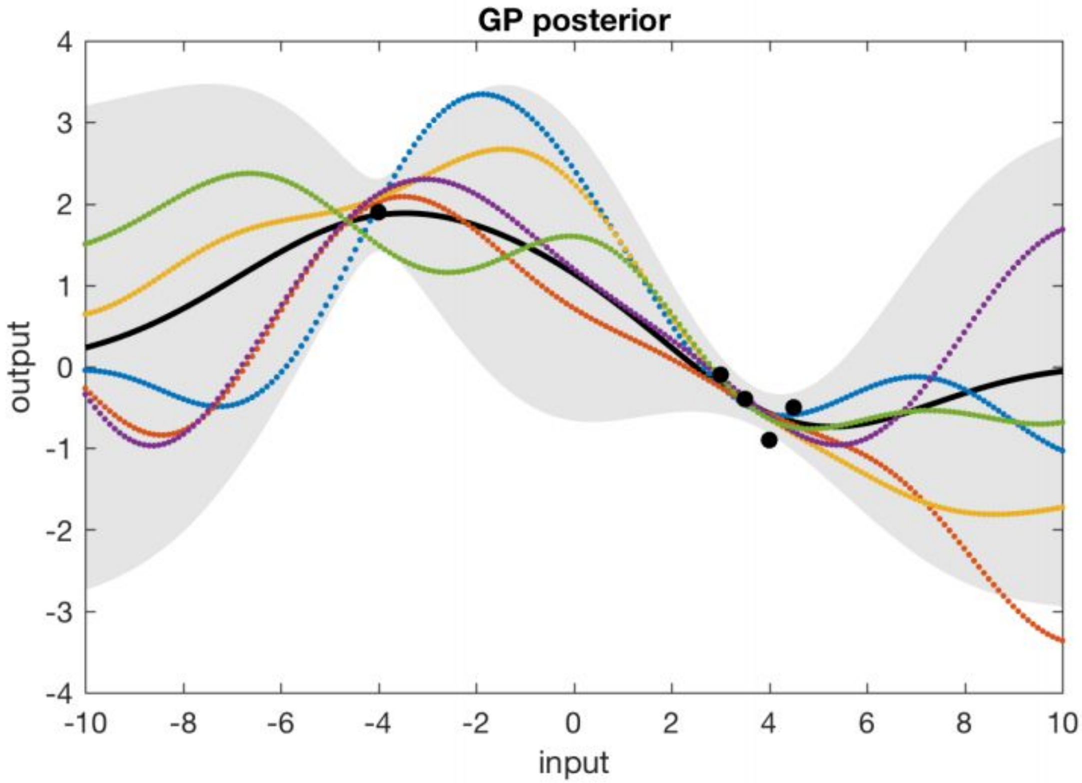
$$\Sigma = \begin{bmatrix} A & C \\ C^T & B \end{bmatrix}$$

$$\Sigma_{x|y} = A - CB^{-1}C^T$$

$$\mu_{x|y} = \mu_x + CB^{-1}(y - \mu_y)$$

Случайный процесс

$$\{Y(x) : x \in X\}$$



Гауссовский случайный процесс

$$\forall n, x_1, \ldots x_n : \vec{Y} \sim \mathbb{N}(\mu_y, \Sigma_y)$$

В задаче регрессии

$$D = \{(x_1, f_1), (x_2, f_2), \ldots, (x_k, f_k)\}$$ - обучающая выборка

$$f(\mathbf{x}) | D \sim N(\mu_*, \Sigma_*)$$ - находим условное распределение

Определение среднего и функции ковариации

- Можем использовать любую функцию среднего $m(\mathbf{x}) = \mathbb{E}(x)$
  обычно используют m(x) = 0

- В качестве функции ковариации обычно используются функции от расстояния между точками

$$k(x, y) = Cov(f(x), f(y))$$

Таким образом функция ковариации определяет гладкость функции - определяя константу Липшеца в выражении

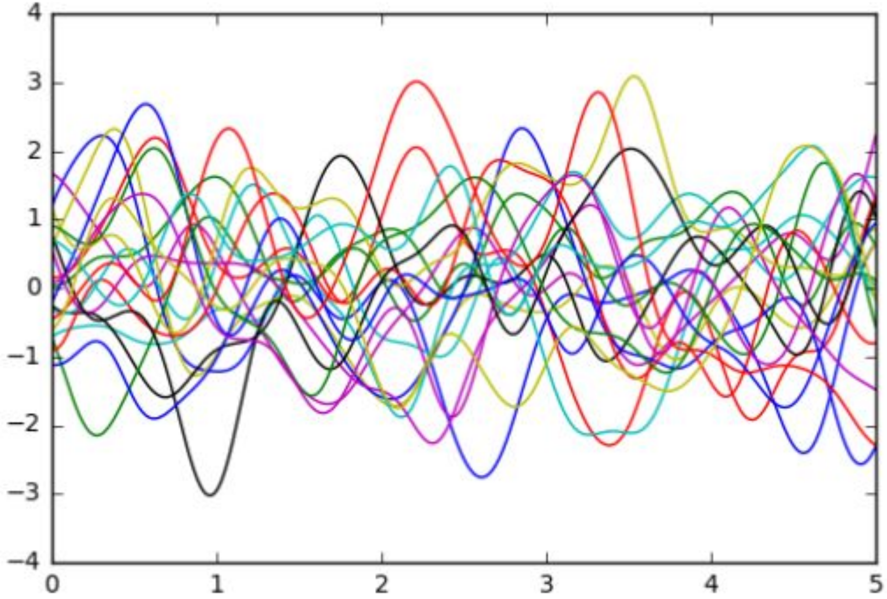
$$|f(x) - f(y)| \leq L|x - y|$$

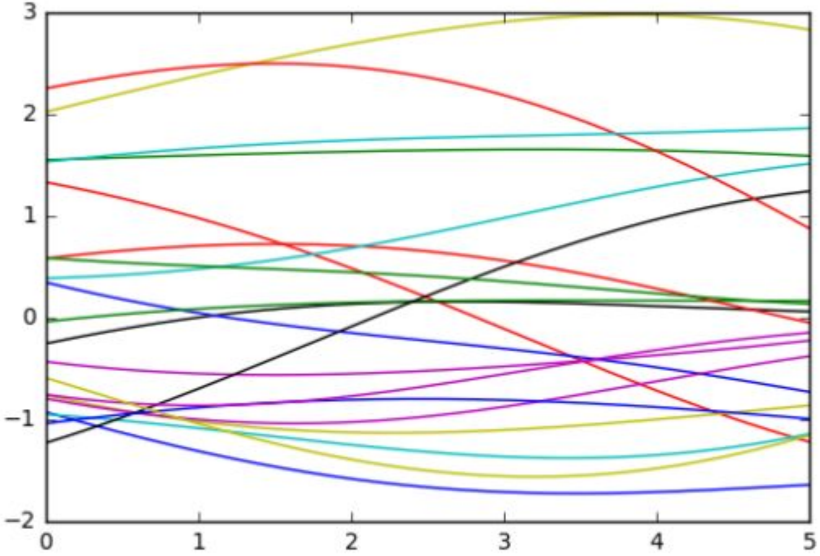
Пример: RBF

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp(-\frac{1}{2l^2}(\mathbf{x}_i - \mathbf{x}_j)^T(\mathbf{x}_i - \mathbf{x}_j))$$

$$k(x, x') = \exp\left(-\frac{1}{2}\frac{(x-x')^2}{0.25^2}\right)$$

$$k(x, x') = \exp\left(-\frac{1}{2}\frac{(x-x')^2}{4^2}\right)$$

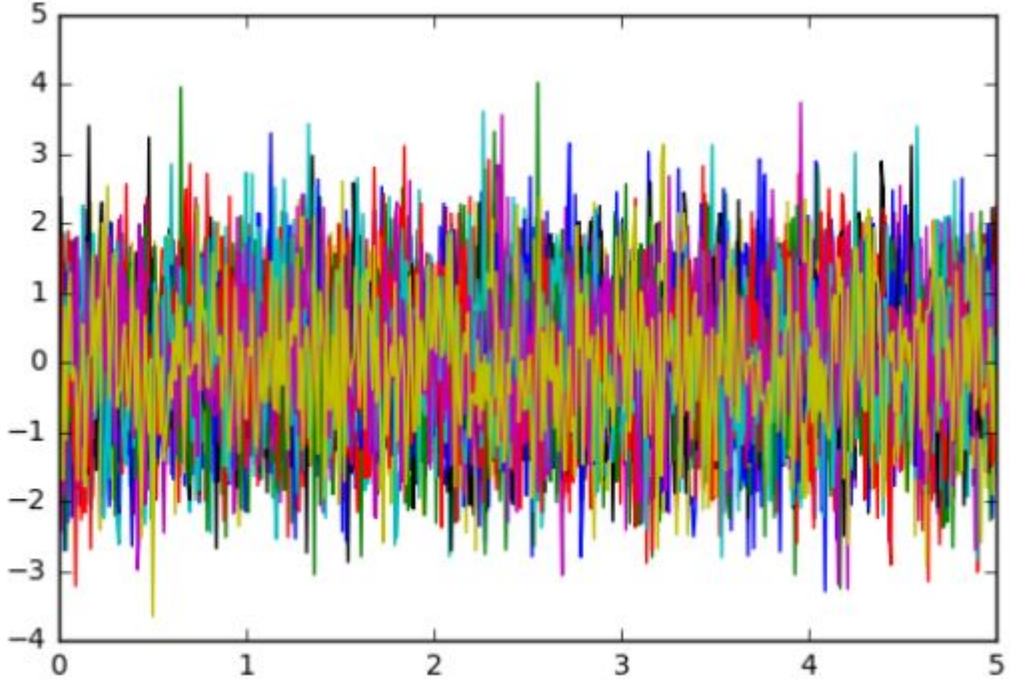




Richard Wilkinson

Пример: белый шум

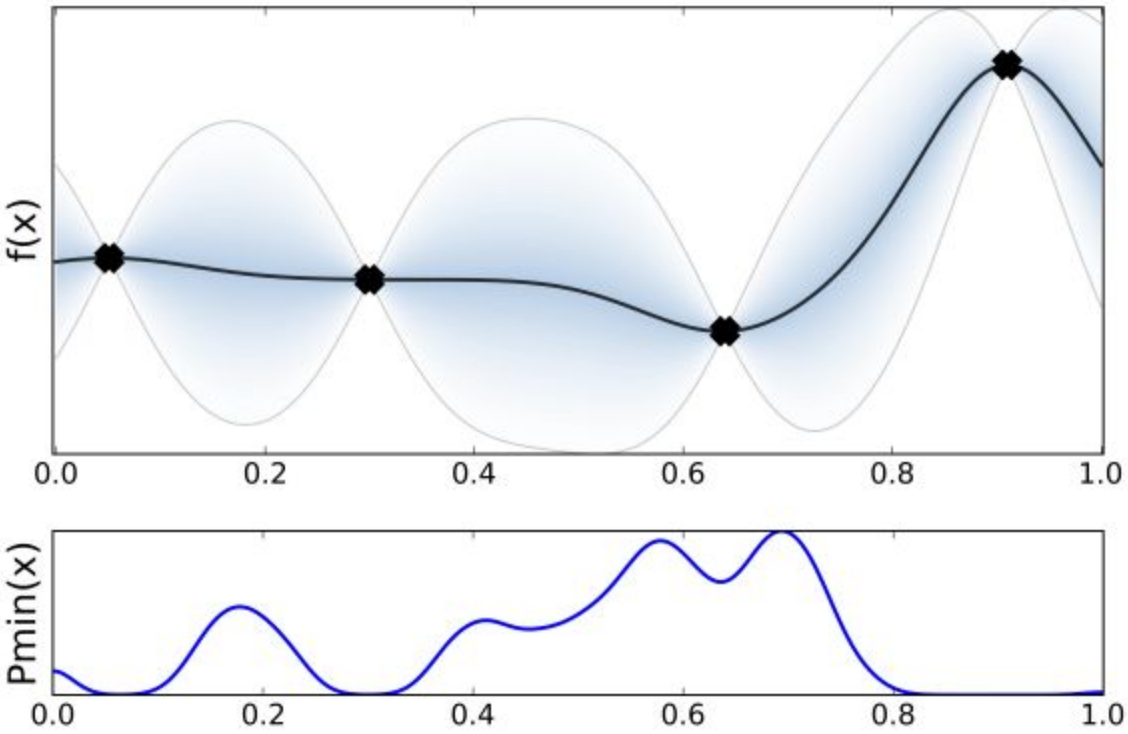
$$k(x, y) = \begin{cases} 1, & \text{if x = y} \\ 0, & \text{otherwise} \end{cases}$$



Richard Wilkinson

байесовская оптимизация

Байесовская оптимизация: интуитивно



Javier Gonzalez

Байесовская оптимизация: алгоритм

1. Определяем **априорное** распределение на возможные функции
2. Обновляем распределение с учётом наблюдений
3. Выбираем новые места для сэмплирования с помощью **acquisition** функции
4. Добавляем данные выборку

Повторяем пункты 2-4, пока возможно.

задача оптимизации -> задача выбора

Формализуем оптимизацию выбора

Мы свели задачу оптимизации к задаче многорукого бандита

$$r_t = f(\mathbf{x}^*) - f(\mathbf{x}_t)$$

$$R_T = \sum_{t=1}^{T} r_t$$ - cumulative regret

Теория информации

Энтропия

$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$ - для дискретной величины

$h(X) = -\int_X f(x) \log_2 f(x) dx$ - для непрерывной

$H(N(\mu, \mathbf{\Sigma})) = \frac{1}{2} \log |2\pi e \mathbf{\Sigma}|$ - энтропия нормально распределенной величины

Прирост информации (information gain)

$$I(X;Y) = H(X) - H(X|Y)$$

Байесовский дизайн экспериментов

$$A \subset D$$

$$\mathbf{y}_A = \mathbf{f}_A + \epsilon_A,\ \epsilon_A \sim N(\mathbf{0}, \sigma^2 \mathbf{I})$$

$$I(\mathbf{y}_A; f) = H(\mathbf{y}_A) - H(\mathbf{y}_A | f)$$

$\mathbf{y}_A$ - распределения на значения, полученные из ГП

$\mathbf{y}_A | f$ - распределения, полученные от оракула

$$\Downarrow$$

$$I(\mathbf{y}_A; f) = \frac{1}{2} \log \left| \mathbf{I} + \sigma^{-2} \mathbf{K}_A \right|$$

Необходимо найти A, максимизирующее данное выражение.

Оптимизация information gain’a

$I(\mathbf{y}_A, f)$ является submodular функцией, т.е.

$$\forall X, Y \subset \Omega,\ X \subset Y \ \rightarrow\ f(X \cup \{x\}) - f(X) \geq f(Y \cup \{x\}) - f(Y)$$

значит можно решить её жадным алгоритмом с хорошей точностью

$$\mathbf{x}_t = argmax_{x \in D} F(A_{t-1} \cup \mathbf{x})$$
$$F(A_T) \geq (1 - 1/e) \max_{|A| \leq T} F(A)$$

$$\Downarrow$$

$$\mathbf{x}_t = arg\max_{\mathbf{x} \in D} \sigma_{t-1}(\mathbf{x})$$

Near-optimal Nonmyopic Value of Information in Graphical Models, A. Krause, C. Guestrin

exploration/exploitation tradeoff

$$\mathbf{x}_t = argmax_{x \in D} \mu_{t-1}(\mathbf{x})$$

- оптимизируясь по такому правилу рискуем застрять в локальном оптимуме

нужно совмещать

GP-Upper Confidence Bound

$$\mathbf{x}_t = argmax_{x \in D} \mu_{t-1}(\mathbf{x}) + \beta_t^{1/2} \sigma_{t-1}(\mathbf{x})$$

β_t - параметр уверенности

выбираем точку с максимальной вероятной оценкой на значение функции оценки точности

Общие обозначения

$$\gamma_T := \max_{A \subset D : |A| = T} I(\mathbf{y}_A; \mathbf{f}_A)$$

максимальный information gain (его тоже нужно оценить)

O^* - O асимптотика с точностью до логарифма

Для конечного D

Th.

$$\delta \in (0,1), \beta_t = 2\log(|D|t^2\pi^2/6\delta)$$

Cumulative regret алгоритма GP-UCB с параметром уверенности β_t медианой 0, функцией ковариации $k(\mathbf{x}, \mathbf{x}^*)$ ограничен $O^*(\sqrt{T\gamma_T \log|D|})$ с высокой вероятностью.

$$\mathbb{P}\{R_T \le \sqrt{C_1 T \beta_T \gamma_T}\ \forall T \ge 1\} \ge 1 - \delta$$

Lemma1. : Возьмём $\delta \in (0,1)$, $\beta_t = 2\log(|D|\pi_t/\sigma)$, где $\sum_{t\ge1}\pi_t^{-1} = 1, \pi_t > 0 \Rightarrow$

$$|f(\mathbf{x}) - \mu_{t-1}(\mathbf{x})| \le \beta_t^{1/2}\sigma_{t-1}(\mathbf{x})\ \forall \mathbf{x} \in D \forall t \ge 1$$

справедливо с вероятностью $\ge 1 - \delta$

Lemma2. : Зафиксируем $t \ge 1$. Если $\forall \mathbf{x} \in D:\ |f(\mathbf{x}) - \mu_{t-1}(\mathbf{x})| \le \beta_t^{1/2}\sigma_{t-1}(\mathbf{x}) \Rightarrow$ проигрыш $r_t \le 2\beta_t^{1/2}\sigma_{t-1}(\mathbf{x}_t)$.

Lemma3. : Можно выразить информационный прирост через отклонение распределения. $\mathbf{f}_T = (f(\mathbf{x}_t)) \in \mathbb{R}^T$:

$$I(\mathbf{y}_T; \mathbf{f}_T) = \frac{1}{2}\sum_{t=1}^{T}\log(1 + \sigma^{-2}\sigma_{t-1}^2(\mathbf{x}_t))$$

Lemma4. : С вероятностью $\ge 1 - \delta$:

$$\sum_{t=1}^{T} r_t^2 \le \beta_T C_1 I(\mathbf{y}_T; \mathbf{f}_T) \le C_1\beta_T\gamma_T\ \ \forall T \ge 1$$

где $C_1 := 8/\log(1 + \sigma^{-2}) \ge 8\sigma^2$

неравенство Коши-Буняковского

Lemma1. : Возьмём $\delta \in (0,1)$, $\beta_t = 2\log(|D|\pi_t/\sigma)$, где $\sum_{t\geq 1} \pi_t^{-1} = 1, \pi_t > 0 \Rightarrow$

$$|f(\mathbf{x}) - \mu_{t-1}(\mathbf{x})| \leq \beta_t^{1/2}\sigma_{t-1}(\mathbf{x}) \ \forall \mathbf{x} \in D \forall t \geq 1$$

справедливо с вероятностью $\geq 1 - \delta$

- Зафиксируем $t \geq 1$ и $\mathbf{x} \in D$. $f(\mathbf{x}) \sim N(\mu_{t-1}(\mathbf{x}), \sigma_{t-1}^2(\mathbf{x}))$.
- $r \sim N(0,1)$ тогда
  $\mathbb{P}\{r > c\} = \frac{1}{\sqrt{2\pi}} \int_c^{+\infty} e^{\frac{1}{2}r^2} dr = \frac{1}{\sqrt{2\pi}} e^{-c^2/2} \int_c^{+\infty} e^{\frac{1}{2}(r-c)^2 - c(r-c)} dr \leq |$ т.к. $e^{-c(r-c)} \leq 1$
  при $r \geq c$ и $\int_c^{+\infty} e^{-(r-c)^2} dr = \int_0^{+\infty} e^{-p^2} dp \ | \leq e^{-c^2/2}\mathbb{P}\{r > 0\} = (1/2)e^{-c^2/2}$
- Приведём распределение $f(\mathbf{x})$ к стандартному $r = f(\mathbf{x}) - \mu_{t-1}(\mathbf{x})/\sigma_{t-1}(\mathbf{x}) \sim N(0,1)$ и $c = \beta_t^{1/2}$ получаем $\mathbb{P}\{|f(\mathbf{x}) - \mu_{t-1}(\mathbf{x})| > \beta_t^{1/2}\sigma_{t-1}(\mathbf{x})\} \leq e^{-\beta_t/2}$
- Далее, используя правило union bound ($\mathbb{P}\{\cup A_i\} \leq \sum \mathbb{P}\{A_i\}$) получаем
  $$|f(\mathbf{x}) - \mu_{t-1}(\mathbf{x})| \leq \beta_t^{1/2}\sigma_{t-1}(\mathbf{x}) \ \forall \mathbf{x} \in D$$
  с вероятностью $\geq 1 - |D|e^{-\beta_t/2}$
- Возьмём $|D|e^{-\beta_t/2} = \delta/\pi_t$ и применяя union bound для $t \in \mathbb{N}$ получаем доказываемое удтверждение.

Lemma2. : Зафиксируем $t \geq 1$. Если $\forall \mathbf{x} \in D: \; |f(\mathbf{x}) - \mu_{t-1}(\mathbf{x})| \leq \beta_t^{1/2}\sigma_{t-1}(\mathbf{x}) \Rightarrow$ проигрыш $r_t \leq 2\beta_t^{1/2}\sigma_{t-1}(\mathbf{x}_t)$.

- $\mathbf{x_t}$ максимизирует соответствующее выражение, поэтому получаем, что $\mu_{t-1}(\mathbf{x_t}) + \beta_t^{1/2}\sigma_{t-1}(\mathbf{x}^*) \geq \mu_{t-1}(\mathbf{x}^*) + \beta_t^{1/2}\sigma_{t-1}(\mathbf{x}^*) \geq f(\mathbf{x}^*)$
- Отсюда получаем

$$r_t = f(\mathbf{x}^*) - f(\mathbf{x}_t) \leq \beta_t^{1/2}\sigma_{t-1}(\mathbf{x}_t) + \mu_{t-1}(\mathbf{x}_t - f(\mathbf{x}_t)) \leq 2\beta_t^{1/2}\sigma_{t-1}(\mathbf{x}_t)$$

Lemma3. : Можно выразить информационный прирост через отклонение распределения. $\mathbf{f}_T = (f(\mathbf{x}_t)) \in \mathbb{R}^T$:

$$I(\mathbf{y}_T; \mathbf{f}_T) = \frac{1}{2} \sum_{t=1}^{T} \log(1 + \sigma^{-2} \sigma_{t-1}^2(\mathbf{x}_t))$$

- $I(\mathbf{y}_T; \mathbf{f}_T) = H(\mathbf{y}_T) - (1/2) \log |2\pi e \sigma^2 E|$.
  $H(\mathbf{y}_T) = H(\mathbf{y}_{T-1}) + H(\mathbf{y}_{T-1}) = H(\mathbf{y}_{T-1}) + 1/2 \log(2\pi e(\sigma^2 + \sigma_{t-1}^2(\mathbf{x}_T)))$ далее по индукции.

Lemma4. : С вероятностью $\geq 1-\delta$:

$$\sum_{t=1}^{T} r_t^2 \leq \beta_T C_1 I(\mathbf{y}_T; \mathbf{f}_T) \leq C_1 \beta_T \gamma_T \ \forall T \geq 1$$

где $C_1 := 8/\log(1+\sigma^{-2}) \geq 8\sigma^2$

- Из предыдущих лемм имеем, что $\mathbb{P}\{r_t^2 \leq 4\beta_t \sigma_{t-1}^2(\mathbf{x}_t) \forall t \geq 1\} \geq 1-\delta$
- β_t - неуменьшающаяся функция
  $\Rightarrow 4\beta_t \sigma_{t-1}^2(\mathbf{x}_t) \leq 4\beta_T \sigma^2(\sigma^{-2}\sigma_{t-1}^2(\mathbf{x}_t)) \leq 4\beta_T \sigma^2 C_2 \log(1+\sigma^{-2}\sigma_{t-1}^2(\mathbf{x}_t))$ где
  $C_2 = \sigma^{-2}/\log(1+\sigma^{-2})$
- Т.к. $\forall s \in [0, \sigma^{-2}]$ и $\sigma^{-2}\sigma_{t-1}^2(\mathbf{x}_t) \leq \sigma^{-2} k(\mathbf{x}_t, \mathbf{x}_t) \leq \sigma_{-2}$ подставляя неравенства в 3 лемму получаем удтверждение леммы.

Далее применяем неравенство Коши-Буняковского $R_T^2 \leq T \sum_{t=1}^{T} r_t^2$ и доказываем теорему.

Для непрерывного D

Th.

- $D \subset [0, r]^d$ - компактное, выпуклое множество, $d \in \mathbb{N},\ r > 0$
- Сэмплируемые функции f из GP с функцией ковариации $k(\mathbf{x}, \mathbf{x}^*)$ с высокой вероятностью имеют ограниченные производные
$$\mathbb{P}\{\sup_{\mathbf{x}\in D} |\partial f/\partial x_j| > L\} \leq a e^{-(L/b)^2}$$
для каких-то констант a и b.
- Выбираем $\delta \in (0, 1)$, тогда
$$\beta_t = 2\log(t^2 2\pi^2/(3\delta)) + 2d\log(t^2 dbr\sqrt{\log(4da/\delta)})$$
Cumulative regret алгоритма GP-UCB с параметром уверенности β_t, медианой 0, функцией ковариации $k(\mathbf{x}, \mathbf{x}^*)$ ограничен $\mathbb{O}^*(\sqrt{dT\gamma_T})$ с высокой вероятностью.
$$\mathbb{P}\{R_T \leq \sqrt{C_1 T\beta_T\gamma_T} + 2\ \forall T \geq 1\} \geq 1 - \delta$$
где $C_1 = 8/\log(1 + \sigma^{-2})$

Оценка максимального information gain’a

γ_T - оценивается точно жадным алгоритмом для конкретной задачи

аналитические выражения представлены в статье

Материалы

1. Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design. N. Srinivas https://arxiv.org/pdf/0912.3995.pdf
2. Another introduction to Gaussian Processes. Richard Wilkinson http://gpss.cc/gpss17/slides/Wilkinson2017.pdf
3. Gaussian proccesses tutorial. Martiт Krasser https://krasserm.github.io/2018/03/19/gaussian-processes/
4. Introduction to bayesian optimization. Javier Gonzalez http://gpss.cc/gpss17/slides/gpss_bayesopt2017.pdf
5. Bayesian optimization tutorial. Martin Krasser https://krasserm.github.io/2018/03/21/bayesian-optimization/
6. Multivariate gaussian distributions. Peter Roelants. https://peterroelants.github.io/posts/multivariate-normal-primer/
7. Gaussian processes lecture. UBC https://www.cs.ubc.ca/~nando/540-2013/lectures/l6.pdf
8. Теория вероятностей. Натан, Горбачёв, Гуз http://www.combook.ru/product/10006534/
9. Learning with spectral kernels. Markus Heinonen https://www.hiit.fi/wp-content/uploads/2018/04/Spectral-Kernels-S12.pdf

спасибо