Лекция 7. Гауссовские процессы

Петр Мостовский

МКН СПбГУ

31 марта 2022

Факультет
математики
и компьютерных
наук
СПбГУ

Пятиминутка

- Выпишите разложение параметров тематической модели через матрицы вероятностей терминов в теме и тем в документе
- Опишите, что такое распределение Дирихле (Latent Dirichlet Allocation, LDA), можно без формул
- Приведите ключевые идеи модели word2vec

Напоминание о теореме Байеса

$$P(A|B) = \frac{P(B|A)P(A)}{\int P(B|A)P(A)dA} = \frac{P(B|A)P(A)}{P(B)} = \frac{P(A,B)}{P(B)}$$

Гауссовские случайные вектора

Случайный вектор $\mathbf{X}$ *нормально распределен*, если любая линейная комбинация его компонент нормально распределена.

Обозначение:

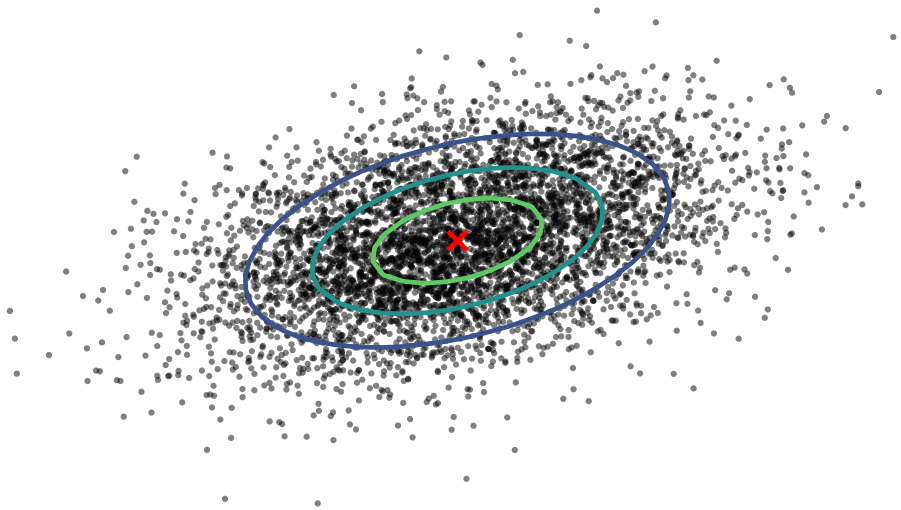
$$\mathbf{X} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$$

$$\mathbb{E}(\mathbf{X}[k]) = \mu[k]$$

$$\mathbb{E}((\mathbf{X}[i] - \mu[i])(\mathbf{X}[j] - \mu[j])) = \mathbf{\Sigma}[i,j] = \mathrm{Cov}(\mathbf{X}[i], \mathbf{X}[j])$$

Гауссовские случайные вектора

Случайный вектор $\mathbf{X}$ *нормально распределен*, если любая линейная комбинация его компонент нормально распределена.

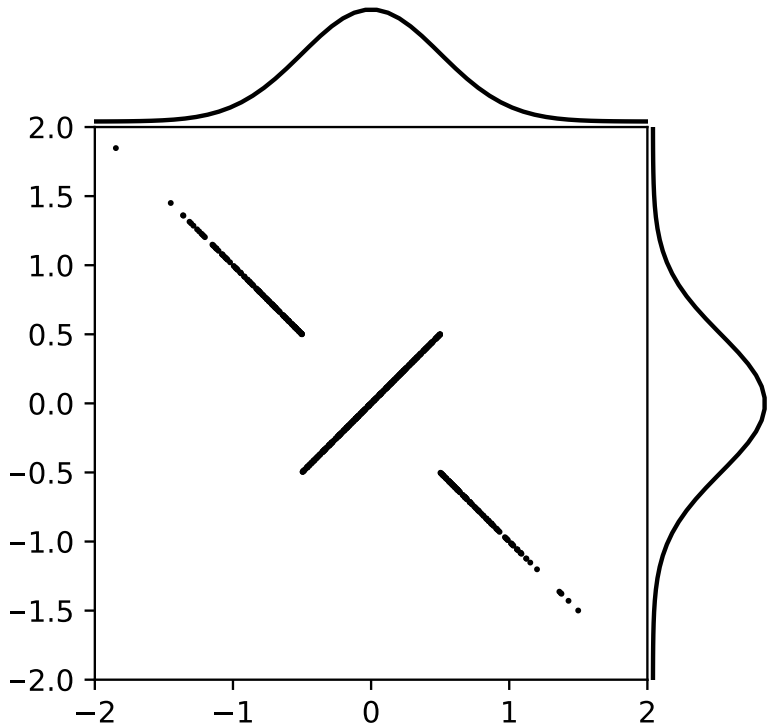


Совместно нормальные величины

Вектора $\mathbf{X}$ и $\mathbf{Y}$ *совместно* нормально распределены, если вектор $[\mathbf{X}, \mathbf{Y}]$ нормально распределен.

Совместно нормальные величины

Вектора $\mathbf{X}$ и $\mathbf{Y}$ *совместно* нормально распределены, если вектор $[\mathbf{X}, \mathbf{Y}]$ нормально распределен.
Pro-tip: не всякая пара нормальных векторов $\mathbf{X}$ и $\mathbf{Y}$ совместно нормальна.



Некоторые полезные свойства

«Подвекторы» нормального вектора $X \sim \mathcal{N}(\mu, \Sigma)$ тоже, конечно, нормальны.
Если $X = [X_1, X_2]$ и

$$\mu = [\mu_1, \mu_2]$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

то

$$X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}), \quad X_2 \sim \mathcal{N}(\mu_2, \Sigma_{22})$$

Некоторые полезные свойства

«Подвекторы» нормального вектора $X \sim \mathcal{N}(\mu, \Sigma)$ тоже, конечно, нормальны.
Если $X = [X_1, X_2]$ и

$$\mu = [\mu_1, \mu_2]$$

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix},$$

то

$$X_1 \sim \mathcal{N}(\mu_1, \Sigma_{11}), \quad X_2 \sim \mathcal{N}(\mu_2, \Sigma_{22})$$

В дальнейшем мы будем работать векторами (и процессами) с нулевым матожиданием, поскольку

$$X \sim (\mu, \Sigma) \Leftrightarrow X = \mu + Y, \quad Y \sim (0, \Sigma)$$

Обуславливание

Пусть X, Y – совместно нормальные векторы:

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

Обуславливание

Пусть X, Y – совместно нормальные векторы:

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix} \right)$$

Тогда *условный* вектор $X|(Y = y)$ тоже нормален:

$$X|(Y = y) \sim \mathcal{N}\left(\Sigma_{12}\Sigma_{22}^{-1}y, \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$$

Обуславливание с шумом

Более естественная постановка: пусть X, Y — совместно нормальны. Пусть $Y = y + \varepsilon$, где $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ – гауссовский шум.

Обуславливание с шумом

Более естественная постановка: пусть X, Y — совместно нормальны. Пусть $Y = y + \varepsilon$, где $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ – гауссовский шум.
Каково распределение $X|(Y = y + \varepsilon)$?

$$\begin{bmatrix} X \\ Y + \varepsilon \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} + \sigma^2 I \end{bmatrix} \right)$$

Обуславливание с шумом

Более естественная постановка: пусть X, Y — совместно нормальны. Пусть $Y = y + \varepsilon$, где $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ – гауссовский шум.
Каково распределение $X|(Y = y + \varepsilon)$?

$$\begin{bmatrix} X \\ Y + \varepsilon \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} + \sigma^2 I \end{bmatrix} \right)$$

$$p(X|Y + \varepsilon) = \frac{p(Y + \varepsilon, X)}{\int p(Y + \varepsilon, X) dX} =$$

$$\mathcal{N}\left(\Sigma_{12}(\Sigma_{22} + \sigma^2 I)^{-1} Y,\ \Sigma_{11} - \Sigma_{12}(\Sigma_{22} + \sigma^2 I)^{-1}\Sigma_{21}\right)$$

(считается аналитически)

Гауссовские процессы

- Нормальные случайные вектора – это обобщение нормальных случайных величин. А что если мы хотим говорить о бесконечномерных обобщениях – о случайных функциях?

Гауссовские процессы

- Нормальные случайные вектора – это обобщение нормальных случайных величин. А что если мы хотим говорить о бесконечномерных обобщениях – о случайных функциях?
- *Гауссовский процесс* на пространстве $\mathcal{X}$ – это набор нормальных случайных величин $\{f(x)\}_{x \in \mathcal{X}}$, таких что *для любых* $x_1, \ldots, x_N$ конечномерный вектор $[f(x_1), \ldots, f(x_N)]$ нормально распределен.

Гауссовские процессы

- Нормальные случайные вектора – это обобщение нормальных случайных величин. А что если мы хотим говорить о бесконечномерных обобщениях – о случайных функциях?
- *Гауссовский процесс* на пространстве $\mathcal{X}$ – это набор нормальных случайных величин $\{f(x)\}_{x \in \mathcal{X}}$, таких что *для любых* $x_1, \ldots, x_N$ конечномерный вектор $[f(x_1), \ldots, f(x_N)]$ нормально распределен.
- Обозначение $f(x)$ выбрано не случайно – можно думать о гауссовском процессе как о *распределении* на функциях в пространстве $\mathcal{X}$.

Гауссовские процессы

Гауссовский процесс однозначно задается своим матожиданием и функцией ковариации (также называемой *ядром*).

$$
\begin{aligned}
f(x) &\sim \mathcal{GP}(m(x), k(x, x')) \\
\mathbb{E}(f(x)) &= m(x) && \forall x \in \mathcal{X} \\
\mathrm{Cov}(f(x), f(x')) &= k(x, x') && \forall x, x' \in \mathcal{X}
\end{aligned}
$$

Гауссовские процессы

Гауссовский процесс однозначно задается своим матожиданием и функцией ковариации (также называемой *ядром*).

$$
\begin{aligned}
f(x) &\sim \mathcal{GP}(m(x), k(x, x')) \\
\mathbb{E}(f(x)) &= m(x) && \forall x \in \mathcal{X} \\
\mathrm{Cov}(f(x), f(x')) &= k(x, x') && \forall x, x' \in \mathcal{X}
\end{aligned}
$$

Для любого $X := [x_1, \ldots, x_N]$ вектор $f_X := [f(x_1), \ldots, f(x_N)]$ гауссовский:

$$f_X \sim \mathcal{N}(m_X, K_{XX}),$$

где

$$m_X := [m(x_1), \ldots, m(x_N)], \quad K_{XX}[i, j] := k(x_i, x_j)$$

Причем здесь машинное обучение?

- Байесовский подход заключается в введении предположений о природе данных и обновлении этих предположений с учетом собственно данных с помощью теоремы Байеса. Мы видели, как такой подход можно применить в параметрических моделях — предположения строятся относительно параметров модели.

Причем здесь машинное обучение?

- Байесовский подход заключается в введении предположений о природе данных и обновлении этих предположений с учетом собственно данных с помощью теоремы Байеса. Мы видели, как такой подход можно применить в параметрических моделях — предположения строятся относительно параметров модели.
- Гауссовские же процессы позволяют мыслить о данных *непараметрически*. А именно, задавать априорное распределение *функций*, которые порождают наблюдаемые данные, и обновлять это распределение с помощью теоремы Байеса. Это приводит к *Gaussian Process Regression* – регрессии, основанной на гауссовских процессах.

Gaussian Process Regression

- Пусть есть некий датасет (X, Y), где $X = [x_1, \ldots, x_N]$, $x_i \in \mathbb{R}^d$, а $Y = [y_1, \ldots, y_N]$, $y_i \in \mathbb{R}$

Gaussian Process Regression

- Пусть есть некий датасет (X, Y), где $X = [x_1, \ldots, x_N]$, $x_i \in \mathbb{R}^d$, а $Y = [y_1, \ldots, y_N]$, $y_i \in \mathbb{R}$
- Данные связаны некоей функциональной зависимостью $y_i = f(x_i)$

Gaussian Process Regression

- Пусть есть некий датасет (X, Y), где $X = [x_1, \ldots, x_N]$, $x_i \in \mathbb{R}^d$, а $Y = [y_1, \ldots, y_N]$, $y_i \in \mathbb{R}$
- Данные связаны некоей функциональной зависимостью $y_i = f(x_i)$
- Функцию f мы не знаем и хотим оценить из данных. Иными словами, каково значение $f(x_*)$ в какой-то произвольной точке x_*

Gaussian Process Regression

Введем априорное предположение, что $f \sim \mathcal{GP}(0, k)$ – некий гауссовский процесс, и предположим, что мы знаем ковариационную функцию k (позже мы увидим, как оценить k из данных).

Gaussian Process Regression

Вектор $f(X) = [f(x_1), \ldots, f(x_N)]$ будет нормальным $f(X) \sim \mathcal{N}(0, K_{XX})$ (по определению гауссовского процесса).
При этом для произвольного x_* вектор $[f(x_*), f(X)] = [f(x_*), f(x_1), \ldots, f(x_N)]$ тоже будет нормальным (по тому же определению):

$$\begin{bmatrix} f(x_*) \\ f(X) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K_{**} & K_{*X} \\ K_{X*} & K_{XX} \end{bmatrix}\right)$$

Gaussian Process Regression

Вектор $f(X) = [f(x_1), \ldots, f(x_N)]$ будет нормальным $f(X) \sim \mathcal{N}(0, K_{XX})$ (по определению гауссовского процесса). При этом для произвольного x_* вектор $[f(x_*), f(X)] = [f(x_*), f(x_1), \ldots, f(x_N)]$ тоже будет нормальным (по тому же определению):

$$\begin{bmatrix} f(x_*) \\ f(X) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K_{**} & K_{*X} \\ K_{X*} & K_{XX} \end{bmatrix}\right)$$

Кроме того, мы знаем, что $f(X) = Y$. Мы можем посчитать условное распределение:

$$f(x_*)|(f(X) = Y) \sim \mathcal{N}\left(K_{*X}K_{XX}^{-1}Y, K_{**} - K_{*X}K_{XX}^{-1}K_{X*}\right)$$

Gaussian Process Regression

Вектор $f(X) = [f(x_1), \ldots, f(x_N)]$ будет нормальным $f(X) \sim \mathcal{N}(0, K_{XX})$ (по определению гауссовского процесса). При этом для произвольного x_* вектор $[f(x_*), f(X)] = [f(x_*), f(x_1), \ldots, f(x_N)]$ тоже будет нормальным (по тому же определению):

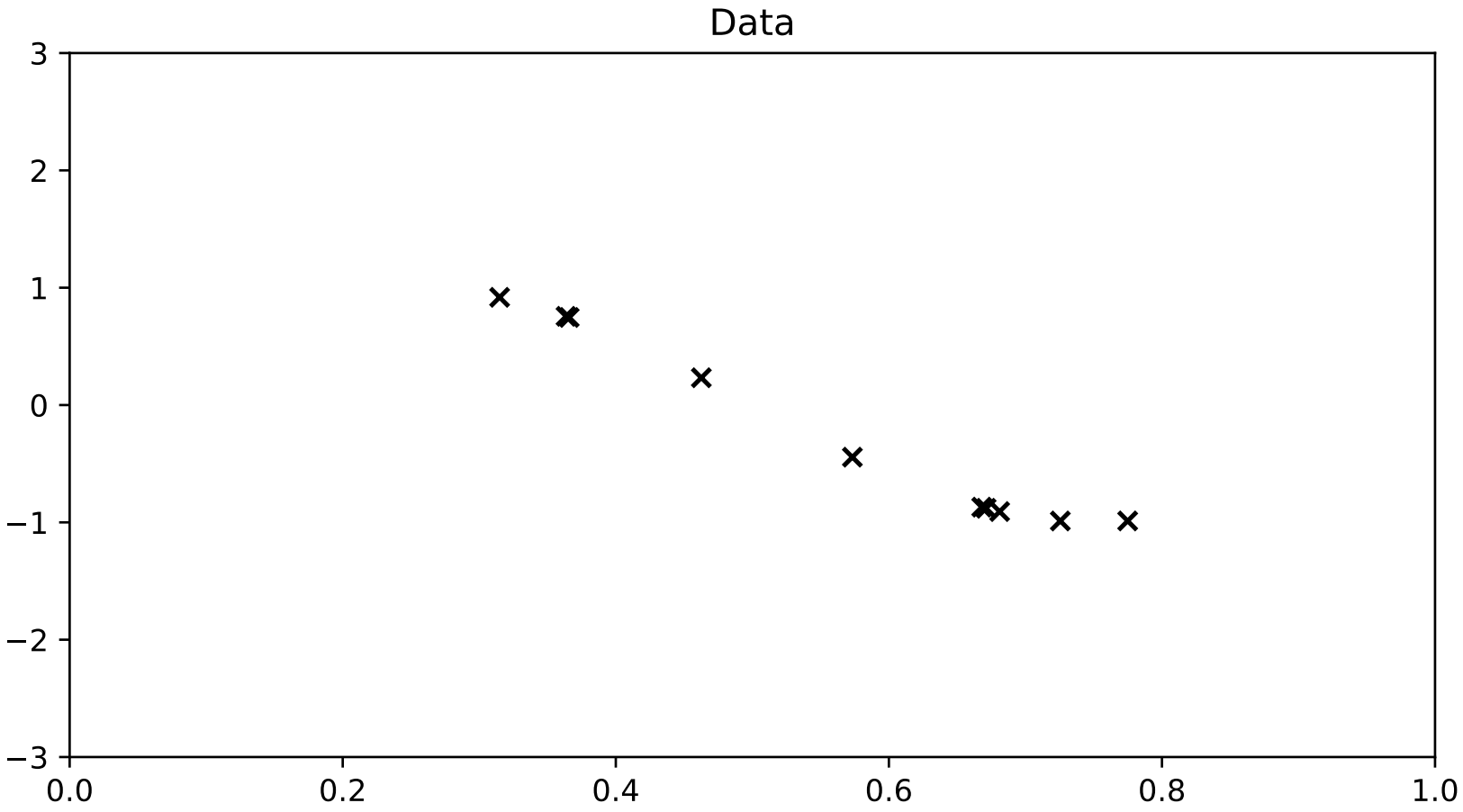
$$\begin{bmatrix} f(x_*) \\ f(X) \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K_{**} & K_{*X} \\ K_{X*} & K_{XX} \end{bmatrix}\right)$$

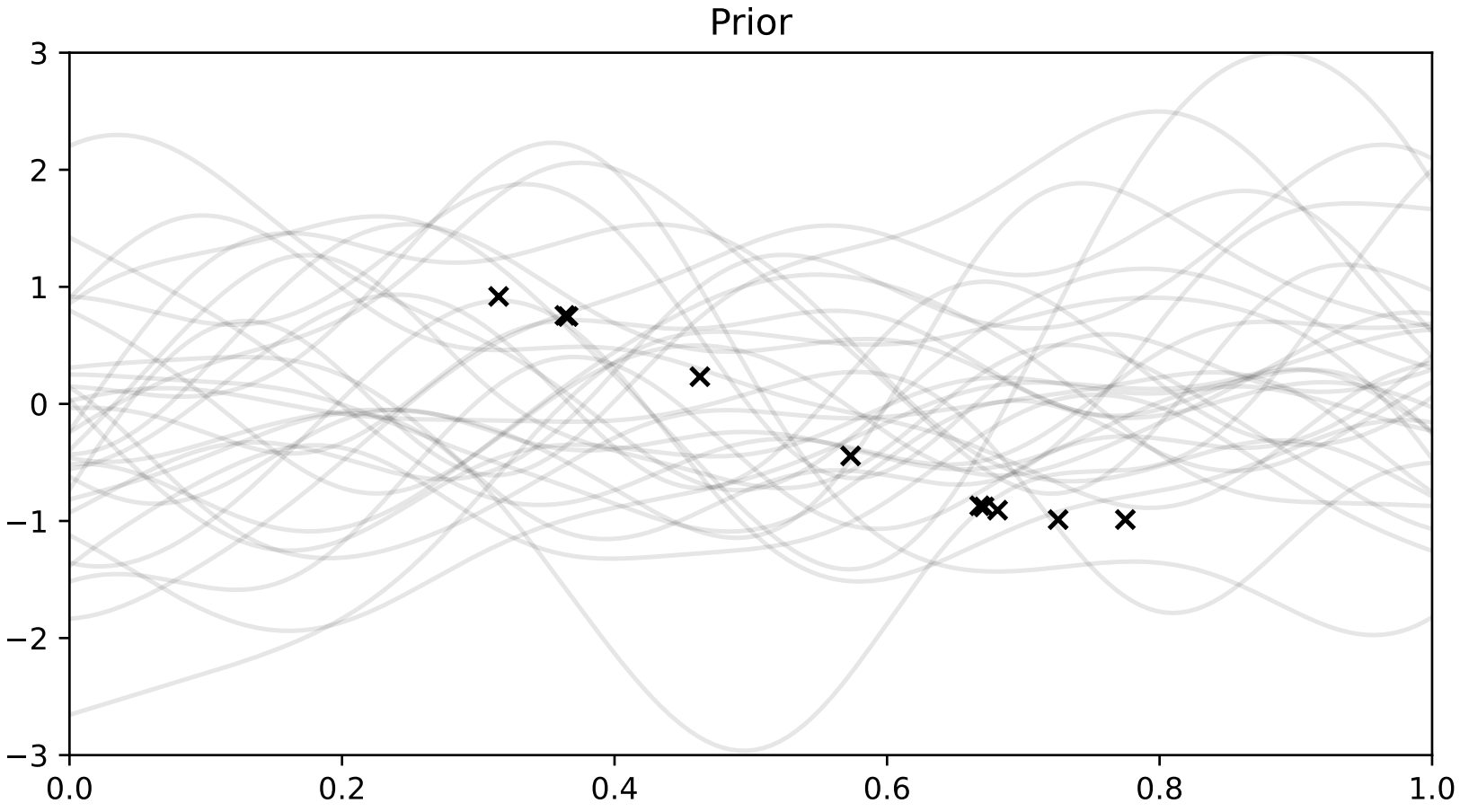
Кроме того, мы знаем, что $f(X) = Y$. Мы можем посчитать условное распределение:

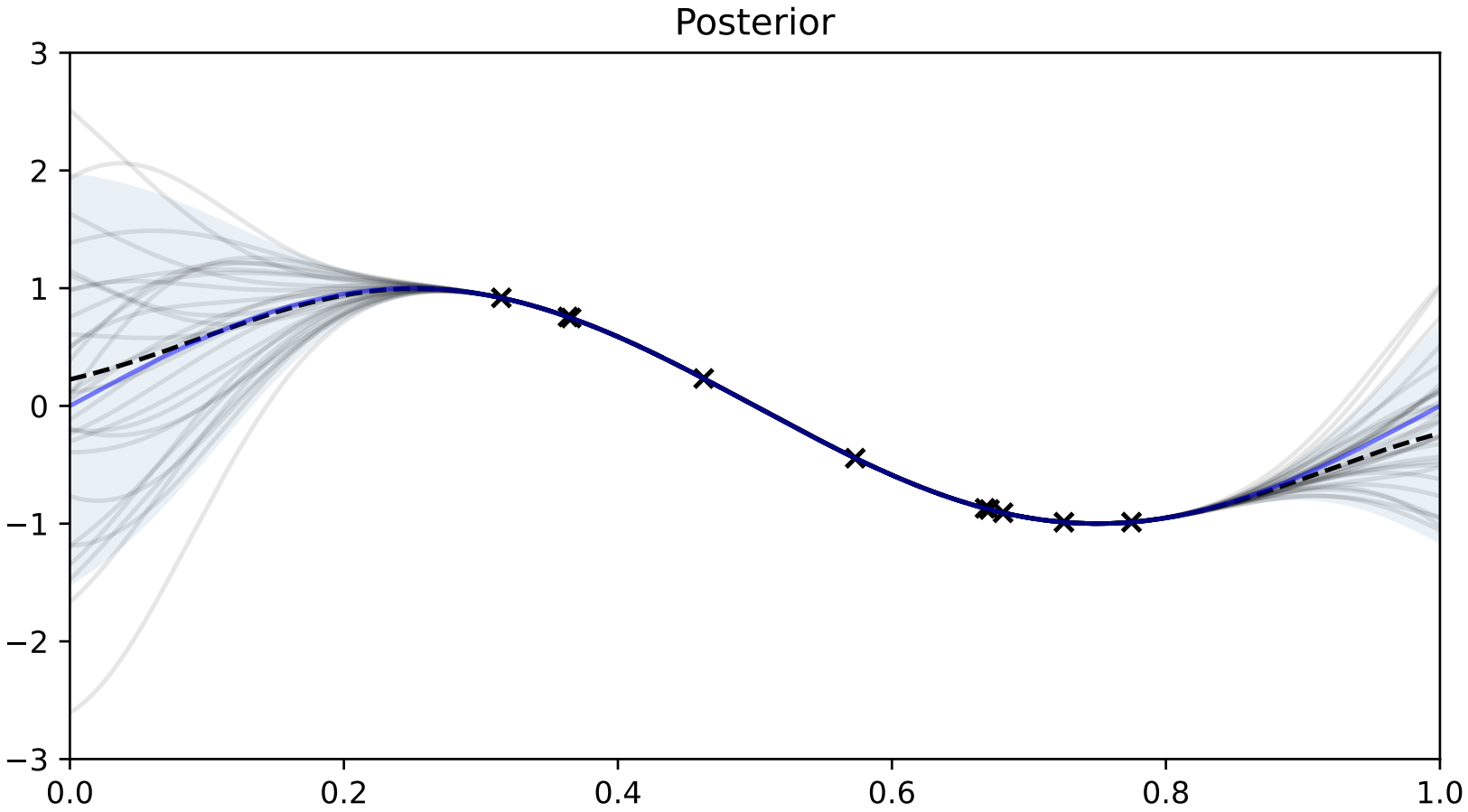
$$f(x_*)|(f(X) = Y) \sim \mathcal{N}\left(K_{*X}K_{XX}^{-1}Y, K_{**} - K_{*X}K_{XX}^{-1}K_{X*}\right)$$

Поскольку эти рассуждения верны для *любого* набора точек $X_* = [x_{*1}, \ldots, x_{*T}]$, мы приходим к *условному* (или *апостериорному*) процессу

$$f(x)|(f(X) = Y) \sim \mathcal{GP}\left(k(x, X)K_{XX}^{-1}Y, k(x, X)K_{XX}^{-1}k(X, x)\right)$$


Data
3
2
1
0
−1
−2
−3
0.0
0.2
0.4
0.6
0.8
1.0


Prior
3
2
1
0
−1
−2
−3
0.0
0.2
0.4
0.6
0.8
1.0


Posterior
3
2
1
0
−1
−2
−3
0.0
0.2
0.4
0.6
0.8
1.0

Апостериорный гауссовский процесс

- Апостериорный гауссовский процесс предоставляет *распределение* функций, согласованных с данными.

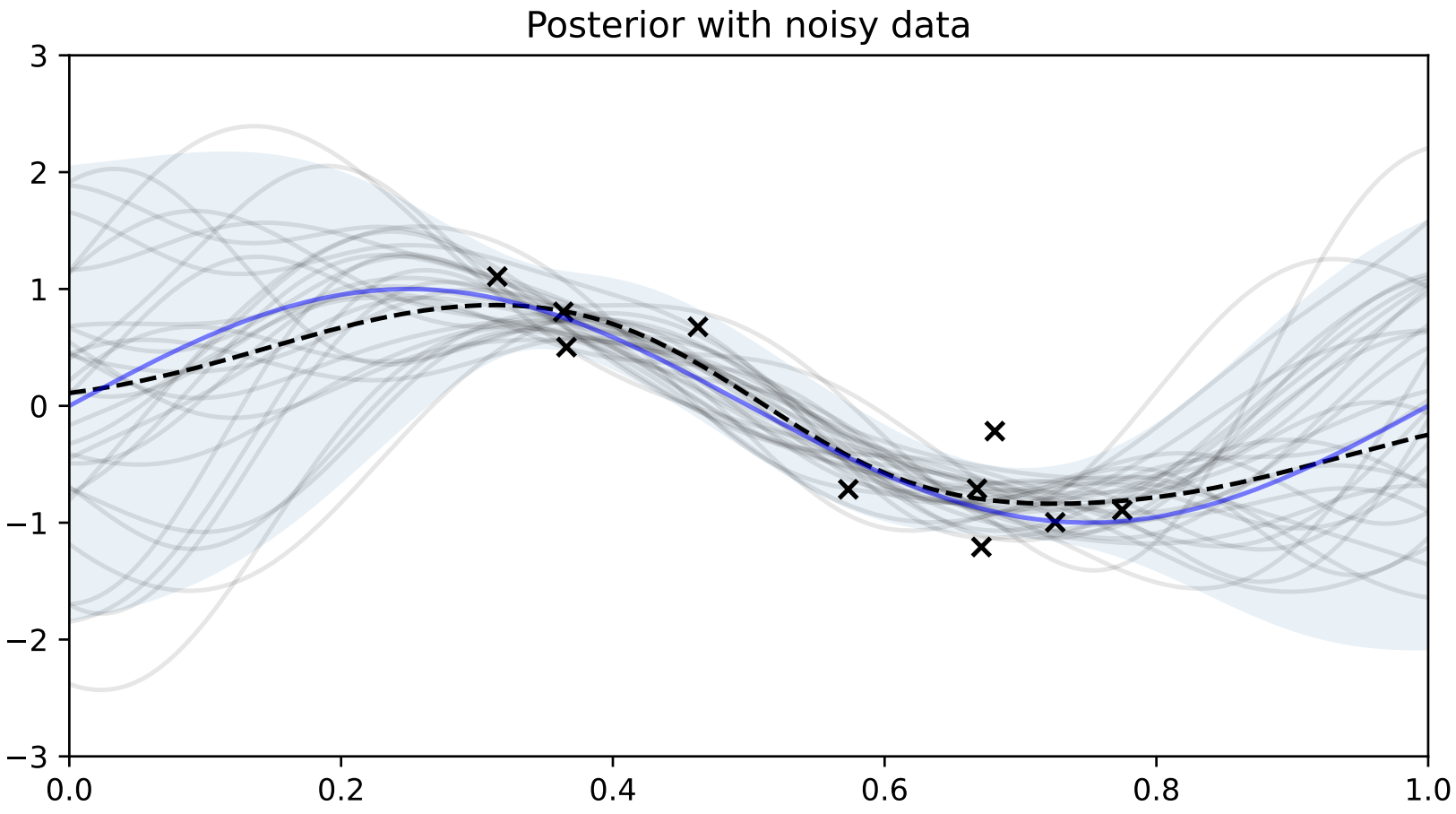
Апостериорный гауссовский процесс

- Апостериорный гауссовский процесс предоставляет *распределение* функций, согласованных с данными.
- Дисперсия служит численной мерой неопределенности модели – чем выше дисперсия, тем более неуверенна модель в своем ответе.

Шумные данные

Кроме того, обычно предполагают зашумленность данных: $f(X) = Y + \varepsilon$, где $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ – гауссовский шум. В таком случае

$$f(x)|(f(X) = Y + \varepsilon) \sim \mathcal{GP}\Big(k(x, X)(K_{XX} + \sigma^2 I)^{-1} Y,$$
$$k(x, X)(K_{XX} + \sigma^2 I)^{-1} k(X, x)\Big)$$


Posterior with noisy data
3
2
1
0
−1
−2
−3
0.0
0.2
0.4
0.6
0.8
1.0

Где Байес?

Введем обозначения: $F := f(X)$

априорное распределение (prior):

$$p(F) = \mathcal{N}\Big(0, K_{XX}\Big)$$

Где Байес?

Введем обозначения: $F := f(X)$

правдоподобие (likelihood):

$$p(Y|F) = \mathcal{N}\left(F, \sigma^2 I\right)$$

Где Байес?

Введем обозначения: $F := f(X)$

апостериорное распределение (posterior):

$$p(F|Y) = \frac{p(Y|F)p(F)}{\int p(Y|F)p(F)}$$

Где Байес?

Введем обозначения: $F := f(X)$

предсказательное распредение (predictive distribution):

$$p(f(x_*)|Y) = \int p(f(x_*)|F)p(F|Y)dF$$

Где Байес?

Введем обозначения: $F := f(X)$

Поскольку все гауссовское, все считается аналитически и:

$$p(f(x_*)|Y) = \mathcal{N}\Big(k(x_*, X)(K_{XX} + \sigma^2 I)^{-1} Y,$$
$$k(x_*, X)(K_{XX} + \sigma^2 I)^{-1} k(X, x_*)\Big)$$

«Обучение» гауссовского процесса

Для проведения регрессии на основе гауссовских процессов необходимо задать априорный гауссовский процесс, то есть задать его матожидание и функцию ковариации. Как правило, матожидание полагают равным нулю – всегда можно отнормировать данные. Свойства гауссовского процесса (например, гладкость траекторий) кодируются ядром. Выбор же ядра – сложный процесс и во многом искусство (хотя и существуют методы автоматического подбора ядра).

Максимизация правдоподобия

Как правило, предполагают, что ядро k принадлежит некоторому параметрическому семейству $k = k_\theta, \theta \in \Theta$.

Максимизация правдоподобия

Как правило, предполагают, что ядро k принадлежит некоторому параметрическому семейству $k = k_\theta, \theta \in \Theta$. Для подбора оптимальных параметров ядра используется метод максимизации правдоподобия.

$$L(Y;\theta) = \log p(Y;\theta) = \log \int p(Y|F;\theta)p(F;\theta)dF \longrightarrow \max$$

Максимизация правдоподобия

Как правило, предполагают, что ядро k принадлежит некоторому параметрическому семейству $k = k_\theta, \theta \in \Theta$. Для подбора оптимальных параметров ядра используется метод максимизации правдоподобия.

$$L(Y;\theta) = \log p(Y;\theta) = \log \int p(Y|F;\theta)p(F;\theta)dF \longrightarrow \max$$

А именно,

$$L(Y;\theta) = -\frac{1}{2}Y^\top K_y^{-1} Y - \frac{1}{2}\log\det K_Y - \frac{n}{2}\log 2\pi \longrightarrow \max$$

где $K_y = K_{XX} + \sigma^2 I$, а n – количество данных.

Ядра Матерна

Наиболее популярное семейство ядер – ядра Матерна.

$$k(x, x') = k(||x - x'||) = k_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{d}{\rho} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{d}{\rho} \right),$$

где Γ – гамма-функция, K_ν – модифицированная функция Бесселя второго рода.
Параметры ядра:

- σ^2 задает дисперсию гауссовского процесса

Ядра Матерна

Наиболее популярное семейство ядер – ядра Матерна.

$$k(x, x') = k(||x - x'||) = k_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{d}{\rho} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{d}{\rho} \right),$$

где Γ – гамма-функция, K_ν – модифицированная функция Бесселя второго рода.
Параметры ядра:

- σ^2 задает дисперсию гауссовского процесса
- ν отвечает за гладкость траекторий гауссовского процесса

Ядра Матерна

Наиболее популярное семейство ядер – ядра Матерна.

$$k(x, x') = k(||x - x'||) = k_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{d}{\rho} \right)^\nu K_\nu \left(\sqrt{2\nu} \frac{d}{\rho} \right),$$

где Γ – гамма-функция, K_ν – модифицированная функция Бесселя второго рода.
Параметры ядра:

- σ^2 задает дисперсию гауссовского процесса
- ν отвечает за гладкость траекторий гауссовского процесса
- ρ (параметр масштаба, lengthscale) отвечает за растяжение пространства

Ядра Матерна

При $\nu = k + 1/2$, $k = 0, 1, 2, \ldots$ формулы упрощаются и наиболее часто используемые ядра это Матерн-1/2, Матерн-3/2, Матерн-5/2.
При $\nu \to \infty$ ядро Матерна совпадает с другим известным ядром – гауссовским:

$$k(x, x') = \sigma^2 \exp\left(-\frac{||x - x'||^2}{2\rho^2}\right)$$

Другие ядра

Рассматривают и другие семейства ядер, такие как Rational Quadratic, Piecewise Polynomial, Standard Periodic, и т.д.
В целом любая положительно-определенная функция может быть ядром.

Другие ядра

Рассматривают и другие семейства ядер, такие как Rational Quadratic, Piecewise Polynomial, Standard Periodic, и т.д.
В целом любая положительно-определенная функция может быть ядром.
Кроме того, если $k_1(x, x'), k_2(x, x')$ – ядра, то
$k_1(x, x') + k_2(x, x')$, $k_1(x, x')k_2(x, x')$ – ядра.
Если $k_1(x, x')$, $k_2(y, y')$ – ядра, то $k(z, z') = k_1(x, x') + k(y, y')$,
$k(z, z') = k(x, x')k(y, y')$ – ядра, где $z = [x, y]$.

Формула Матерона

Апостериорный процесс $f|Y$ задается аналитическими матожиданием и ковариационной функцией.

$$f(X_*) \sim \mathcal{N}\Big(\mu := k(x_*, X)(K_{XX} + \sigma^2 I)^{-1} Y,$$
$$\Sigma := k(x_*, X)(K_{XX} + \sigma^2 I)^{-1} k(X, x_*)\Big)$$

Формула Матерона

Апостериорный процесс $f|Y$ задается аналитическими матожиданием и ковариационной функцией.

$$f(X_*) \sim \mathcal{N}\Big(\mu := k(x_*, X)(K_{XX} + \sigma^2 I)^{-1} Y,$$
$$\Sigma := k(x_*, X)(K_{XX} + \sigma^2 I)^{-1} k(X, x_*)\Big)$$

$$f(X_*) = \mu + \Sigma^{1/2} u, \quad u \sim \mathcal{N}(0, I)$$

Это требует вычисления корня из матрицы Σ размера (T, T) (сложность $O(T^3)$ операций). Если T велико, то это фактически невозможно.

Формула Матерона

Альтернативный способ предоставляет формула Матерона.

$$f(x_*)|Y = f(x_*) + (K_{XX} + \sigma^2 I)^{-1}(Y - \varepsilon)$$

Пусть мы умеем каким-то образом эффективно получать траектории априорного процесса f как функции от x. Тогда можно автоматически получать траектории процесса $f|Y$ как функции от x_*, *корректируя* априорные сэмплы поправкой, вносимой наблюдаемыми данными.

Зачем

- Байесовский подход, в котором всё можно посчитать

Зачем

- Байесовский подход, в котором всё можно посчитать (почти).

Зачем

- Байесовский подход, в котором всё можно посчитать (почти).
- Оценка неопределенности.

Зачем

- Байесовский подход, в котором всё можно посчитать (почти).
- Оценка неопределенности.
- Small data.

Зачем

- Байесовский подход, в котором всё можно посчитать (почти).
- Оценка неопределенности.
- Small data.
- Ядра существуют для необычных пространств — графы, римановы многообразия и т.п.

Проблемы ГП-регрессии

- **Big data**. Поскольку «обучение» ГП требует обращения матрицы K_{XX}, размер которой зависит от количества данных, использование ГП-регрессии на больших данных затруднено. Обращение большой плотной плохо обусловленной матрицы на каждой итерации градиентного спуска – вычислительно тяжелая задача. Существует методы *sparse Gaussian processes*, которые в определенной степени решают эту проблему.

Проблемы ГП-регрессии

- **Big data**. Поскольку «обучение» ГП требует обращения матрицы K_{XX}, размер которой зависит от количества данных, использование ГП-регрессии на больших данных затруднено. Обращение большой плотной плохо обусловленной матрицы на каждой итерации градиентного спуска – вычислительно тяжелая задача. Существует методы *sparse Gaussian processes*, которые в определенной степени решают эту проблему.
- **Large dimensions**. ГП-регрессия плохо работает на данных в пространствах большой размерности (например, картинки). Существуют подходы для работы с такими данными.

Это еще не всё

GP in Machine Learning — активно развивающаяся область науки.
Стандартная книжка по гауссовским процессам в машинном обучении – Gaussian Processes for Machine Learning (доступна онлайн).

Спасибо!