Эффективное применение гауссовских процессов к задаче классификации

Вайсер Кирилл Олегович

Московский физико-технический институт

Группа 774, осень 2020

Научный руководитель: Панов М.Е.

Задача эффективного применения гауссовских процессов

Цель

Предложить модель активного обучения в решении задач классификации, использующую гауссовский процесс в качестве латентной функции.

Решаемая проблема

Для того, чтобы учесть информацию, содержащуюся в обучающей выборке, необходимо вычислить апостериорное распределение гауссовского процесса. Однако вычисление параметров апостериорного распределения и сэмплирование из него - дорогие по числу вычислений операции.

Постановка задачи : построение модели

1. выборка

$$\mathfrak{D} = \{\boldsymbol{x}_i, y_i\} \quad i = 1, \ldots, n, \quad \boldsymbol{x}_i \in \mathbb{R}^d \quad y_i \in \{-1, 1\}$$

2. оракул $g : \mathbb{R}^d \mapsto \{-1, 1\}$ —функция, которая возвращает правильные ответы, но обращение к которой дорогостоящее в некотором смысле, например долго вычисляется или требует много ресурсов. Например командировка бригады бурильщиков для выявления наличия нефти в определенной местности.

3. модель, которая аппроксимирует вероятность объекта $\boldsymbol{x}$ принадлежать к классу $g(\boldsymbol{x})$:

$$\boldsymbol{f}(\boldsymbol{x}, \boldsymbol{\theta}) : \mathbb{R}^d \times \mathbb{R}^h \mapsto [0, 1],$$

где $\boldsymbol{\theta} \in \mathbb{R}^h$ — параметры модели.

Постановка задачи : метрики неопределенности

Для определения точек, в которых необходимо узнать истинную метку для улучшения модели, можно использовать значения метрик неопределенности.

1. $\mathbb{E}H(p(x)) = \mathbb{E}\sum_{c\in C} p_c(x)\log p_c(x)$,
   где $p_c(x)$ это вероятность принадлежности объекта x к классу c. Эта метрика соответствует средней энтропии в точке.
2. $\mathbb{E}H(p(x)) - H(\mathbb{E}p(x))$
   Эпистемическая неопределенность в точке (epistemic uncertainty).
3. $\max_c p_c(x)$
   Максимальная вероятность класса.

Гауссовский процесс

Гауссовский процесс —случайный процесс, любая конечная совокупность сечений которого имеет совместное нормальное распределение.
Гауссовский процесс характеризуется функцией среднего и ковариационной функцией:

$$m(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})]$$
$$k\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \mathbb{E}\left[\left(f(\boldsymbol{x}) - m(\boldsymbol{x})\right)\left(f\left(\boldsymbol{x}'\right) - m\left(\boldsymbol{x}'\right)\right)\right].$$

Гауссовский процесс задает тем самым распределение функций на своей области определения. Это свойство существенно увеличивает обобщающую способность модели, построенной на гауссовском процессе, а так же делает его хорошим выбором в задачах, требующих расчета метрик неопределенности.

Традиционный подход

Обозначим $\boldsymbol{f}_m \mid \boldsymbol{y}$ за апостериорное распределение гауссовского процесса в точках $\{\boldsymbol{x'}\}_{i=1}^m = X_m$ после наблюдения выборки $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$. Тогда его среднее и матрица ковариации записываются как

$$\boldsymbol{m}_{m|n} = \boldsymbol{K}_{m,n}\left(\boldsymbol{K}_{n,n} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{y}$$
$$\boldsymbol{K}_{m,m|n} = \boldsymbol{K}_{m,m} - \boldsymbol{K}_{m,n}\left(\boldsymbol{K}_{n,n} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{K}_{n,m},$$

а сэмплирование производится с помощью стандартной схемы

$$\boldsymbol{f}_m \mid \boldsymbol{y} = \boldsymbol{m}_{m|n} + \boldsymbol{K}_{m,m|n}^{1/2}\zeta, \quad \zeta \sim \mathbb{N}(0,1).$$

Такой подход требует произвести $O(n^3) + O(m^3)$ вычислений, что существенно снижает его работоспособность для больших обучающих и тестовых выборок.

Взвешенный подход

Вместо прямого точного вычисления моментов апостериорного распределения и сэмплирования из него факторизуем его на два слагаемых.
Взвешенный подход проистекает из байесовской линейной модели

$$f = \phi(\boldsymbol{x})^\top \boldsymbol{w},$$

где $\phi : \mathbb{R}^d \mapsto \mathbb{R}^k$, а $\boldsymbol{w}$ — параметры модели с априорным распределением $\mathbb{N}(0, \Sigma_p)$. Тогда

$$f_m \mid \boldsymbol{x}_m, X, \boldsymbol{y} \sim \mathcal{N}(\phi_m^\top \Sigma_p \Phi \left(K + \sigma_n^2 I\right)^{-1} \boldsymbol{y}$$
$$\phi_m^\top \Sigma_p \phi_m - \phi_m^\top \Sigma_p \Phi \left(K + \sigma_n^2 I\right)^{-1} \Phi^\top \Sigma_p \phi_m\Big),$$

где $K = \Phi^\top \Sigma_p \Phi$. Тогда мы можем определить ковариационную функцию как

$$k(x, x') = \phi(x)^\top \Sigma_p \phi(x') = \psi(x)^\top \psi(x'),$$

где $\psi(x) = \Sigma_p^{1/2} \phi(x)$.

Функциональный подход

Гауссовский процесс может быть рассмотрен как распределение над функциями на заданной области определения. Если эта область очень большая, то можно рассмотреть подмножество *индуцирующих* точек $\boldsymbol{Z} = \{z_j\}_{j=1}1^v$. Пусть $u = f(\boldsymbol{Z}) \sim \mathbb{N}(\mu_u, \Sigma_u)$ Тогда $f_m \mid \boldsymbol{u}$ имеет следующие моменты

$$\boldsymbol{\mu}_{m|v} = \boldsymbol{K}_{m,v}\boldsymbol{K}_{v,v}^{-1}\boldsymbol{\mu}_v$$

$$\boldsymbol{K}_{m,m|v} = \boldsymbol{K}_{m,m} + \boldsymbol{K}_{m,v}\boldsymbol{K}_{v,v}^{-1}\left(\Sigma_{\boldsymbol{u}} - \boldsymbol{K}_{v,v}\right)\boldsymbol{K}_{v,v}^{-1}\boldsymbol{K}_{v,m}$$

Такой подход позволяет снизить сложность обучения с $O(n^3)$ до $O(v^3)$, что, при выборе подходящего индуцирующего множества, может существенно снизить затраты на вычисления. Поскольку используются не все точки из обучающей выборки, она тем самым *разрежается*, поэтому назовем гауссовский процесс, обученный таким образом, *разреженным*.

Как объединить эти два подхода, чтобы использовать их сильные стороны?

Правило Маферона

Пусть и имеют совместное нормальное многомерное распределение. Тогда условное распределение на при $= \boldsymbol{\beta}$ вычисляется как

$$(\boldsymbol{a} \mid \boldsymbol{b} = \boldsymbol{\beta}) \stackrel{\mathrm{d}}{=} \boldsymbol{a} + \operatorname{Cov}(\boldsymbol{a}, \boldsymbol{b}) \operatorname{Cov}(\boldsymbol{b}, \boldsymbol{b})^{-1} (\boldsymbol{\beta} - \boldsymbol{b})$$

Эта теорема позволяет разложить апостериорное распределение в сумму априорного и обновления из-за несоответствия данных априорному распределению.

Применение теоремы Маферона

Применение теоремы Маферона к взвешенному и функциональному подходам позволяет получить следующий вид распределений:

$$\boldsymbol{w} \mid \boldsymbol{y} \stackrel{\mathrm{d}}{=} \boldsymbol{w} + \Phi^\top \left(\Phi\Phi^\top + \sigma^2 \boldsymbol{I}\right)^{-1} (\boldsymbol{y} - \Phi\boldsymbol{w} - \varepsilon)$$
$$f_m \mid u \stackrel{\mathrm{d}}{=} f_m + \boldsymbol{K}_{m,v} \boldsymbol{K}_{v,v}^{-1} (\boldsymbol{u} - \boldsymbol{f}_v)$$

Отсюда можно увидеть, что в взвешенном подходе вся сложность приходится на вычисление слагаемого обновления, в то время как в функциональном обновление линейно по длине входа m, но имеет кубическую сложность в сэмплировании из априорного распределения. Таким образом, мы можем совместить два подхода, чтобы существенно ускорить сэмплирование.

Модель эффективного сэмплирования

Итоговая модель, совмещающая в себе сильные стороны обоих подходов, выглядит следующим образом

$$(f \mid \boldsymbol{u})(\cdot) \overset{\mathrm{d}}{\approx} \sum_{i=1}^{\ell} w_i \phi_i(\cdot) + \sum_{j=1}^{v} h_j k\left(\cdot, \boldsymbol{z}_j\right),$$

где $\boldsymbol{h} = \boldsymbol{K}_{v,v}^{-1}(\boldsymbol{u} - \Phi \boldsymbol{w})$. Таким образом, разложение с помощью теоремы Маферона позволяет существенно ускорить сэмплирование.

Постановка задачи

Критерий качества:

1. Точность

$$A = \frac{1}{m}\sum_{i=1}^{m}[\hat{y}_i = y_i]$$

2. AUROC score
3. Скорость работы

Эксперимент

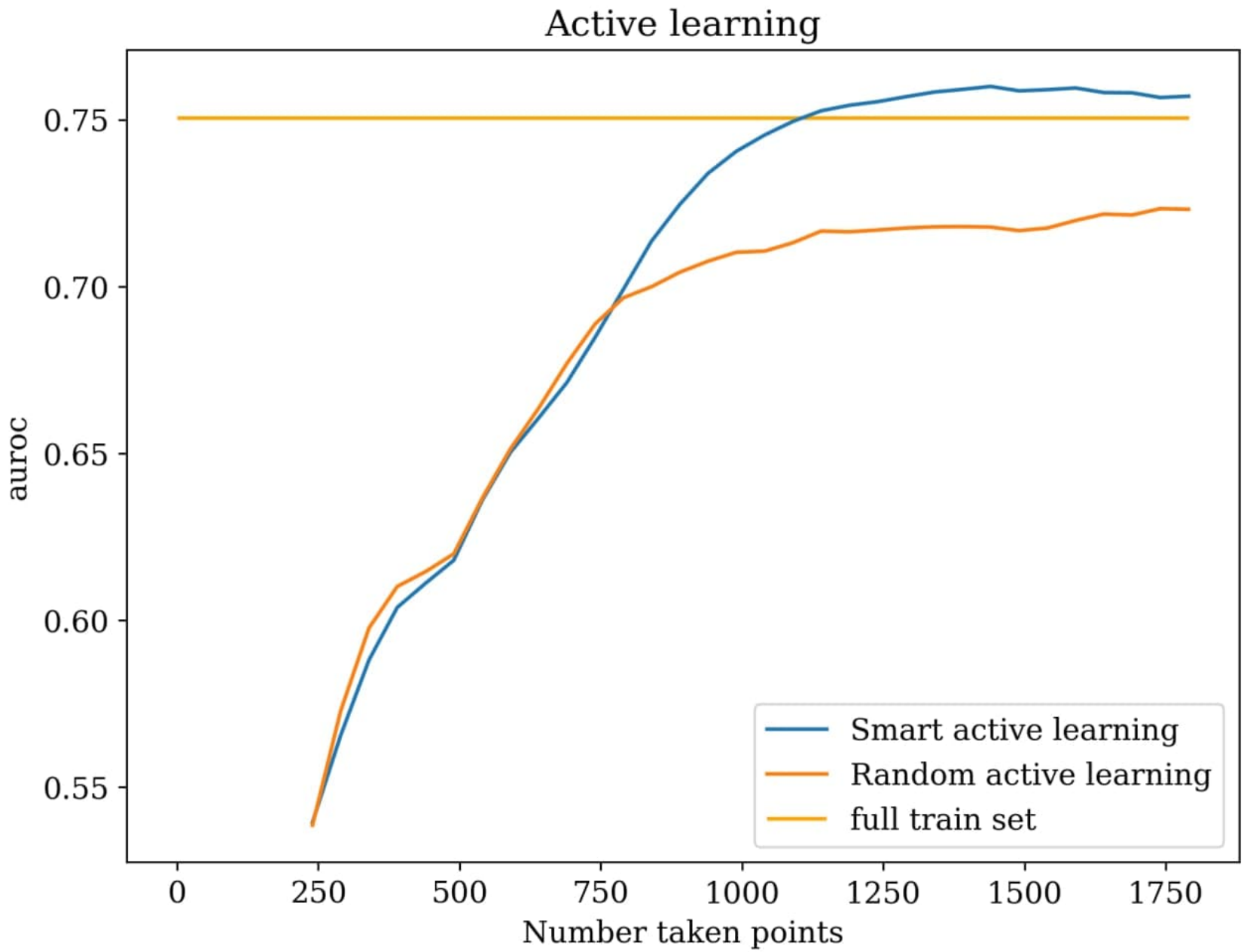
Для эксперимента был рассмотрен датасет default of credit card clients.

Data Set Characteristics:	Multivariate	**Number of Instances:**	30000	**Area:**	Business
Attribute Characteristics:	Integer, Real	**Number of Attributes:**	24	**Date Donated**	2016-01-26
Associated Tasks:	Classification	**Missing Values?**	N/A	**Number of Web Hits:**	568344

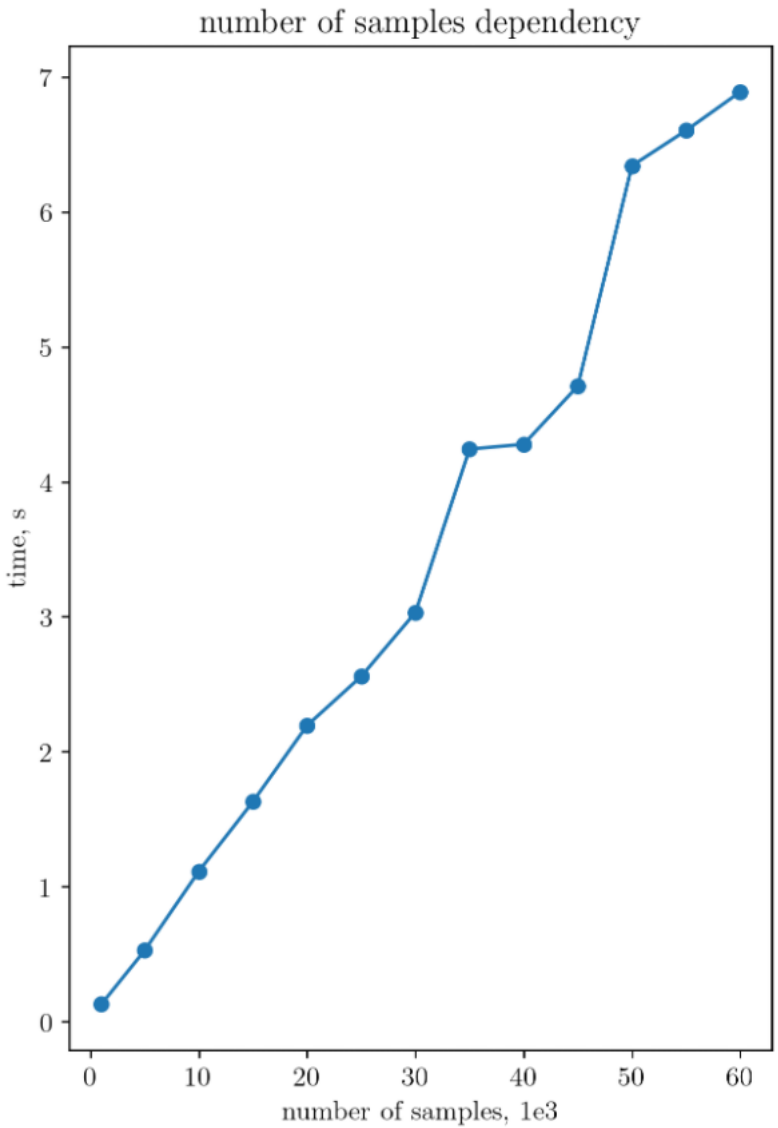
Эксперимент проводился следующим образом:

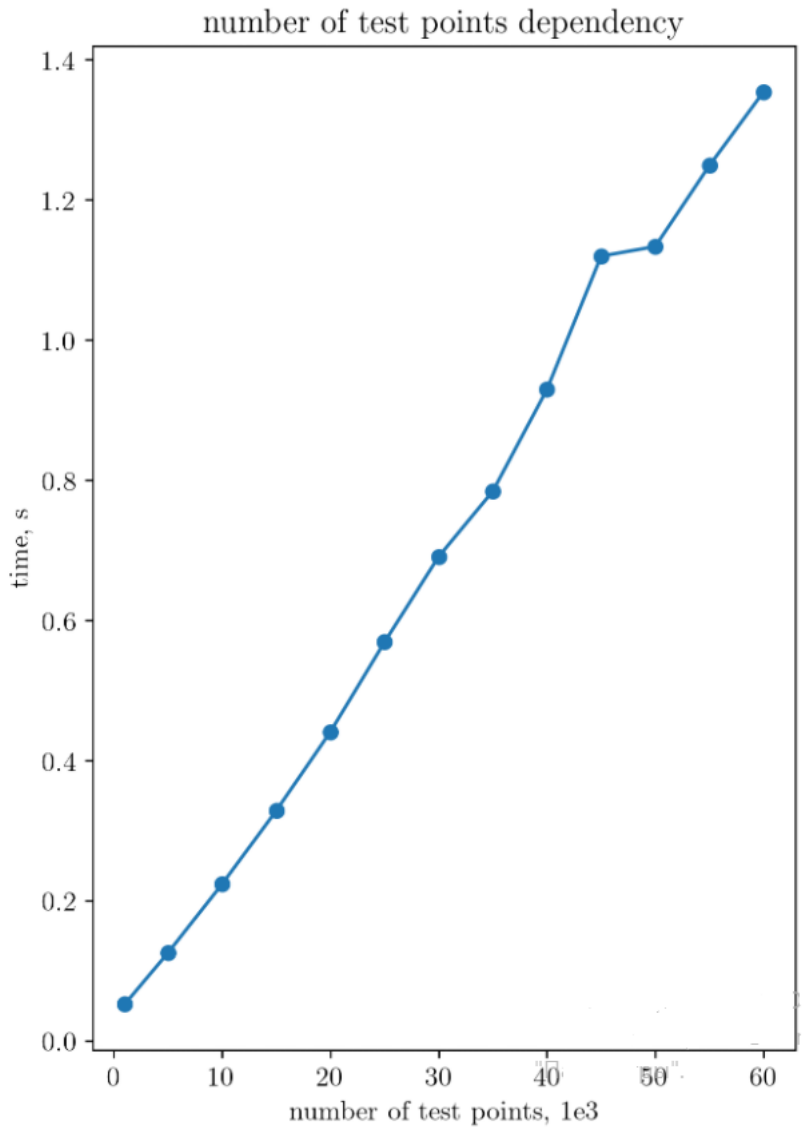
1. Обучение модели на всей обучающей выборке и определение метрики качества (auroc).
2. Последовательное обучение на малой подвыборке с добавлением в выборку новых точек, отобранных по значению метрики неопределенности.
3. Последовательное обучение на малой подвыборке с добавлением в выборку новых точек, отобранных случайно.
4. Сравнение результатов.
5. Тестирование скорости сэмплирования.

Текущие результаты



Текущие результаты





Анализ ошибки

Обозначим за $GP(0, k)$ гауссовский процесс с средним 0 и ковариационной функцией k. Пусть $f \sim (0, k)$. Обозначим за $f \mid \boldsymbol{y}$ точное апостериорное распределение, $f^{(s)}$ - апостериорное распределение в терминах функционального подхода, $f^{(d)}$ - апостериорное распределение в терминах эффективного подхода и $f^{(w)}$ - априорное распределение в терминах взвешенного подхода.Тогда

$$
\begin{aligned}
& W_{2,L^2(\mathcal{X})}\left(f^{(d)}, f \mid \boldsymbol{y}\right) \leq \\
\leq \quad & W_{2,L^2(\mathcal{X})}\left(f^{(s)}, f \mid \boldsymbol{y}\right) + C_1 W_{2,C(\mathcal{X})}\left(f^{(w)}, f\right),
\end{aligned}
$$

где $W_{2,L^2(\mathcal{X})}, W_{2,C(\mathcal{X})}$ - расстояния Вассерштейна в пространствах $L^2(\mathcal{X})$ и $C(\mathcal{X})$ и

$$
C_1 = \sqrt{2\left(1 + \|k\|^2_{C(\mathcal{X}^2)} \left\|\boldsymbol{K}^{-1}_{mm}\right\|^2_{L(\ell^\infty;\ell^1)}\right)}.
$$

Ссылки

Ключевые работы

Rasmussen, C. E., Williams, C. K. I. (2005). Gaussian processes for machine learning. MIT Press.
J.T. Wilson, V. Borovitskiy, A. Terenin, P. Mostowsky, M. P. Deisenroth. Efficiently Sampling Functions from Gaussian Process Posteriors. International Conference on Machine Learning, 2020