Министерство образования и науки Российской Федерации
Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Студент Вайсер Кирилл Олегович

Вычислительно эффективное сэмплирование из гауссовского процесса в задаче активного обучения

03.03.01 — Прикладные математика и физика

Бакалаврская диссертация

Научный руководитель: Панов Максим Евгеньевич
кандидат физико-математических наук

Москва
2021 г.

Аннотация

В данной работе рассматривается применение гауссовских процессов к задачам активного обучения. В метода отбора точек используются функции неопределенности. Для их вычисления необходимо сэмплировать большое число реализаций случайных величин. Исследуется задача быстрого сэмплирования из гауссовских процессов. В методе быстрого сэмплирования используется идея разложения апостериорного распределения. Получена линейная скорость сэмплирования. Получены результаты, подтверждающие состоятельность метода. Разработана методика отбора точек.

Ключевые слова: активное обучение, гауссовские процессы, правило Маферона, быстрое сэмплирование.

Содержание

1 Введение

Одним из важнейших классов моделей машинного обучения являются байесовские модели [1], в которых вводится априорное распределение на параметры модели. В процессе обучения определяется апостериорное распределение этих параметров, которое позволяет получить распределение результата выхода модели. Подобный подход позволяет извлекать намного больше информации из обучающей выборки, поскольку становится возможным вычисление уверенности модели в своем ответе, а так же расчет обоснованности модели. Примечательным представителем класса байесовских моделей являются гауссовские процессы. Они служат мощным инструментом для задания распределения функций на заданной области. После обучения гауссовский процесс задает распределение на функции с фиксированным значением в рассмотренных точках, позволяя находить лучшую аппроксимацию истинной зависимости.

Свое применение гауссовские процессы находят в активном обучении — одной из областей машинного обучения [2]. Идея методов этой области заключается в выборе определенных точек для обучения в задачах, где затруднительно получить в распоряжение большие объемы размеченных данных. Такие трудности могут быть вызваны большой стоимостью получения этих данных. Например, сбор данных для обучения модели распознавания образов требует, чтобы большое число изображений было размечено человеком. Если количество изображений исчисляется сотнями тысяч, то использование человеческого труда будет очень дорогим и долгим. Однако если у модели будет возможность выбирать более информативные изображения для разметки, то число изображений, которые требуют разметки, существенно снизится.

Гауссовские процессы позволяют получить распределение ответов модели, что дает возможность учитывать неуверенность модели в своем ответе. Это

открывает возможность разумного метода выбора точек для активного обучения. Чем выше неуверенность модели в том, какой ответ выдать на определенной точке, тем больше информации эта точка даст модели при включении в обучающую выборку. Однако для вычисления метрик, определяющих неуверенность модели, как правило необходимо вычислять значение функций от выхода модели. В силу того, что выход модели в таком подходе —случайная величина, вычислить это значение невозможно аналитически. В таком случае часто используются методы Монте-Карло. Однако для их использования необходимо сэмплировать большое количество реализаций случайной величины. Но в случае использования гауссовского процесса такое сэмплирование — долгая операция. Цель данной работы — рассмотрение метода быстрого сэмплирования и его использование в задаче активного обучения. Ключевой идеей предлагаемого метода является разложение апостериорного распределения в сумму априорного распределения и некоторой поправки.

2 Постановка задачи

2.1 Активное обучение

Пусть дана выборка

$$\mathfrak{D} = \{\mathbf{x}_i\} \quad i = 1, \ldots, n, \quad \mathbf{x}_i \in \mathbf{R}^d$$

и оракул

$$g : \mathbf{R}^d \mapsto \{-1, 1\}.$$

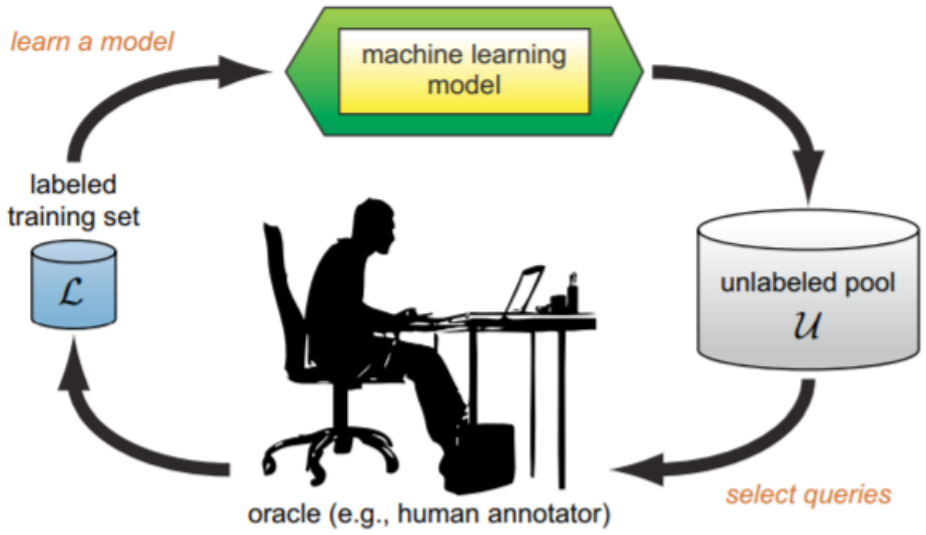


Рис. 1: Процесс машинного обучения. Изображение взято из [2]

Оракул — это функция, возвращающая истинную метку объекта. Однако, в рассмотрении модели активного обучения, вызов оракула - дорогая операция, поэтому цель конструируемой модели - достигнуть требуемого качества, обратившись к оракулу как можно меньшее количество раз. Мы задаемся целью построить модель

$$\boldsymbol{f}(\mathbf{x}, \boldsymbol{\theta}) : \mathbf{R}^d \times \mathbf{R}^h \mapsto [0, 1], \tag{2.1}$$

где $\mathbf{R}^h$ — пространство парамеров модели. Модель (2.1) аппроксимирует вероятность принадлежности объекта $\mathbf{x}$ классу 1. При построении модели мы решаем задачу максимизации критерия качества AUC [3]:

$$AUC = \frac{\sum_{\mathbf{x}_0 \in \mathcal{D}^0} \sum_{\mathbf{x}_1 \in \mathcal{D}^1} \mathbf{1}\left[f\left(\mathbf{x}_0, \boldsymbol{\theta}\right) < f\left(\mathbf{x}_1, \boldsymbol{\theta}\right)\right]}{|\mathcal{D}^0| \cdot |\mathcal{D}^1|},$$

где $\mathcal{D}^0, \mathcal{D}^1$ —множества объектов с негативной и позитивной меткой соответственно, а $\boldsymbol{\theta}$ — параметры модели.

В задаче активного обучения критически важным является способ выбора новых точек. Поскольку мы стремимся снизить число обращений к оракулу, мы должны выбирать точки, которые будут наиболее информативными для модели. Информативность точки мы будем интерпретировать как значение неопределенности модели в этой точке. Соответственно, чем выше неопределенность, тем более информативна эта точка для модели.

Для оценки неопределенности используются различные функции. Например, функция уверенности:

$$u_1\left(\mathbf{x}\right) = -\left|f\left(\mathbf{x}\right) - \frac{1}{2}\right|,$$

где $f\left(\mathbf{x}\right)$ — вероятность объекта $\mathbf{x}$ принадлежать к классу 1.

Также широко используется энтропия Шеннона [4]:

$$\mathcal{H}\left(f\left(\mathbf{x}\right)\right) = -\left(f\left(\mathbf{x}\right)\log f\left(\mathbf{x}\right) + \left(1 - f\left(\mathbf{x}\right)\right)\log\left(1 - f\left(\mathbf{x}\right)\right)\right).$$

. В данной работе мы будем использовать функцию эпистемической энтропии, которая отвечает взаимной информации [5]:

$$MI = \mathbf{E}\mathcal{H}\left(f\left(\mathbf{x}\right)\right) - \mathcal{H}\left(\mathbf{E}f\left(\mathbf{x}\right)\right).$$

Обычно аналитическое вычисление матожидания энтропии невозможно. Для того, чтобы вычислить его численно, используется метод Монте-Карло [6]:

$$\mathbf{E}X \approx \frac{1}{N}\sum_{i=1}^{N} X_i$$

Для использования этого метода необходимо сэмплировать большое количество реализаций случайной величины для снижения ошибки аппроксимации [7]. Зависимость дисперсии от числа сэмплов выглядит следующим образом:

$$\sigma \leqslant \frac{1}{2}\sqrt{\frac{1}{s}},$$

где σ —дисперсия аппроксимируемой величины и s —число реализаций. Поэтому используемая модель должна обладать возможностью быстро сэмплировать на выбранном множестве точек.

Таким образом, процесс активного обучения можно записать в следующем виде:

1. Инициализировать начальную выборку $D = \{(x_i, y_i)\}_{i=1}^{n_0}$, функцию неопределенности $u : R^d \mapsto R$, модель $f : R^d \mapsto [0, 1]$

2. Обучить модель на выборке D.

3. Выбрать из выборки $\mathfrak{D}$ точки, на которых u достигает максимально возможного значения, вычислить в них истинную метку, добавить получившиеся пары в выборку D.

4. Повторять шаги 2, 3 до выполнения критерия останова. Критерием останова обычно служит стабилизация значения критерия качества на одном уровне.

2.2 Гауссовские процессы

В качестве модели f мы будем рассматривать гауссовский процесс.

Определение 1 (Гауссовский процесс). Гауссовский процесс —случайный процесс, любая конечная совокупность сечений которого имеет совместное нормальное распределение.

Гауссовский процесс характеризуется функцией среднего и ковариационной функцией:

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

Если f — гауссовский процесс с функцией среднего m и ковариационной функцией k, то он записывается как $f \sim GP(\mu, k)$ и распределение вектора $\mathbf{f}_* = f(X_*), X_* \subset R^d$ есть многомерное нормальное:

$$\mathbf{f}_* \sim \mathcal{N}(m(X_*), k(X_*, X_*)).$$

Обучение гауссовского процесса означает вычисление апостериорного распределения на основе наблюдаемых обучающих данных и оптимизацию параметров ковариационной функции для минимизации функции потерь. Обозначим $\boldsymbol{f}_m \mid \mathbf{y}$ за апостериорное распределение гауссовского процесса в точках $\{\boldsymbol{x'}\}_{i=1}^m = X_m$ после наблюдения выборки $\{\mathbf{x}_i, y_i\}_{i=1}^n$. Тогда его среднее и матрица ковариации записываются как

$$\begin{aligned} \boldsymbol{m}_{m|n} &= \boldsymbol{K}_{m,n}\left(\boldsymbol{K}_{n,n} + \sigma^2\boldsymbol{I}\right)^{-1}\mathbf{y} \\ \boldsymbol{K}_{m,m|n} &= \boldsymbol{K}_{m,m} - \boldsymbol{K}_{m,n}\left(\boldsymbol{K}_{n,n} + \sigma^2\boldsymbol{I}\right)^{-1}\boldsymbol{K}_{n,m}, \end{aligned} \tag{2.2}$$

а сэмплирование производится с помощью стандартной схемы

$$\boldsymbol{f}_m \mid \mathbf{y} = \boldsymbol{m}_{m|n} + \boldsymbol{K}_{m,m|n}^{1/2}\boldsymbol{\zeta}, \quad \boldsymbol{\zeta} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}). \tag{2.3}$$

Оптимизация параметров ковариацинной функции проводится в процессе максимизации маргинального правдоподобия обучающей выборки:

$$\log p(\mathbf{y} \mid X, \eta) = -\frac{1}{2}\mathbf{y}^\top\left(K(\eta) + \sigma_n^2 I\right)^{-1}\mathbf{y} - \frac{1}{2}\log\left|K(\eta) + \sigma_n^2 I\right| - \frac{n}{2}\log 2\pi \to \max_\eta,$$

где η — параметры ковариационной функции.

Так как выход модели, основанной на гауссовском процессе, зависит от реализации случайной величины ξ в (2.3), он так же является случайной величиной. Это позволяет получить как результат целое распределение значений на тестовой выборке. Благодаря этому становится удобно вычислять

неопределенность модели в точках. В то же время такая модель становится чувствительной к скорости сэмплирования.

Как видно, точное вычисление апостериорных моментов (2.2) требует не менее, чем $O\left(n^3\right)$ операций, поскольку необходимо вычислить обратную матрицу размера $n \times n$. Для сэмплирования необходимо вычислить разложение Холецкого [8] для ковариационной матрицы тестовой выборки, что так же требует $O\left(m^3\right)$ операций. Таким образом, традиционная схема обладает кубической сложностью по размеру как обучающей, так и тестовой выборки. Это делает прямое применение гауссовских процессов невозможным, поскольку уже даже для небольшого числа данных вычислительные затраты становятся слишком большими.

3 Быстрое сэмплирование

Чтобы решить эту проблему, рассмотрим два подхода к рассмотрению гауссовских процессов [9].

3.1 Функциональный подход

Как известно из (2.2), сложность вычисления ковариационной матрицы апостериорного распределения равняется $O\left(n^3\right)$, где n — размер обучающей выборки. Если n велико, то вычисление такой матрицы потребует большое количество вычислительных ресурсов. В таком случае можно рассмотреть подмножество *индуцирующих* точек $\boldsymbol{Z} = \{z_j\}_{j=1}^{v}$. В данной работе не рассматривается способ выбора индуцирующего подмножества. Разные методы такого выбора описаны в [10–12]. Мы предполагаем наличие возможности сэмплировать $\mathbf{u} = f\left(\boldsymbol{Z}\right) \sim \mathcal{N}\left(\mu_u, \Sigma_u\right)$. Тогда апостериорное распределение на тестовой выборке $f_m \mid \boldsymbol{u}$ имеет следующие моменты

$$\begin{aligned} \boldsymbol{\mu}_{m|v} &= \boldsymbol{K}_{m,v}\boldsymbol{K}_{v,v}^{-1}\boldsymbol{\mu}_v \\ \boldsymbol{K}_{m,m|v} &= \boldsymbol{K}_{m,m} + \boldsymbol{K}_{m,v}\boldsymbol{K}_{v,v}^{-1}\left(\boldsymbol{\Sigma}_{\boldsymbol{u}} - \boldsymbol{K}_{v,v}\right)\boldsymbol{K}_{v,v}^{-1}\boldsymbol{K}_{v,m} \end{aligned} \tag{3.1}$$

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

Однако использование функционального подхода позволяет упростить лишь обучение процесса. Сэмплирование по-прежнему предполагает использование традиционной схемы (2.3), которая имеет кубическую сложность.

3.2 Взвешенный подход

Помимо функционального подхода, существует так же взвешенный подход к рассмотрению гауссовского процесса. Рассмотрим байесовскую линейную модель

$$f(\mathbf{x}) = \boldsymbol{\varphi}(\mathbf{x})^\top \mathbf{w}, \quad y = f(\mathbf{x}) + \varepsilon, \quad \varepsilon \sim \mathcal{N}\left(0, \sigma_n^2\right) \tag{3.2}$$

где $\mathbf{w}$ — параметры модели с априорным распределением $\mathcal{N}(0, \Sigma_p)$. Тогда апостериорное распределение вектора ответов:

$$\begin{aligned} f_m \mid \mathbf{x}_m, X, \mathbf{y} \sim \mathcal{N}(&\boldsymbol{\varphi}_m^\top \Sigma_p \boldsymbol{\Phi} \left(K + \sigma_n^2 I\right)^{-1} \mathbf{y} \\ &\boldsymbol{\varphi}_m^\top \Sigma_p \boldsymbol{\varphi}_m - \boldsymbol{\varphi}_m^\top \Sigma_p \boldsymbol{\Phi} \left(K + \sigma_n^2 I\right)^{-1} \boldsymbol{\Phi}^\top \Sigma_p \boldsymbol{\varphi}_m), \end{aligned} \tag{3.3}$$

где $\boldsymbol{\Phi} = \boldsymbol{\varphi}(X), \quad K = \boldsymbol{\Phi}^\top \Sigma_p \boldsymbol{\Phi}$. Здесь мы можем определить ковариационную функцию как

$$k(\mathbf{x}, \mathbf{x}') = \boldsymbol{\varphi}(\mathbf{x})^\top \Sigma_p \boldsymbol{\varphi}(x') = \boldsymbol{\psi}(\mathbf{x})^\top \boldsymbol{\psi}(x'),$$

где $\boldsymbol{\psi}(\mathbf{x}) = \Sigma_p^{1/2} \boldsymbol{\varphi}(\mathbf{x})$. Апостериорное распределение гауссовского процесса с такой ковариационной функцией будет совпадать с (3.3)

В взвешенном подходе случайной величиной являются веса $\mathbf{w}$ функций $\boldsymbol{\varphi}$. Сэмплирование реализации этих весов позволяет получить неслучайную модель. Это означает, что значение модели в точке $f(\mathbf{x})$ вычисляется за $O(1)$, следовательно вычисление значения функции на тестовой выборке размера m затребует $O(m)$ операций. Это существенный прирост производительности по сравнению с $O\left(m^3\right)$ в традиционном методе.

В работе [13] было показано, что ядро k может быть рассмотрено, как скалярное произведение в порожденном ядром гильбертовом пространстве (RKHS) [9] $\mathcal{H}$ с отображением $\boldsymbol{\psi} : X \mapsto \mathcal{H}$. Если $\mathcal{H}$ сепарабельно, то это скалярное произведение можно аппроксимировать конечномерным отображением $\boldsymbol{\varphi} : X \mapsto \mathbf{R}^l$:

$$k(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\psi}(\mathbf{x}), \boldsymbol{\psi}(\mathbf{x}') \rangle_{\mathbf{H}} \approx \boldsymbol{\varphi}(\mathbf{x})^\top \boldsymbol{\varphi}(\mathbf{x}')$$

Отображение $\boldsymbol{\varphi}$ можно построить, используя следующую теорему.

Теорема 1 (Бохнера). Пусть функция $f : \mathbf{R}^n \mapsto \mathbf{C}$ положительно определенная, непрерывная и $f(\mathbf{0}) = 1$. Тогда существует борелевская вероятностная мера μ такая, что

$$f(t) = \int_{\mathbf{R}^n} e^{-itx} d\mu(\mathbf{x})$$

Из теоремы Бохнера следует, что для ковариационной функции стационарного процесса существует разложение Фурье [14].Из этого следует, что в качестве функций $\boldsymbol{\varphi}$ в (3.2) можно взять функции следующего типа:

$$\boldsymbol{\varphi}(\mathbf{x}) = \sqrt{\frac{2}{l}} \cos \boldsymbol{\theta}_i^\top \mathbf{x} + \tau_i,$$
$$\tau_i \sim U(0, 2\pi),$$

$\boldsymbol{\theta}_i$ сэмплируется пропорциально спектральной плотности ядра и l —выбранная размерность аппроксимации. Как было изложено выше, f сохраняет нормальное распределение своих сечений. Однако теперь случайность процесса контролируется распределением весов $\mathbf{w}$.

Однако взвешенный подход обладает рядом недостатков [15]. Аппроксимации на основе методов разложения в Фурье испытывают вырождение дисперсии с увеличением n [16]. Это происходит из-за нарушения требования стационарности процессов. В общем случае процессы, происходящие в реальных задачах, не обладают таким свойством.

3.3 Правило Маферона

Взвешенный и функциональный подходы обладают разными преимуществами и недостатками. Функциональный подход позволяет эксплуатировать большие обучающие выборки благодаря выбору индуцирующего множества. Однако в этом подходе сэмплирование все так же имеет кубическую сложность по длине входа. Взвешенный подход, напротив, позволяет сэмплировать

за линейное время, однако приводит к большим потерям качества при увеличении выборки. Благодаря тому, что эти подходы обладают противоположными достоинствами и недостатками, их можно объединить и использовать сильные стороны каждого из них. Ключевой идее для такого объединения будет служить

Теорема 2 (Правило Маферона). Пусть $\mathbf{a}$ и $\mathbf{b}$ имеют совместное нормальное многомерное распределение. Тогда условное распределение на $\mathbf{a}$ при $\mathbf{b} = \boldsymbol{\beta}$ вычисляется как

$$(\boldsymbol{a} \mid \boldsymbol{b} = \boldsymbol{\beta}) \stackrel{\mathrm{d}}{=} \boldsymbol{a} + \operatorname{Cov}(\boldsymbol{a}, \boldsymbol{b}) \operatorname{Cov}(\boldsymbol{b}, \boldsymbol{b})^{-1} (\boldsymbol{\beta} - \boldsymbol{b}) \tag{3.4}$$

Эта теорема позволяет разложить апостериорное распределение в сумму двух слагаемых. Первое слагаемое соответствует априорному распределению, а второе является поправкой после наблюдения данных. Применение (3.4) к взвешенному (3.3) и функциональному (3.1) подходам позволяет получить следующий вид распределений:

$$\begin{aligned} &\mathbf{w} \mid \mathbf{y} \stackrel{\mathrm{d}}{=} \mathbf{w} + \boldsymbol{\varphi}^\top \left(\boldsymbol{\varphi}\boldsymbol{\varphi}^\top + \sigma^2 \boldsymbol{I}\right)^{-1} (\mathbf{y} - \boldsymbol{\varphi}\mathbf{w} - \boldsymbol{\varepsilon}) \\ &f_m \mid u \stackrel{\mathrm{d}}{=} f_m + \boldsymbol{K}_{m,v} \boldsymbol{K}_{v,v}^{-1} (\boldsymbol{u} - \boldsymbol{f}_v) \end{aligned} \tag{3.5}$$

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.

Итоговая модель, совмещающая в себе априорное распределение взвешенного подхода и обновление функционального (3.5), выглядит следующим образом

$$(f \mid \boldsymbol{u})(\cdot) \stackrel{\mathrm{d}}{\approx} \sum_{i=1}^{\ell} w_i \boldsymbol{\varphi}_i(\cdot) + \sum_{j=1}^{v} h_j k(\cdot, \boldsymbol{z}_j), \tag{3.6}$$

где $\boldsymbol{h} = \boldsymbol{K}_{v,v}^{-1} (\boldsymbol{u} - \boldsymbol{\varphi}\mathbf{w})$.

Таким образом, итоговая модель аппроксимации гауссовского процесса имеет сложность сэмплирования $O(m)$ и сложность обучения $O\left(v^3\right)$, где m — размер тестовой выборки и v — размер индуцирующего множества.

Предложенная идея разложения универсальна и позволяет использовать также и другие подходы к рассмотрению гауссовских процессов [17]

3.4 Ошибка аппроксимации

В работе [18] показано, что ошибка аппроксимации гауссовского процесса моделью (3.6) может быть оценена с помощью следующей

Теорема 3. Пусть $\mathcal{X} \subseteq \mathbf{R}^d$ —компакт и $f \sim GP(0, k)$- стохастически непрерывный стационарный гауссовский процесс. Обозначим за $f \mid \mathbf{y}$ точное апостериорное распределение, $f^{(s)}$ - апостериорное распределение в терминах функционального подхода, $f^{(d)}$ - апостериорное распределение в терминах объединенного подхода и $f^{(w)}$ - априорное распределение в терминах базисного подхода.Тогда

$$W_{2,L^2(\mathcal{X})}\left(f^{(d)}, f \mid \mathbf{y}\right) \leqslant \underbrace{W_{2,L^2(\mathcal{X})}\left(f^{(s)}, f \mid \mathbf{y}\right)}_{\text{ошибка в апостериорном распределении}} + \underbrace{C_1 W_{2,C(\mathcal{X})}\left(f^{(w)}, f\right)}_{\text{ошибка в априорном распределении}},$$

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

4 Вычислительный эксперимент

4.1 Алгоритм эксперимента

Проведение вычислительного эксперимента преследует следующие цели:

1. Исследование количества точек, требуемого для достижения качества, достигаемого на всей выборке.
2. Сравнение методов выбора точек
3. Сравнение скорости сэмплирования традиционного (2.3) и быстрого (3.6) методов.

4.2 Используемые модели

Были рассмотрены следующие модели:

1. Full model

   Обучение модели на всей обучающей выборке

2. Smart active model

   Модель активного обучения. На каждом шаге активного обучения выполняет E шагов обучения задачи классификации, после чего добавляет k точек в текущую обучающую выборку на основе значения функции неопределенности.

3. Random active model

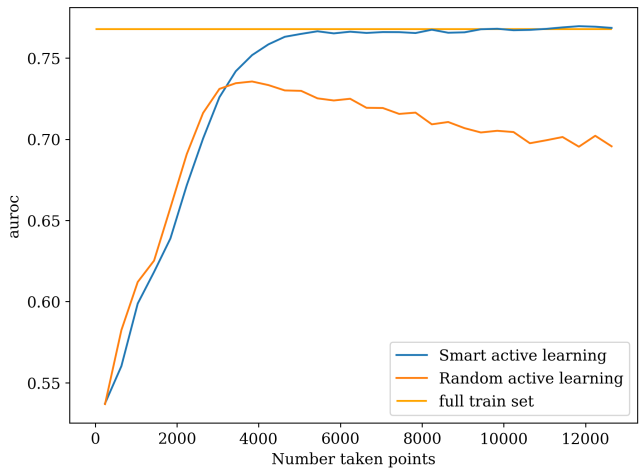
   Модель активного обучения. На каждом шаге активного обучения выполняет E шагов обучения задачи классификации, после чего вычисляет случайным образом выбирает k из всех точек полной обучающей выборки и добавляет в текущую обучающую выборку.

4.3 Отбор точек

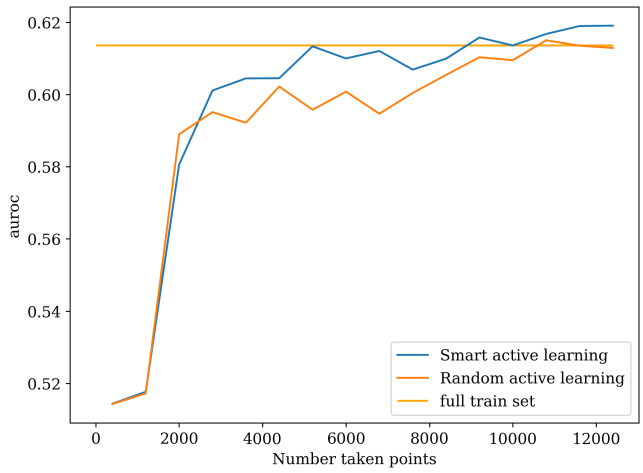
Необходимо предложить методику отбора точек для добавления в выборку. Естественным способом будет добавление в выборку точек с наибольшей неопределенностью. Однако такой метод обладает рядом недостатков. Поскольку новые точки в выборку добавляюся батчами, то если в одном батче будут точки, близкие друг к другу, то это приведет к ухудшению качества итоговой аппроксимации, так как эти точки несут информацию об одной и той же области. Вследствие этого будет происходить переобучение. Чтобы избежать этой проблемы, предлагаются два метода:

1. Отсортировать все множество точек по норме L_2 и разбить на K областей. Выбрать из каждой области k/K точек с наибольшей неопределенностью.

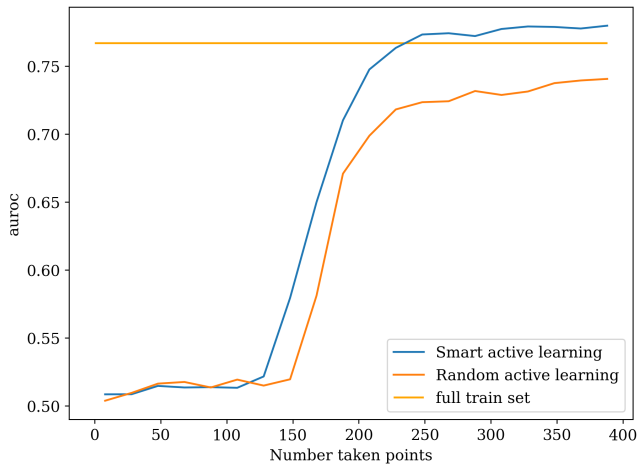
2. Вычислить значение неопределенности для всей выборки, после чего отсортировать точки по убыванию неопределенности. Добавлять точки в выборку до тех пор, пока не будет отобрано k точек, поддерживая условие превосходства разности норм между любой парой из добавленных точек выше заранее заданного порога.



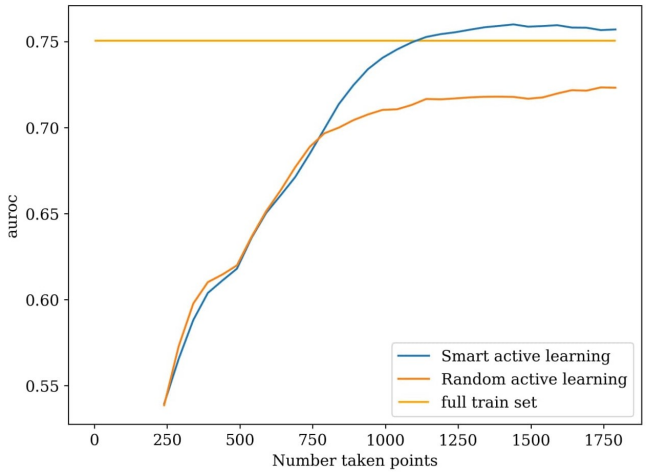
(a) датасет DCCC



(b) датасет DOTA



(c) датасет SGC



(d) датасет ROM

Рис. 2: Сравнение качества при обучении на всей доступной выборке, выборке, полученной с помощью функции неопределенности и выборке, полученной случайным выбором точек.

Результаты Предложенный метод обладает требуемыми свойствами.Во-первых, он позволяет достичь и даже превзойти качество, получанное на всей выборке, при этом обратившись к оракулу меньшее число раз. Во-вторых, использование точек с наивысшей неопределенностью позволяет повышать качество с увеличением числа взятых точек, в то время как выбор точек случайным образом ведет к ухудшению качества.

Было проведено так же сравнение скорости сэмплинга традиционным и предлагаемым способом.

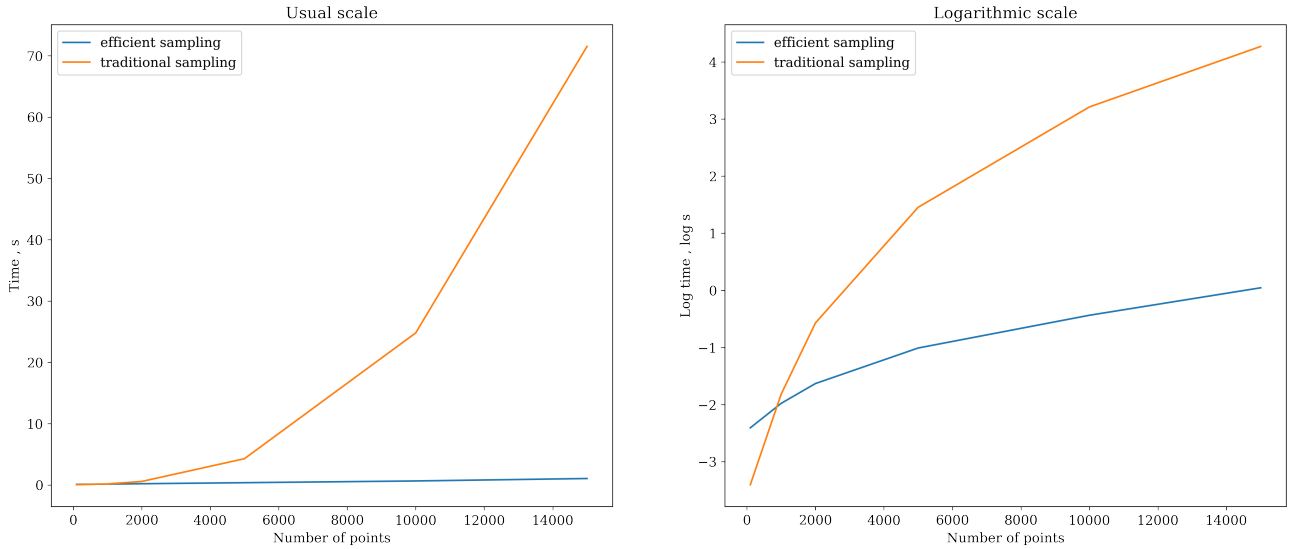


Рис. 3: Зависимость скорости сэмплирования от числа точек.

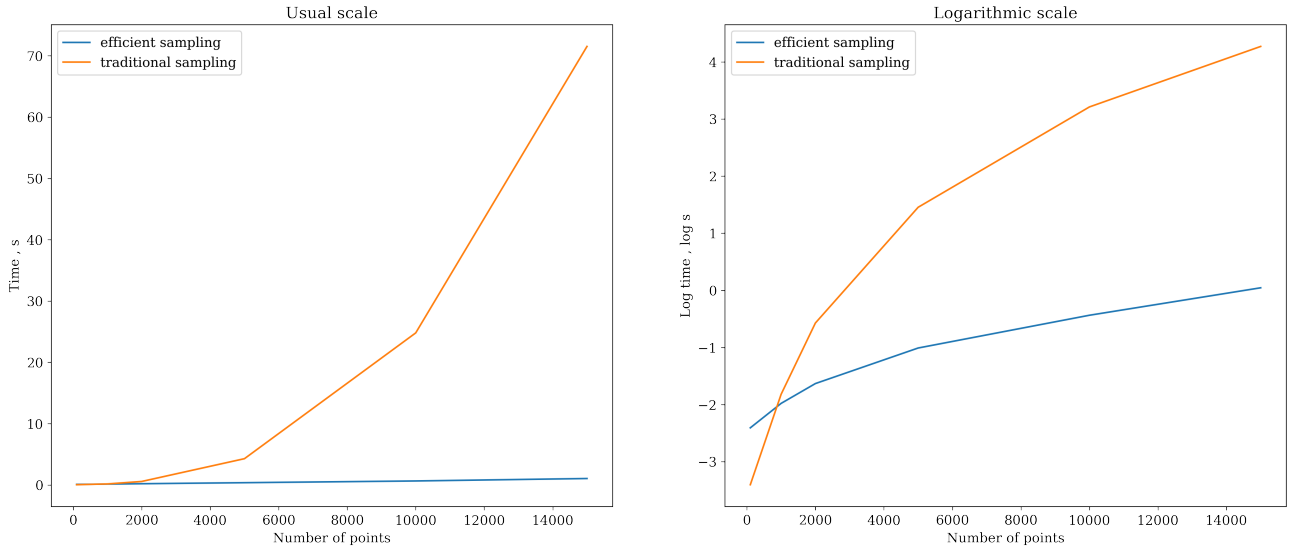


Рис. 4: Зависимость скорости сэмплирования от числа сэмплов.

Таблица 1: Описание выборок для экспериментов

Выборка $\mathfrak{D}$	Размер train	Размер test	Объекты	Признаки
DCCC	24000	6000	30000	13
SGC	800	200	1000	21
DOTA	50000	20000	70000	116
ROM	4000	2000	8000	17

5 Заключение

Предложенный метод позволяет достичь существенного ускорения сэмплирования по сравнению с традиционными методами. Сложность сэмплирования понизилась с кубической до линейной по длине входа. Также в рамках данной работы были проведены эксперименты на обучающих выборках разного размера, подтверждающие состоятельность предложенного метода. Быстрое сэмплирование необходимо в задачах, где невозможно производить аналитические расчеты. Предложенный метод позволяет приобщить класс гауссовских процессов к моделям, удобных и эффективных в применении к таким задачам.

Список литературы

[1] D. Barber. *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[2] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[3] Kai Feng, Han Hong, Ke Tang, and Jingyuan Wang. Decision making with machine learning and roc curves. *Available at SSRN 3382962*, 2019.

[4] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.

[5] Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *arXiv preprint arXiv:1112.5745*, 2011.

[6] Dirk P. Kroese, T. Brereton, T. Taimre, and Z. Botev. Why the monte carlo method is so important today. *Wiley Interdisciplinary Reviews: Computational Statistics*, 6:386–392, 2014.

[7] Robert Eric Beard, Teivo Pentikäinen, and Erkki Pesonen. *Risk Theory*. Springer Science, 1984.

[8] Claude Brezinski. La méthode de cholesky. *Revue d'histoire des mathématiques*, 11(2):205–238, 2005.

[9] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

[10] James Hensman, Nicolas Durrande, and Arno Solin. Variational fourier features for gaussian processes. *Journal of Machine Learning Research*, 18(151):1–52, 2018.

[11] Michalis Titsias. Variational learning of inducing variables in sparse gaussian processes. In David van Dyk and Max Welling, editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 16–18 Apr 2009. PMLR.

[12] Edward Snelson and Zoubin Ghahramani. Sparse gaussian processes using pseudo-inputs. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2006.

[13] Bernhard Schölkopf and A.J. Smola. *Smola, A.: Learning with Kernels - Support Vector Machines, Regularization, Optimization and Beyond. MIT Press, Cambridge, MA*, volume 98. 01 2001.

[14] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2008.

[15] Carl Edward Rasmussen and Joaquin Quiñonero Candela. Healing the relevance vector machine through augmentation. In *Proceedings of the 22nd International Conference on Machine Learning*, ICML '05, page 689–696, New York, NY, USA, 2005. Association for Computing Machinery.

[16] Zi Wang, Clement Gehring, Pushmeet Kohli, and Stefanie Jegelka. Batched large-scale bayesian optimization in high-dimensional spaces. In Amos Storkey and Fernando Perez-Cruz, editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 745–754. PMLR, 09–11 Apr 2018.

[17] Andrew Gordon Wilson and Hannes Nickisch. Kernel interpolation for scalable structured gaussian processes (kiss-gp). In Francis Bach and David Blei, editors, *32nd International Conference on Machine Learning, ICML 2015*, 32nd International Conference on Machine Learning, ICML 2015, pages 1775–1784. International Machine Learning Society (IMLS), 2015. 32nd International Conference on Machine Learning, ICML 2015 ; Conference date: 06-07-2015 Through 11-07-2015.

[18] James Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Deisenroth. Efficiently sampling functions from Gaussian process posteriors. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 10292–10302. PMLR, 13–18 Jul 2020.

6 Приложения

6.1 Доказательство теоремы Маферона

Теорема 4 (Правило Маферона). Пусть $\mathbf{a}$ и $\mathbf{b}$ имеют совместное нормальное многомерное распределение. Тогда условное распределение на $\mathbf{a}$ при $\mathbf{b} = \beta$ вычисляется как

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

Доказательство.

Для удобства проведем переобозначения : $\Sigma_{11} = \operatorname{Cov}(\mathbf{a}, \mathbf{a})$, $\Sigma_{12} = \operatorname{Cov}(\mathbf{a}, \mathbf{b})$, Σ_{21} $\operatorname{Cov}(\mathbf{b}, \mathbf{a})$, $\Sigma_{22} = \operatorname{Cov}(\mathbf{b}, \mathbf{b})$. Пусть также $E\mathbf{a} = \boldsymbol{\mu}_1$, $E\mathbf{b} = \boldsymbol{\mu}_2$

Обозначим $\mathbf{z} = \mathbf{a} + \mathbf{ab}$, где $\mathbf{a} = -\Sigma_{12}\Sigma_{22}^{-1}$. Тогда

$$\begin{aligned} \operatorname{Cov}(\mathbf{z}, \mathbf{b}) &= \operatorname{Cov}(\mathbf{a}, \mathbf{b}) + \operatorname{Cov}(\mathbf{ab}, \mathbf{b}) = \Sigma_{12} + \mathbf{a} \operatorname{var}(\mathbf{b}) \\ &= \Sigma_{12} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22} = 0 \end{aligned}$$

Это значит, что $\mathbf{z}$ и $\mathbf{b}$ некореллированны и, вследствие их нормальности, независимы. Тогда $E\mathbf{z} = \boldsymbol{\mu}_1 + \mathbf{a}\boldsymbol{\mu}_2$. Из этого следует, что

$$\begin{aligned} E(\mathbf{a} \mid \mathbf{b}) &= E(\mathbf{z} - \mathbf{ab} \mid \mathbf{b}) \\ &= E(\mathbf{z} \mid \mathbf{b}) - E(\mathbf{ab} \mid \mathbf{b}) \\ &= E(\mathbf{z}) - \mathbf{ab} \\ &= \boldsymbol{\mu}_1 + \mathbf{a}(\boldsymbol{\mu}_2 - \mathbf{b}) \\ &= \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{b} - \boldsymbol{\mu}_2) \end{aligned}$$

Получается, что матожидания левой и правой части искомого тождества сов-

падают. Осталось рассмотреть ковариацию:

$$\begin{aligned}\operatorname{var}(\mathbf{a} \mid \mathbf{b}) &= \operatorname{var}(\mathbf{z} - \mathbf{ab} \mid \mathbf{b}) \\ &= \operatorname{var}(\mathbf{z} \mid \mathbf{b}) + \operatorname{var}(\mathbf{ab} \mid \mathbf{b}) - \mathbf{a}\operatorname{Cov}(\mathbf{b}, \mathbf{z}) - \operatorname{Cov}(\mathbf{z}, \mathbf{b})\mathbf{a}^\top \\ &= \operatorname{var}(\mathbf{z} \mid \mathbf{b}) \\ &= \operatorname{var}(\mathbf{z})\end{aligned}$$

Раскроем получившееся выражение:

$$\operatorname{var}(\mathbf{a} \mid \mathbf{b}) = \operatorname{var}(\mathbf{z}) = \operatorname{var}(\mathbf{a} + \mathbf{ab})$$

$$\begin{aligned}&= \operatorname{var}(\mathbf{a}) + \mathbf{a}\operatorname{var}(\mathbf{b})\mathbf{a}^\top + \mathbf{a}\operatorname{Cov}(\mathbf{b}, \mathbf{a}) + \operatorname{Cov}(\mathbf{a}, \mathbf{b})\mathbf{a}^\top \\ &= \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{22}\Sigma_{22}^{-1}\Sigma_{21} - 2\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\ &= \Sigma_{11} + \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} - 2\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} \\ &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\end{aligned}$$

Ковариации левой и правой части так же совпадают. В силу свойств нормального распределения отсюда следует искомое утверждение. ■