Метод быстрого сэмплирования из гауссовского процесса в задачах активного обучения

К. О. Вайсер[1], **М. Е. Панов**[2]

Аннотация

1 Постановка задачи

1.1 Активное обучение

Активное обучение является одним из подразделов машинного обучения [1]. Идея методов активного обучения заключается в выборе определенных точек для обучения вместо использования всех возможных данных, которые могут иметь в некотором смысле высокую стоимость. Например, человеческая разметка изображений для обучения модели распознавания образов требует, чтобы большое число изображения было размечено человеком. Если количество изображений исчисляется сотнями тысяч, то использоваие человеческого труда будет очень нецелесообразным. Если же у модели будет возможность выбирать изображения для разметки, которые в некотором смысле будут более полезны для нее, чем остальные, то число изображений, которые требуют разметки существенно снизится.

Пусть дана выборка

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

и оракул

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

Оракул — это функция, которая возвращает истинную метку объекта. Однако, в рассмотрении модели активного обучения, вызов оракула - дорогая операция, поэтому цель

[1]Московский физико-технический институт, vajser.ko@phystech.edu
[2]Сколковский Институт Науки и Технологий, m.panov@skoltech.ru

конструируемой модели - достигнуть требуемого качества, обратившись к оракулу как можно меньшее количество раз. Мы задаемся целью построить модель

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

которая аппроксимирует вероятность объекта $\mathbf{x}$ принадлежать к классу 1. При построении модели мы решаем задачу максимизации критерия качества AUC:

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

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

В задаче активного обучения немаловажным является способ выбора новых точек. Поскольку мы стремимся снизить число обращений к оракулу, мы должны выбирать точки, которые внесут наибольшее количество новой информации (!! рыхлая формулировка). Мы будем отбирать точки, основываясь на оценке неопределенности модели в них. Для оценки неопределенности будем использовать функцию эпистемической энтропии (!! почему именно такую):

$$\mathbb{E}H(p(x)) - H(\mathbb{E}p(x)),$$

где $H(p(x))$ — энтропия Штрассена в точке , которая дается выражением

$$H(p(x)) = -(p(x) \log p(x) + (1 - p(x)) \log 1 - p(x)$$

. Здесь значение функции вероятности $p(x)$ возвращает обученная модель $f(x)$. Обычно аналитическое вычисление матожидания энтропии невозможно (!! объяснить более подробно, почему). Для того, чтобы вычислить его численно, используется метод Монте-Карло (!! ссылку на метод):

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

Для использования этого метода необходимо сэмплировать большое количество реализаций случайной величины для снижения ошибки аппроксимации (!! сюда ссылку почему). Поэтому используемая модель должна обладать возможностью быстро сэмплировать на выбранном множестве точек.

1.2 Гауссовские процессы

В качестве модели f мы будем рассматривать гауссовский процесс.

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

Гауссовский процесс характеризуется функцией среднего и ковариационной функцией:

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

Гауссовский процесс задает тем самым распределение функций на своей области определения. Это свойство существенно увеличивает обобщающую способность модели (!! рыхлая формулировка), построенной на гауссовском процессе. Выход модели, основанной на гауссовском процессе, недетерминирован и зависит от реализации случайных величин (!! рыхлая формулировка). Это делает такие модели хорошим выбором в задачах, где необходимо использовать методы Монте-Карло.

Обучение гауссовского процесса означает вычисление апостериорного распределения на основе наблюдаемых обучающих данных и оптимизацию параметров ковариационной функции для минимизации функции потерь. Обозначим $\boldsymbol{f}_m \mid \boldsymbol{y}$ 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

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

Как видно, точное вычисление апостериорных моментов требует не менее, чем $O(n^3)$ операций, поскольку необходимо вычислить обратную матрицу размера $n \times n$. Для сэмплирования необходимо вычислить разложение Холецкого для ковариационной матрицы тестовой выборки, что так же требует $O(m^3)$ 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 сложность становится слишком большой.

2 Быстрое сэмплирование

Чтобы решить эту проблему, рассмотрим два подхода к рассмотрению гауссовских процессов [2].

2.1 Взвешенный подход

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

2.2 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 $\boldsymbol{Z} = \{z_j\}_{j=1}^v$. Пусть $u = f(\boldsymbol{Z}) \sim \mathbb{N}(\mu_u, \Sigma_u)$ Тогда $f_m \mid \boldsymbol{u}$ имеет следующие моменты

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

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*.

2.3 Правило Маферона

Как объединить эти два подхода, чтобы использовать их сильные стороны? [3]

Теорема 1 *Пусть* $\mathbf{w}$ *и* $\mathbf{b}$ *имеют совместное нормальное многомерное распределение. Тогда условное распределение на* $\mathbf{w}$ *при* $\mathbf{b} = \boldsymbol{\beta}$ *вычисляется как*

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

Эта теорема позволяет разложить апостериорное распределение в сумму двух слагаемых, одно из которых отражает априорное распределение, а другой отвечает поправк из-за несоответствия наблюдаемых данных априорному распределению. Применение теоремы Маферона к взвешенному и функциональному подходам позволяет получить следующий вид распределений:

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

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

3 Вычислительный эксперимент

Проведение вычислительного эксперимента преследует следующие цели:

1. Сравнение гауссовского процесса с другими классификаторами в терминах критерия качества AUC.
2. Исследование количества точек, требуемого для достижения качества, достигаемого на всей выборке.
3. Сравнение методов выбора точек
4. Сравнение скорости сэмплирования традиционного и быстрого метода.

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 вычисляет значение неопределенности для всех точек полной обучающей выборки и добавляет k точек с максимальным значением неопределенности в текущую обучающую выборку.

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.

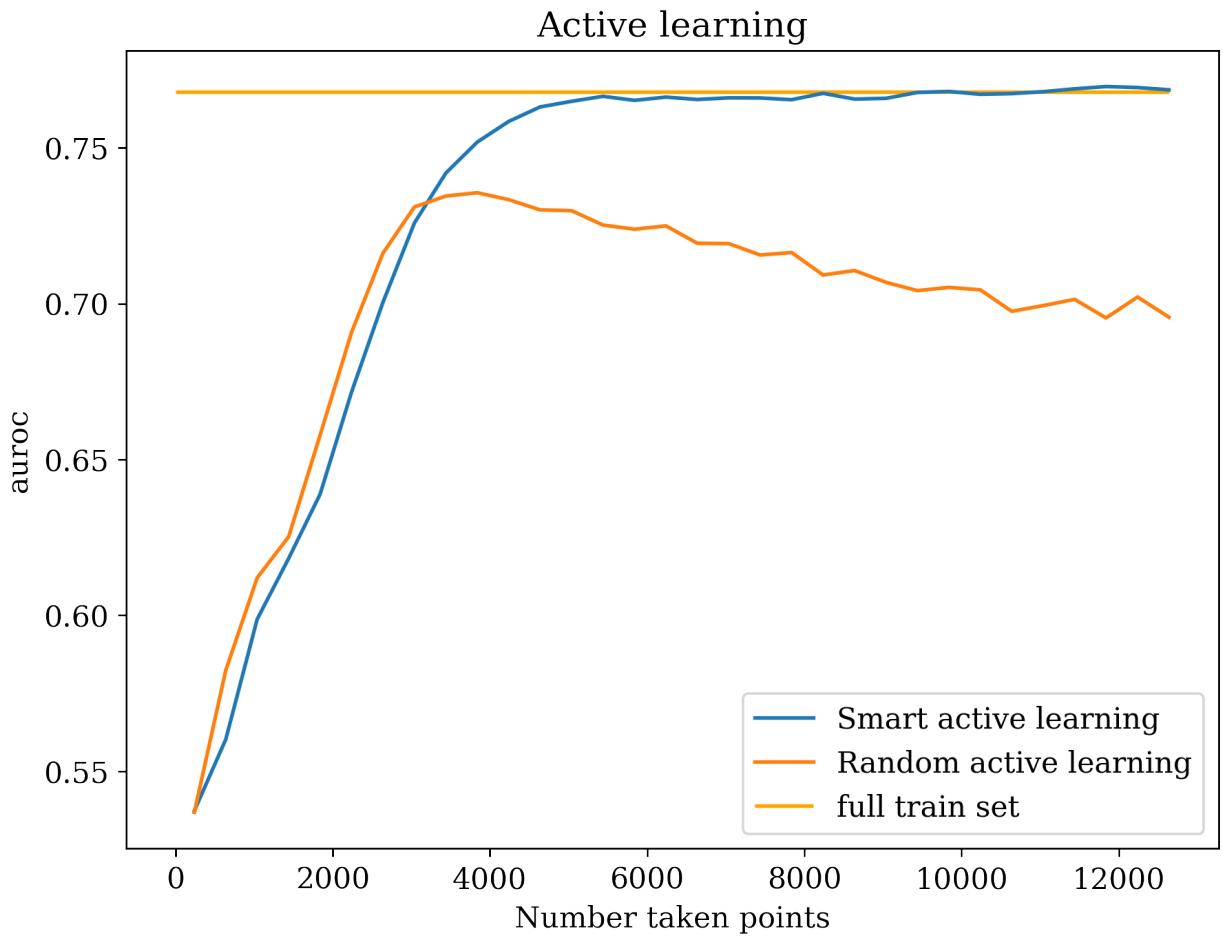

Рис. 1: Сравнение критерия AUC для трех методов обучения.

Полученные результаты демонстрируют, что использование значения неопределенности для выбора новых точек в активном обучение оправдало себя (!! рыхлая формулировка). Во-первых, 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. (!! рыхлая формулировка, интерпретация).

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

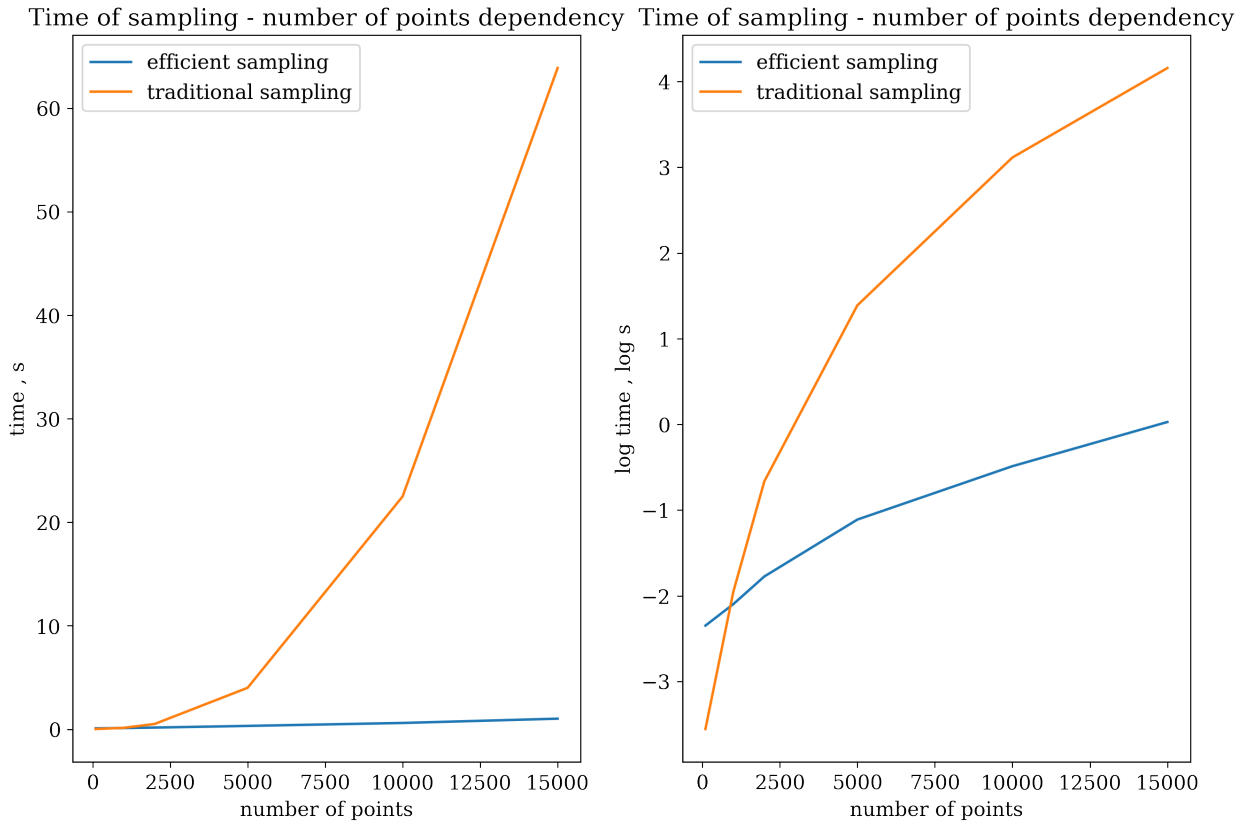

Рис. 2: Сравнение скорости сэмплинга.

Таблица 1: Описание выборок для экспериментов

Выборка $\mathfrak{D}$	Размер train	Размер test	Объекты	Признаки
dccc	24000	6000	30000	13

4 Заключение

Список литературы

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

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

[3] James T Wilson, Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Peter Deisenroth. Efficiently sampling functions from gaussian process posteriors.

arXiv preprint arXiv:2002.09309, 2020.