Вычислительно-эффективное сэмплирование из гауссовского процесса в задаче активного обучения

Вайсер Кирилл Олегович

Московский физико-технический институт

Группа 774

Научный руководитель: к.ф.-м.н. Панов М.Е.

Применение гауссовских процессов в активном обучении

Проблема

- Активное обучение предполагает недостаток исходных обучающих данных и постепенное пополнение обучающей выборки.
- Для пополнения выборки используются функции неопределенности, показывающие уверенность модели в своем ответе. Для их вычисления необходимо сэмплирование.
- Для сэмплирования необходимо обращать ковариационную матрицу и вычислять разложение Холецкого. Сложность этих операций — кубическая по длине выборки.

Решение

Использовать декомпозицию апостериорного распределения по правилу Маферона. Это позволит сэмплировать за линейное время по длине выборки.

Постановка задачи: выбор модели классификации

Задана выборка

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

и оракул $g : \mathbb{R}^d \mapsto \{-1, 1\}$ — функция, которая возвращает

правильные ответы.

Требуется найти модель, которая аппроксимирует вероятность принадлежности объекта $\boldsymbol{x}$ 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.

Функция стоимости $c : \{f, g\} \mapsto \mathbb{R}$ стоимости определяет затраты на вычисление функции в точке $\boldsymbol{x}$. В моделях активного обучения $c(g) \gg c(f)$

Функции неопределенности

Точки в выборку добавляются на основе значения функций неопределенности:

1) средняя энтропия в точке,

$$\mathrm{EH}\big(p(\boldsymbol{x})\big) = \mathrm{E} \sum_{y \in \{0,1\}} p_y(\boldsymbol{x}) \log p_y(\boldsymbol{x}),$$

где $p_y(\boldsymbol{x})$ — вероятность принадлежности объекта $\boldsymbol{x}$ к классу y

2) эпистемическая неопределенность в точке,

$$\mathrm{EH}\big(p(\boldsymbol{x})\big) - \mathrm{H}\big(\mathrm{E}p(\boldsymbol{x})\big),$$

3) максимальная вероятность класса

$$\max_y p_y(\boldsymbol{x}).$$

Апостериорное распределение гауссовского процесса

Обозначим $\boldsymbol{f}_m \mid \boldsymbol{y}$ — апостериорное распределение гауссовского процесса на тестовой выборке $\boldsymbol{X}_m$ после наблюдения обучающей выборки $(\boldsymbol{X}_n, \boldsymbol{y})$. Апостериорные моменты записываются как

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

где $\boldsymbol{\mu}_{m|n}$ — матожидание полученного вектора, а $\boldsymbol{K}$ — ковариационная матрица.
Сэмплирование производится с помощью схемы:

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

Такой подход требует $O\left(n^3\right)$ операций для вычисления $\left(\boldsymbol{K}_{n,n} + \sigma^2 \boldsymbol{I}\right)^{-1}$ и $O\left(m^3\right)$ операций для вычисления $\boldsymbol{K}^{1/2}_{m,m|n}$.

Взвешенная сумма спрямляющих функций

Байесовская линейная модель

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

где $\phi : \mathbb{R}^d \mapsto \mathbb{R}^\ell$, а $\boldsymbol{w}$ — параметры модели с априорным распределением $\mathcal{N}(0, \Sigma_p)$. Апостериорное распределение вектора ответов:

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

где $\Phi = \phi(\boldsymbol{X})$, $K = \Phi^\top \Sigma_p \Phi$. Это распределение соответствует гауссовскому процессу с ковариационной функцией

$$k(\boldsymbol{x}, \boldsymbol{x}) = \phi(\boldsymbol{x})^\top \Sigma_p \phi\left(\boldsymbol{x}'\right) = \psi(\boldsymbol{x})^\top \psi\left(\boldsymbol{x}'\right),$$

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

Разреженный гауссовский процесс

Для снижения сложности обучения гауссовского процесса возможен выбор подмножества точек обучающей выборки для вычисления апостериорных моментов. Пусть выбрано множество точек $\boldsymbol{Z} = \{z_j\}_{j=1}^{v}$ и $u = f(\boldsymbol{Z}) \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{u}}, \Sigma_u)$.

Тогда апостериорное распределение на тестовой выборке $f_m \mid \boldsymbol{u}$ имеет моменты

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

Сложность вычисления ковариационной матрицы снижается с $O\left(n^3\right)$ до $O\left(v^3\right), \quad n \gg v$.

Правило Маферона

Теорема (Вайсер, 2021, на основе работ Маферона, 1960)

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

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

Апостериорное распределение может быть представлено суммой априорного распредления и поправки на наблюдаемые данные.

Применение теоремы Маферона

Записав апостериорные распределения весов $\boldsymbol{w}$ из байесовской модели и вектора ответов $\boldsymbol{f}_m \mid \boldsymbol{u}$ разреженнего процесса в виде разложения по правилу Маферона, получим

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

Сэмплирование из априорного распределения $\boldsymbol{w}$ имеет сложность $O(\ell)$ и не зависит от размера выборки m. В то же время вычисление обновления существенно сложнее. В разреженном подходе напротив, сэмплирование из априорного распределения по-прежнему имеет сложность $O(m^3)$, однако поправка вычисляется за $O(m)$.

Эффективное сэмплирование из гауссовского процесса

Итоговая модель, совмещающая в себе априорное распределение байесовской модели и обновление разреженного процесса

$$(f \mid \boldsymbol{u})(\cdot) \overset{\mathrm{d}}{\approx} \boldsymbol{w}^\top \phi(\cdot) + \boldsymbol{h}^\top \boldsymbol{k}(\cdot, \boldsymbol{Z}),$$

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

Сложность сэмплирования из апостериорного распределения $O\left(v^3\right) + O(m)$.

Критерии качества модели классификации

1. AUROC score

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

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

2. Общая стоимость вызовов оракула

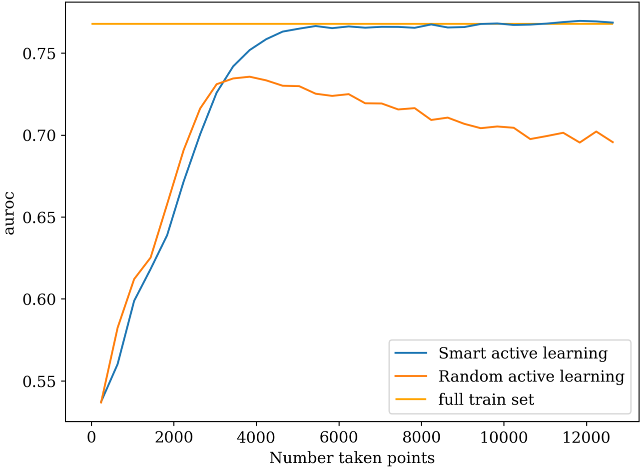
$$\sum_{i=1}^{E} n_i c(g),$$

   где E — число шагов активного обучения и n_i — число добавляемых в выборку точек на i-ом шаге.

Эксперимент

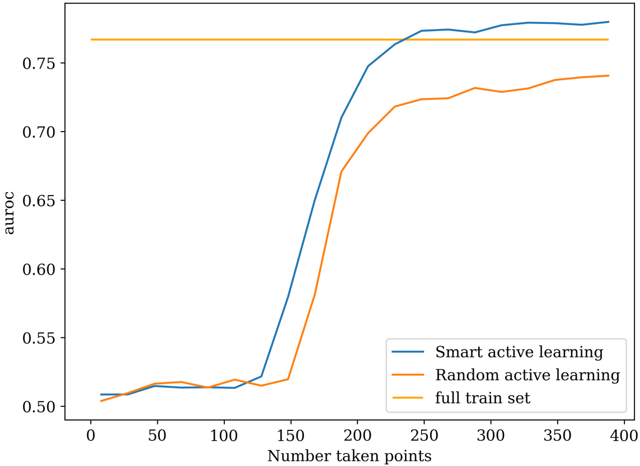
1. Обучение модели на всей обучающей выборке и определение метрики качества (AUROC).
2. Последовательное обучение на малой подвыборке с добавлением в выборку новых точек, отобранных по значению функции неопределенности.
3. Последовательное обучение на малой подвыборке с добавлением в выборку новых точек,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

Зависимость качества AUROC от числа точек в выборке



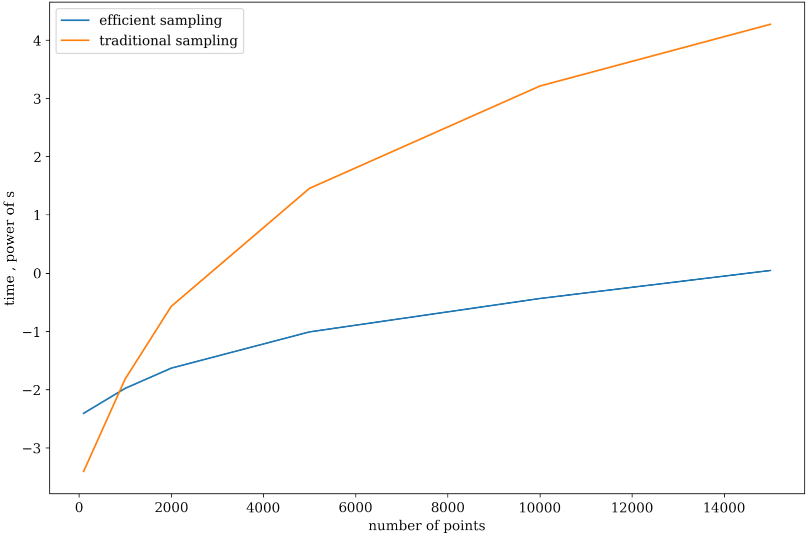
Добавление точек в выборку на основе функций неопределенности позволяет избежать переобучения и достичь уровня качества модели, обученной на всех доступных точках.

Зависимость качества AUROC от числа точек в выборке



Добавление точек в выборку на основе функций неопределенности позволяет превзойти качество, полученное на всей доступной выборке.

Сложность сэмплирования (логарифмическая шкала)



Выносится на защиту

1. Реализация метода эффективного сэмплирования через разложение апостериорного распределения в сумму априорного и поправки. Использование быстрого сэмплирования из априорного распределения байесовской модели и линейной сложности расчета поправки разреженного гауссовского процесса.

2. Применение метода эффективного сэмплирования к задачам активного обучения. Получение превосходящего результата в терминах заданного критерия качества при использовании меньшей обучающей выборки.

3. Разработка методики выбора точек для включения в обучающую выборку.

Публикации

1. Deep learning model structure optimization. Potanin M.S., Vayser K.O., Zholobov V.A., Strijov V.V, Informatics and Applications, 2020, 14(2): 58-65

2. Regularization schedule for neural architecture search Potanin M.S., Vayser K.O., Strijov V.V. Manuscript submitted for publication.