

Федеративное обучение и эволюция по Дарвину

С.В. Козырев

Дарвин предполагал (неправильно), что при генетической рекомбинации признаки (например, родителей) усредняются.

Такой подход связан с федеративным обучением.

Будет предъявлено уравнение дарвиновской эволюции с учетом генетической рекомбинации усреднением.

Теория обучения

Пусть задана обучающая выборка $\{z^l\}$, $l = 1, \dots, L$ (набор испытаний) и функция потерь $\mathcal{L}(z, x) \geq 0$ для испытания z и гипотезы x (пусть пространство гипотез есть $\mathbb{R}^d$).

Минимизация эмпирического риска есть задача минимизации по гипотезам суммы функции потерь по выборке

$$f(\{z\}, x) = \frac{1}{L} \sum_{l=1}^L \mathcal{L}(z^l, x) \rightarrow \min_x. \quad (1)$$

Градиентный спуск — численное решение дифференциального уравнения

$$\frac{dx}{dt} = -\nabla f(x),$$

траектория идёт в локальный минимум f .

При численной итерации спуска вектор x будет изменяться как

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k).$$

Стохастический градиентный спуск

1) Динамика Ланжевена стохастического градиента
(stochastic gradient Langevin dynamics, **SGLD**)

$$x_{k+1} = x_k + w_k - \alpha_k \nabla f(x_k),$$

w_k — независимые гауссовские случайные вектора.

Стохастическое дифференциальное уравнение (СДУ)

$$d\xi^i(t) = \sqrt{2\theta}dw^i(t) - \frac{\partial f(\xi(t))}{\partial x^i}dt, \quad (2)$$

$dw^i(t)$ стохастический дифференциал винеровского процесса

$$dw^i(t)dw^j(t) = \delta_{ij}dt.$$

G. Parisi, Correlation functions and computer simulations, Nuclear Physics B, 180(3), 378–384 (1981).

G. Parisi, Correlation functions and computer simulations II, Nuclear Physics B, 205 (3), 337–344 (1982).

2) **Mini-Batch stochastic gradient descent.** Вспомним, что функционал эмпирического риска есть сумма по выборке.

Для задачи минимизации (1) выберем случайно подмножество $\mathcal{I}$ (mini-batch) выборки и составим стохастический "пробатченный" градиент, где $f^i(x) = \mathcal{L}(z^i, x)$

$$\nabla f(x, \mathcal{I}) = \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} \nabla f^i(x). \quad (3)$$

Стохастический градиентный спуск есть процедура с итерацией вида (подмножество $\mathcal{I}_k$ выбирается случайно на каждом шаге)

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k, \mathcal{I}_k). \quad (4)$$

Случайный выбор подмножества $\mathcal{I}_k$ вносит шум в процедуру оптимизации.

Robbins, H.; Monro, S. A Stochastic Approximation Method. The Annals of Mathematical Statistics. 22 (3), 400 (1951).

Федеративное обучение (Federated Learning)

Рассмотрим теперь популяцию обучающихся агентов с параметрами x^i ; стартуют обучение с одного значения параметра x_0 ; совершают шаги стохастического градиентного спуска вида

$$x_{k+1}^i = x_k^i - \alpha_k^i \nabla f(x_k^i, z_k^i), \quad i = 1, \dots, m \quad (5)$$

z_k^i есть видная i -му агенту на k -м шаге часть выборки, $\nabla f(x_k^i, z_k^i)$ соответствующий стохастический градиент вида (3).

Коммуникация:

Раз в K шагов СГС обмениваются информацией: параметры каждого из агентов на K -м шаге устанавливаются как среднее

$$x_K'^i = \frac{1}{m} \sum_{i=1}^m x_K^i. \quad (6)$$

Потом процедура повторяется до следующей коммуникации.

Используется для распределённого обучения в интернете — информация о выборке может храниться на разных серверах, коммуникация есть обмен информацией о параметрах x_K^i .

J. Konecny, B. McMahan, D. Ramage, Federated optimization: Distributed optimization beyond the datacenter, arXiv:1511.03575 (2015).

C. A. Uribe, S. Lee, A. Gasnikov, A. Nedic, Optimal Algorithms for Distributed Optimization, arXiv:1712.00232 (2017).

Advances and Open Problems in Federated Learning, P. Kairouz, H. B. McMahan et al., in: Foundations and Trends in Machine Learning, 14 (1–2), (2021).

A Field Guide to Federated Optimization, J. Wang, Z. Charles et al., arXiv:2107.06917 (2021).

Популяция агентов есть аналог популяции особей в биологии, коммуникация для федеративного обучения — аналог генетической рекомбинации в популяционной генетике.

Альтернативные протоколы коммуникации: Дарвин предполагал усреднение параметров при рекомбинации, то есть попарную коммуникацию со случайным выбором агентов, и преобразование параметров

$$z = \frac{1}{2}(x + y). \quad (7)$$

В реальности в генетике рекомбинация идёт дискретно (законы Менделя).

Уравнение дарвиновской эволюции

Уравнение Фоккера–Планка для случайного блуждания SGLD

$$d\xi^i(t) = \sqrt{2\theta}dw^i(t) - \frac{\partial f(\xi(t))}{\partial x^i}dt.$$

Уравнение диффузии в потенциале

$$\frac{\partial u}{\partial t} = \theta \Delta u + \nabla u \cdot \nabla f + u \Delta f, \quad (8)$$

где $x \in \mathbb{R}^d$, $u = u(x, t)$ есть функция распределения, $f = f(x)$ есть потенциал, $f \in C^2(\mathbb{R}^d)$, $\theta > 0$ есть температура.

Эквивалентно,

$$\frac{\partial u}{\partial t} = \theta \operatorname{div} \left[e^{-\beta f} \operatorname{grad} \left[u e^{\beta f} \right] \right],$$

гиббсовское распределение $e^{-\beta f}$, $\beta = 1/\theta$ является стационарным решением уравнения, и решение сходится к гиббсовскому распределению (при некоторых условиях на f).

Добавим член, отвечающий рекомбинации — частицы с параметрами x и y заменяются на две частицы с параметрами $z = (x + y)/2$.

$$2 \int \delta(x + y - 2z) u(x) u(y) dx dy - \int u(z) u(x) dx.$$

Условие $z = (x + y)/2$ — признаки потомков есть среднее от признаков предков (по Дарвину).

Первый член описывает возникновение частицы с координатой $z = (x + y)/2$ из частиц с координатами x и y . Второй член описывает исчезновение частицы с координатой z за счёт рекомбинации с другими частицами. Число частиц сохраняется (интеграл от члена по z равен нулю).

Получаем уравнение, которое будет описывать федеративное обучение с коммуникацией (7) и дарвиновскую эволюцию, если её понимать буквально по Дарвину (при рекомбинации признаки потомков есть среднее признаков предков).

$$\frac{\partial u(z)}{\partial t} = \theta \Delta u + \nabla u \cdot \nabla f + u \Delta f + \quad (9)$$

$$+ \eta \left[2 \int \delta(x + y - 2z) u(x) u(y) dx dy - \int u(z) u(x) dx \right].$$

Здесь η есть скорость рекомбинации.

Второй член можно переписать как

$$\int u(x)(2u(2z - x) - u(z)) dx.$$

Дарвиновская дарвиновская эволюция (не только мутации с отбором, но и рекомбинация по Дарвину).

Кошмар Дженкина (swamping argument) — отличающиеся признаки будут усредняться, популяция будет стремиться к узкому распределению.