Машинное обучение: Методы оптимизации по Ю. Е. Нестерову

Кураленок И.Е.

Яндекс

18 апреля 2012 г.

Содержание

1. Общая постановка задачи
2. Области оптимизации
3. Локальные методы безусловной оптимизации

ML и оптимизация

$$F_0 = \underset{F}{\operatorname{argmax}}\ T(X|F)$$
$$F = F(x, \lambda)$$
$$\lambda \in \mathbb{B} \subset \mathbb{R}^n$$
$$\lambda_0 = \underset{\lambda \in \mathbb{B}}{\operatorname{argmax}}\ T(\lambda)$$

Кажется это задача нелинейной (в общем случае) оптимизации.

Постановка задачи оптимизации

$$\min f_0(x)$$
$$f_j(x)?0, j = 1 \ldots m$$
$$x \in S \subset \mathbb{R}^n$$

Вместо ? ставим $\geq$, $\leq$, или $=$. S — базовое допустимое множество.

$$Q = \{x | x \in S, f_j(x) \leq 0, j = 1 \ldots m\}$$

Q — допустимое множество

Типы задач оптимизации

условные $Q \subset \mathbb{R}^n$;

безусловные $Q \equiv \mathbb{R}^n$;

гладкие f_j — дифференцируемы;

негладкие $\exists k : f_k$ — не дифференцируема;

целочисленные $f_i = sin(\Pi x_i) = 0, i = 1 \ldots n$;

etc.

Алгоритм оптимизации

$$\mathbb{P} = (\Sigma, \Omega, \Upsilon)$$

Эффективность метода M на задаче $P \in \mathbb{P}$ – это необходимые вычислительные затраты для приближенного решения задачи с заданной точностью $\epsilon > 0$

Общая итеративная схема

Вход: $x_0, \epsilon > 0, k = 0, I_{-1} = \emptyset$

Основной цикл:

1. $\Omega(x_k)$;
2. $I_k = I_{k-1} \cup (x_k, \Omega(x_k))$;
3. применяем правила M к $x_{k+1} = M(I_k)$;
4. проверяем $\Upsilon(x_(k+1), \epsilon)$, если не выполнено $k++$ и к 1;

Сложность метода

Аналитическая сложность: число обращений к оракулу

Арифметическая сложность: общее число вычислений

Виды оракулов

Нулевого порядка: $f(x_k)$ – сэмплирование, генетика, random walk, etc.

Первого порядка: $(f(x_k), f'(x_k))$ – градиентный спуск, TWIST, FISTA, Fobos, etc.

Второго порядка: $(f(x_k), f'(x_k), f''(x_k))$ – Ньютона.

Оценки сложности задач глобальной оптимизации

$$f^* = \min x \in \mathbb{B}^{\ltimes} f(x)$$
$$\mathbb{B}^{\ltimes} = \{x | x \in \mathbb{R}^n, x_i \in [0, 1], i = 1 \dots n\}$$
$$f \in C^{0,0}_{L_\infty}$$

Будем решать равномерными сетками.

$$f(\overline{x}) - f^* \leq \frac{L}{2p}$$

$\Rightarrow$ $A \leq \left(\left[\frac{L}{2\epsilon}\right] + 2\right)^n$

$\Rightarrow$ применим сопротивляющийся оракул $A \geq \left(\left[\frac{L}{2\epsilon}\right]\right)^n$

Пример

Пусть $L = 2$, $n = 10$, $\epsilon = 0.01$
Пусть мы умеем вычислять f за 100ms

- $\Rightarrow A \geq \left(\left[\frac{L}{2\epsilon}\right]\right)^n = 10^20$
- $\Rightarrow$ процесс сойдется за 10^19c » 10^11 лет (np-полные/трудные задачи нервно курят!).
- $\Rightarrow$ наверное нужно искать более эффективные методы/ограничения на задачи.

Общая глобальная оптимизация

Цель: найти глобальный экстремум

Класс функций: непрерывные

Оракул: 0-2

Чего хотим: хоть какая-то сходимость к глобальному решению

Особенности: в теории — не работает. Этих кошек надо учиться готовить в каждом конкретном случае.

Общая нелинейная оптимизация

Цель: найти локальный минимум

Класс функций: дифференцируемые

Оракул: 1-2

Чего хотим: *быстро* бежать до локального минимума

Особенности: множество решений, не всегда работает на больших размерностях

Выпуклая оптимизация

Цель: найти глобальный минимум

Класс функций: выпуклые функции

Оракул: 1

Чего хотим: приемлемую скорость сходимости к минимуму

Особенности: ограниченная размерность, тяжело привести задачу к выпуклой

Полиномиальные методы внутренней точки

Цель: найти глобальный минимум

Класс функций: выпуклые множества Q, функции с явно заданной структурой

Оракул: 2

Чего хотим: быструю сходимость к глобальному минимум, скорость может зависить от структуры

Особенности: практически неограниченная размерность, задача рассматривается как белый ящик, над target функцией надо работать

Локальная аппроксимация целевой функции Тейлором

Будем надеяться что в окрестности x_k функция хорошо аппроксимируется:

Линейно: $f(y) = f(x) + f'(x) \cdot (y - x) + o(\|y - x\|)$

Квадратично: $f(y) = f(x) + f'(x)^T (y - x) + \frac{1}{2} \left(f''(x)(y - x)\right)^T (y - x) + o(\|y - x\|^2)$

Если $f \in C_L^{k,p}$ то можно делать какие-то выводы о сходимости.

Градиентный метод

$$x_0 \in \mathbb{R}^n$$
$$x_{k+1} = x_k - h_k f'(x_k), k = 0, \ldots$$

Градиентный метод

Есть много методов выбора шага x_k:

- Последовательность не зависит от истории $\{h_k\}_0^\infty$:

$$h_k = h > 0$$
$$h_k = \frac{h}{log(k+2)}$$

- Полная релаксация (градиентный спуск):

$$h_k = \underset{h \geq 0}{\operatorname{argmin}} f\left(x_k - hf'(x_k)\right)$$

- Правило Голдштейна-Армийо: зафиксируем α, β и найдем x_{k+1}

$$\alpha f'^T(x_k - x_{k+1}) \leq f(x_k) - f(x_{k+1})$$
$$\beta f'^T(x_k - x_{k+1}) \geq f(x_k) - f(x_{k+1})$$

Эффективность градиентного метода

$$f(x_k) - f(x_{k+1}) \geq h\left(1 - \frac{1}{2}Lh\right)\|f'(x_k)\|^2$$
$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2L}\|f'(x_k)\|^2$$
$$f(x_k) - f(x_{k+1}) \geq \frac{2}{L}\alpha(1-\beta)f'(x_k)\|^2$$

$$f(x_k) - f(x_{k+1}) \geq \frac{\omega}{L}f'(x_k)\|^2$$

$$\Rightarrow g_N^* \leq \frac{1}{\sqrt{N+1}}\left(\frac{1}{\omega}L(f(x_0) - f^*)\right)^{\frac{1}{2}}$$

В общем случае градиентные методы умеет сходиться не к минимуму, а к стационарной точке!

Метод Ньютона

$$x_{k+1} = x_k - \left(f''(x_k)\right)^{-1} f'(x_k)$$

Работает только близко от точки минимума (попробуйте $f(x) = \frac{x}{\sqrt{1+x^2}}$, при начальной точке $x : |x| > 1$)!
Демпфированный метод Ньютона:

$$x_{k+1} = x_k - h_k \left(f''(x_k)\right)^{-1} f'(x_k)$$

Ньютон сходится квадратично $f \in C_M^{2,2}, f''(x^*) \succcurlyeq l\ E, l > 0$:

$$\|x_k - x^*\| \leq \frac{M\|x_k - x*\|^2}{2(l - M\|x_k - x^*\|)}$$

Методы переменной метрики

$$x_{k+1} = x_k - H_k f'(x_k)$$
$$\lim_{k \to \infty} H_k = f''^{-1}$$
$$H_{k+1}(f'(x_{k+1}) - f'(x_k)) = x_{k+1} - x_k$$

Методы переменной метрики

Есть много способов удовлетворить квазиньютоновское правило:

$$\triangle H_k = H_{k+1} - H_k, \delta_k = x_{k+1} - x_k, \gamma_k = f'(x_k + 1) - f'(x_k)$$

- Правило одноранговой коррекции:

$$\triangle H_k = \frac{(\delta_k - H_k\gamma_k)(\delta_k - H_k\gamma_k)^T}{(\delta_k - H_k\gamma_k)^T\gamma_k}$$

- Правило Давидона-Флетчера-Пауэла (ДФП):

$$\triangle H_k = \frac{\delta_k\delta_k^T}{\gamma_k^T\delta_k} - \frac{H_k\gamma_k\gamma_k^T H_k}{(H_k\gamma_k)^T\gamma_k}$$

Для квадратичных функций не более n итераций.

Методы переменной метрики

Есть много способов удовлетворить квазиньютоновское правило:

$$\triangle H_k = H_{k+1} - H_k, \delta_k = x_{k+1} - x_k, \gamma_k = f'(x_k + 1) - f'(x_k)$$

- Правило Бройдена-Флетчера-Гольдфарба-Шенно (БФГД):

$$\triangle H_k = \frac{H_k \gamma_k \delta_k^T + \delta_k \gamma_k^T H_k}{(H_k \gamma_k)^T \gamma_k} - \beta \frac{H_k \gamma_k \gamma_k^T H_k}{(H_k \gamma_k)^T \gamma_k}$$
$$\beta = 1 + \frac{\gamma^T \delta_k}{(H_k \gamma_k)^T \gamma_k}$$

Для квадратичных функций не более n итераций.