Машинное обучение: Выпуклая оптимизация по Ю. Е. Нестерову

Кураленок И.Е.

Яндекс

12 апреля 2012 г.

Содержание

ML и выпуклая оптимизация

Часто в ML мы можем поставить задачу в терминах выпуклых функций.

$$\|Ax - b\|^2$$
$$\|Ax - b\|^2 + \|x\|_1$$
$$\log \frac{1}{1+e^{-t}}$$
$$\log \frac{1}{1+e^{c-t}}$$

Пример сведения к выпуклой оптимизации

Хотим элемент классификатора h из деревьев решений:

$$\operatorname*{argmin}_{h \in T} \sum_{i=1}^{m} \log \frac{1}{1 + e^{(c - \alpha h(x_i)) y_i}}$$

Деревья — зло, но как их подбирать в случае MSE мы знаем.
Разобьем задачку на 2 части:

$$\operatorname*{argmin}_{\phi \in \mathbb{R}} \sum_{i=1}^{m} \log \frac{1}{1+e^{(c-\alpha\phi_i)*y_i}}$$
$$\operatorname*{argmin}_{h \in T} \sum_{i=1}^{m} (h(x_i) - \phi_i)^2$$

Первая часть выпуклая(на самом деле вогнутая) в $\mathbb{R}^m$, вторая – понятная.
$\Rightarrow$ важно научиться быстро и эффективно решать выпуклые задачки.

Выпуклые функции

Есть много определений $\mathcal{F}^{1,1}$, вот некоторые из них:

$$
\begin{aligned}
&f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha) f(y) && ,\forall x, y \in \mathbb{R}^n, \alpha \in [0,1] \\
&f(y) \geq f(x) + (f'(x))^T (y - x) && ,\forall x, y \in \mathbb{R}^n \\
&(f'(x) - f'(y))^T (x - y) \geq 0 && ,\forall x, y \in \mathbb{R}^n
\end{aligned}
$$

Важнейшее для нас свойство: $f'(x) = 0 \Leftrightarrow x$ — глобальный максимум! Часто хотим чуть большего: $\mathcal{F}_L^{1,1}$, более того, сильной выпуклости $\mathcal{G}_{\mu,L}^{1,1}$:

$$
f(y) \geq f(x) + (f'(x))^T (y - x) + \frac{1}{2}\mu \|x - y\|^2, \quad \forall x, y \in \mathbb{R}^n, \mu \geq 0
$$

Класс методов

Хотим искать решения в классе

Модель: $\underset{x \in \mathbb{R}^n}{\operatorname{argmin}}, f \in \mathcal{F}_L^{1,1}$.

Оракул: локальный черный ящик первого порядка

Решение: $\bar{x} \in \mathbb{R}^n : f(\bar{x}) - f^* \leq \epsilon$

В этом классе решения будут не лучше чем:

$$f(x_k) - f^* \geq \frac{3L\|x_0 - x^*\|^2}{32(k+1)^2}$$

при числе шагов $k < \frac{1}{2}(n-1)$.

Градиентный метод

На классе $\mathcal{F}_L^{1,1}$ и $\mathcal{G}_{\mu,L}^{1,1}$ градиентный метод с фиксированным шагом $h_k = h = \frac{1}{L}$ сходится.

$$f(x_k) - f^* \geq \frac{2L\|x_0 - x^*\|^2}{k+4}$$
$$f(x_k) - f^* \geq \frac{L}{2}\left(\frac{L-\mu}{L+\mu}\right)^{2k}\|x_0 - x^*\|^2$$

Скорость сходимости — линейна на $\mathcal{F}_L^{1,1}$, а далеко не квадратична.

Утверждается, что релаксацией невозможно получить оптимальный метод первого порядка!

Оценивающие последовательности

Последовательности $\{\varphi_k(x)\}_{k=0}^{\infty}$ и $\{\lambda_k\}_{k=0}^{\infty}, \lambda_k \geq 0$ — *оценивающие*, если $\lambda_k \to 0$ и $\forall x \in \mathbb{R}^n, k \geq 0$:

$$\varphi_k(x) \leq (1 - \lambda_k) f(x) + \lambda_k \varphi_0(x)$$

Введенные последовательности хороши этим:

$$f(x_k) - f^* \leq \lambda_k (\varphi_0 - f^*) \to 0$$

Построение оценивающей последовательности

Пусть:

1. $f \in \mathcal{G}^{1,1}_{\mu,L}(\mathbb{R}^n)$,
2. $\varphi_0(x)$ — произвольная функция на $\mathbb{R}^n$,
3. $\{y_k\}_{k=0}^{\infty}$ — произвольная последовательность в $\mathbb{R}^n$,
4. $\{a_k\}_{k=0}^{\infty} : a_k \in (0,1), \sum_k a_k = \infty$,
5. $\lambda_0 = 1$,

тогда последовательности $\{\varphi_k(x)\}_{k=0}^{\infty}$ и $\{\lambda_k\}_{k=0}^{\infty}$:

$$\lambda_{k+1} = (1 - a_k)\lambda_k,$$
$$\varphi_{k+1}(x) = (1 - a_k)\varphi_k(x) + a_k \left(f(y_k) + (f'(y_k))^T (x - y_k) + \frac{\mu}{2}\|x - y_k\|$$

оценивающие, а если $\varphi_0(x) = \varphi_0^* + \frac{\gamma_0}{2}\|x - v_0\|^2$ то и все $\varphi_0(x)$ имеют такой же вид.

Общая схема оптимального метода

1. Выберем $x_0 \in \mathbb{R}^n$ и $\gamma_0 > 0$, возьмем $v_0 = x_0$.
2. k-я итерация:
   1. $a_k \in (0,1) : La_k^2 = (1-a_k)\gamma_k + a_k\mu$, положим $\gamma_{k+1} = (1-a_k)\gamma_k + a_k\mu$
   2. Выберем
   $$y_k = \frac{a_k\gamma_k v_k + \gamma_{k+1}x_k}{\gamma_k + a_k\mu},$$
   вычислим $f(y_k)$ и $f'(y_k)$
   3. найдем $x_{k+1} : f(x_{k+1}) \leq f(y_{k+1}) - \frac{1}{2L}\|f'(y_k)\|^2$
   4. $$v_{k+1} = \frac{(1-a_k)\gamma_k v_k + a_k\mu y_k - a_k f'(y_k)}{\gamma_{k+1}}$$

Скорость сходимости оптимального метода

Если все делать по схеме, то:

$$f(x_k) - f^* \leq Lmin\left\{\left(1 - \sqrt{\frac{\mu}{L}}\right)^k, \frac{4}{(k+2)^2}\right\}\|x_0 - x*\|^2$$

Iterative Shrinkage/Threshoulding Algorithm (ISTA)

Ограничемся задачей

$$\underset{x\in\mathbb{R}^n}{\operatorname{argmin}}\|Ax-b\|+R(x)$$

В основном интересен случай, когда $R(x)=\lambda\|x\|_1$ (l_1 регуляризация) или $R(x)=\lambda\|Lx\|_2$ (регуляризация Тихонова). Заметим, что задача в случае l_1 негладкая.

$$x_{k+1}=\tau_{\lambda t_k}\left(x_k-2t_kA^T(Ax_k-b)\right)$$
$$\tau_\alpha^i=(|x_i|-\alpha)_+sign(x_i)$$

Такая штука, являясь по сути градиентным методом сходится со скоростью:

$$f(x_k)-f^*\leq\alpha\frac{L\|x_k-x^*\|^2}{2k}$$

Если шаг $t_k=t$, то $\alpha=1$, если шаг брать "умнее", то $\alpha\leq 1$.

Fast Iterative Shrinkage/Threshoulding Algorithm (FISTA)

Объединим ISTA и Нестерова.

1. Выберем $x_0 \in \mathbb{R}^n$ и $\gamma_0 = 1$, возьмем $y_0 = x_0$.
2. k-я итерация:
   1. $x_k = \operatorname{argmin}_x \left(R(x) + \left\| x - \left(y_{k-1} - \frac{1}{L} f'(y_{k-1}) \right) \right\|^2 \right)$
   2. $\gamma_{k+1} = \frac{1+\sqrt{1+4\gamma_k^2}}{2}$
   3. $y_k = x_k + \left(\frac{\gamma_k - 1}{\gamma_{k+1}} \right) (x_k - x_{k-1})$

Это добро сходится:

$$f(x_k) - f^* \leq \frac{\alpha L}{(k+1)^2} \|x_0 - x^*\|^2$$

Что еще бывает?

Область богата на исследования и приложения. Стоит почитать: Нестерова (strong recommend), FOBOS (google), TWIST, etc.