

Машинное обучение. Stochastic gradient descent

Алексей Колесов

Белорусский государственный университет

31 октября 2018 г.

План

- вместо $L_S(h)$ будем пытаться оптимизировать $L_D(h)$
- итеративно: каждый шаг в направлении уменьшения функции
- эффективно для выпуклых задач

Содержание

- 1 Gradient descent
 - Идея алгоритма
 - GD для выпукло-липшицевых функций
- 2 Субдифференциал
- 3 Stochastic gradient descent
 - Общая идея
 - Варианты SGD
 - SGD для задач машинного обучения

Общий план алгоритма

- для дифференцируемой $f : \mathbb{R}^d \rightarrow \mathbb{R}$ **градиентом** называется $\nabla f(w) = \left(\frac{\partial f(w)}{\partial w_1}, \dots, \frac{\partial f(w)}{\partial w_d} \right)$
- итеративный алгоритм:
 - $w^{(1)} = 0$
 - для $\eta > 0$, $w^{(t+1)} = w^{(t)} - \eta \nabla f(w^{(t)})$ (update rule)
 - результат: $\bar{w} = \frac{1}{T} \sum_{t=1}^T w^{(t)}$
- для некоторых классов выпуклых функций доказуемо хорошо работает

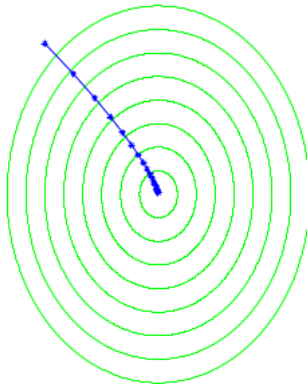
Мотивация update rule

$$w^{(t+1)} = w^{(t)} - \eta \nabla f(w^{(t)})$$

- $-\nabla f(w)$ — направление наибольшего убывания функции в точке w
- $f(u) \approx f(w) + \langle u - w, \nabla f(w) \rangle$, для выпуклых
 $f(u) \geq f(w) + \langle u - w, \nabla f(w) \rangle$
- хотим минимизировать $f(u)$, но не уходить далеко:

$$w^{(t+1)} = \operatorname{argmin}_w \frac{1}{2} \|w - w^{(t)}\|^2 + \eta (f(w^{(t)}) + \langle w - w^{(t)}, \nabla f(w^{(t)}) \rangle)$$

Пример GD



Предположения

- пусть функция f выпуклая и липшицева
- зафиксируем w^* , такой что $\|w^*\| \leq B$
- оценим $f(\bar{w}) - f(w^*)$

Оценка оптимальности

$$f(\bar{w}) - f(w^*) = f\left(\frac{1}{T} \sum_{t=1}^T w^{(t)}\right) - f(w^*) \quad (1)$$

$$\leq \frac{1}{T} \sum_{t=1}^T f(w^{(t)}) - f(w^*) \quad (2)$$

$$= \frac{1}{T} \sum_{t=1}^T \left(f(w^{(t)}) - f(w^*) \right) \quad (3)$$

$$\leq \frac{1}{T} \sum_{t=1}^T \langle w^{(t)} - w^*, \nabla f(w^{(t)}) \rangle \quad (4)$$

Лемма о итеративном процессе

Лемма о итеративном процессе

Пусть v_1, \dots, v_T произвольные векторы. Пусть $w^{(1)} = 0$ и

$$w^{(t+1)} = w^{(t)} - \eta v_t$$

Тогда:

$$\sum_{t=1}^T \langle w^{(t)} - w^*, v_t \rangle \leq \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2$$

Если $\|v_t\| \leq \rho$, $\|w^*\| \leq B$, $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, то

$$\frac{1}{T} \sum_{t=1}^T \langle w^{(t)} - w^*, v_t \rangle \leq \frac{B\rho}{\sqrt{T}}$$

Доказательство леммы

Хотим доказать:

$$\sum_{t=1}^T \langle w^{(t)} - w^*, v_t \rangle \leq \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2$$

Рассмотрим одно слагаемое:

$$\langle w^{(t)} - w^*, v_t \rangle = \frac{1}{\eta} \langle w^{(t)} - w^*, \eta v_t \rangle \quad (5)$$

$$= \frac{1}{2\eta} \left(-\|w^{(t)} - w^* - \eta v_t\|^2 + \|w^{(t)} - w^*\|^2 + \eta^2 \|v_t\|^2 \right) \quad (6)$$

$$= \frac{1}{2\eta} \left(-\|w^{(t+1)} - w^*\|^2 + \|w^{(t)} - w^*\|^2 \right) + \frac{\eta}{2} \|v_t\|^2 \quad (7)$$

Доказательство леммы

Хотим доказать:

$$\sum_{t=1}^T \langle w^{(t)} - w^*, v_t \rangle \leq \frac{\|w^*\|^2}{2\eta} + \frac{\eta}{2} \sum_{t=1}^T \|v_t\|^2$$

Распишем:

$$\sum_{t=1}^T \left(-\|w^{(t+1)} - w^*\|^2 + \|w^{(t)} - w^*\|^2 \right) \quad (8)$$

$$= \|w^{(1)} - w^*\|^2 - \|w^{(T+1)} - w^*\|^2 \quad (9)$$

$$\leq \|w^{(1)} - w^*\|^2 \quad (10)$$

$$= \|w^*\|^2 \quad (11)$$

Следствие для выпукло-ограниченно-липшицевых функций

Следствие для выпукло-ограниченно-липшицевых функций

Пусть f выпуклая, ρ -липшицева функция и $w^* \in \operatorname{argmin}_{\{w: \|w\| \leq B\}} f(w)$. Тогда если запустить T итераций GD с $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, то

$$f(\bar{w}) - f(w^*) \leq \frac{B\rho}{\sqrt{T}}$$

Чтобы получить, что $f(\bar{w}) - f(w^*) \leq \epsilon$, для $\epsilon > 0$ нужно провести T итераций, где

$$T \geq \frac{B^2 \rho^2}{\epsilon^2}$$

Содержание

- 1 Gradient descent
 - Идея алгоритма
 - GD для выпукло-липшицевых функций
- 2 Субдифференциал
- 3 Stochastic gradient descent
 - Общая идея
 - Варианты SGD
 - SGD для задач машинного обучения

Мотивация

- GD применим только для дифференцируемых функций
- попробуем ослабить требование
- необходимо наличие «опорных плоскостей»

Лемма о выпуклости

Лемма о выпуклости

Пусть S — открытое выпуклое множество. Тогда функция $f : S \rightarrow \mathbb{R}$ выпукла тогда и только тогда, когда $\forall w \in S \exists v$, такой что

$$\forall u \in S, f(u) \geq f(w) + \langle u - w, v \rangle$$

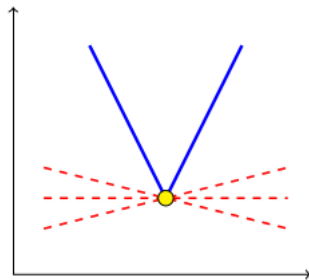
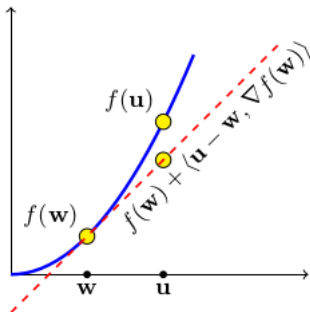
Определение

Вектор v , который удовлетворяет

$$\forall u \in S, f(u) \geq f(w) + \langle u - w, v \rangle$$

называется **субградиентом** (subgradient) функции f в точке w .
Множество всех субградиентов f в w называется **субдифференциалом** и обозначается $\partial f(w)$

Пример



Вычисление субградиентов

Лемма о субградиенте дифференцируемой функции

Если f дифференцируема в w , то $\partial f(w) = \{\nabla f(w)\}$

Лемма о субградиенте максимума выпуклых функций

Пусть $g(w) = \max_{i \in [r]} g_i(w)$, где g_i — выпуклая дифференцируемая функция. Зафиксируем w и найдём $j \in \operatorname{argmax}_i g_i(w)$. Тогда $\nabla g_j(w) \in \partial g(w)$

Пример

Пусть $f(w) = \max\{0, 1 - y\langle w, x \rangle\}$ — hinge loss. Тогда для любого w можно найти $v \in \partial f$:

$$v = \begin{cases} 0 & \text{если } 1 - y\langle w, x \rangle \leq 0 \\ -yx & \text{если } 1 - y\langle w, x \rangle > 0 \end{cases}$$

Субградиент липшицевых функций

Лемма о субградиенте липшицевых функций

Пусть A выпуклое открытое множество и $f : A \rightarrow \mathbb{R}$ выпуклая функция. Тогда, f является ρ -липшицевой тогда и только тогда, когда $\forall w \in A, \forall v \in \partial f(w)$ выполняется $\|v\| \leq \rho$

Итоги

- субградиент обобщает градиент
- можно использовать в GD (заменив $\nabla f(w)$ на субградиент)

Содержание

- 1 Gradient descent
 - Идея алгоритма
 - GD для выпукло-липшицевых функций
- 2 Субдифференциал
- 3 Stochastic gradient descent
 - Общая идея
 - Варианты SGD
 - SGD для задач машинного обучения

План

- в GD необходимо знать градиент в каждой точке
- будем выбирать случайный вектор, который в среднем будет равен градиенту
- выведем простой способ для задачи МО

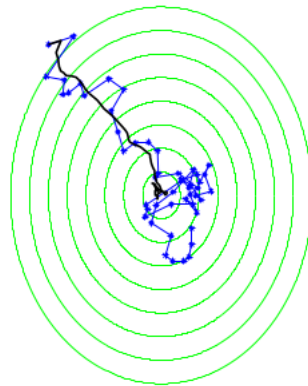
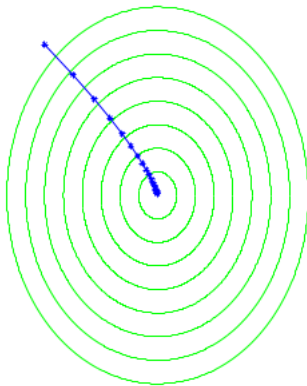
Алгоритм

Алгоритм 1 Stochastic gradient descent для минимизации $f(w)$

Вход: $\eta > 0$, $T > 0$

- 1: $w^{(1)} = 0$
 - 2: **for** $t = 1, \dots, T$ **do**
 - 3: v_t — случайный вектор, т.ч. $\mathbb{E}[v_t | w^{(t)}] \in \partial f(w^{(t)})$
 - 4: $w^{(t+1)} = w^{(t)} - \eta v_t$
 - 5: **end for**
 - 6: **return** $\bar{w} = \sum_{t=1}^T w^{(t)}$
-

Пример



SGD для выпукло-липшицево-ограниченных функций

SGD для выпукло-липшицево-ограниченных функций

Пусть $B > 0$, $\rho > 0$. Пусть f выпуклая функция и $w^* \in \operatorname{argmin}_{\{w: \|w\| \leq B\}} f(w)$. Пусть SGD был запущен на T итерациях с $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$. Кроме того, пусть $\mathbb{P}[\|v_t\| \leq \rho] = 1$. Тогда:

$$\mathbb{E}[f(\bar{w})] - f(w^*) \leq \frac{B\rho}{\sqrt{T}}$$

Проекционный шаг

- в анализе GD и SGD мы требовали, чтоб $\|w\| \leq B$
- после update rule нет гарантий на $\|w\|$
- можно добавить проекцию $w^{(t)}$ на допустимое множество

Алгоритм

- 1 $w^{(t+\frac{1}{2})} = w^{(t)} - \eta v_t$
- 2 $w^{(t+1)} = \operatorname{argmin}_{w \in H} \|w - w^{(t+\frac{1}{2})}\|$

Лемма о проекции

Пусть H замкнутое выпуклое множество и v проекция w на H , т.е. $v = \operatorname{argmin}_{x \in H} \|x - w\|^2$. Тогда, для любого $u \in H$:

$$\|w - u\|^2 \geq \|v - u\|^2$$

Доказательство леммы

Хотим:

$$\|w - u\|^2 \geq \|v - u\|^2$$

Так как H выпукло, то $\forall \alpha \in (0, 1) \ v + \alpha(u - v) \in H$.

Имеем:

$$\|v - w\|^2 \leq \|v + \alpha(u - v) - w\|^2 \quad (12)$$

$$= \|v - w\|^2 + 2\alpha \langle v - w, u - v \rangle + \alpha^2 \|u - v\|^2 \quad (13)$$

Значит:

$$\begin{aligned} 2\langle v - w, u - v \rangle &\geq -\alpha \|u - v\|^2 \\ \langle v - w, u - v \rangle &\geq 0 \end{aligned}$$

Доказательство

Хотим:

$$\|w - u\|^2 \geq \|v - u\|^2$$

Имеем:

$$\langle v - w, u - v \rangle \geq 0$$

Распишем:

$$\|w - u\|^2 = \|w - v + v - u\|^2 \tag{14}$$

$$= \|w - v\|^2 + \|v - u\|^2 + 2\langle v - w, u - v \rangle \tag{15}$$

$$\geq \|v - u\|^2 \tag{16}$$

Справедливость анализа для проекционной версии

$$\|w^{(t+1)} - w^*\|^2 - \|w^{(t)} - w^*\|^2 \quad (17)$$

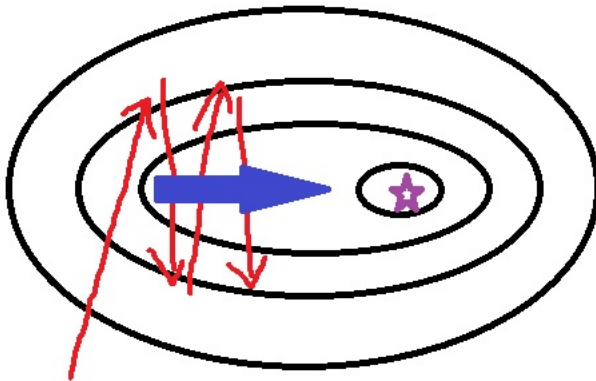
$$= \|w^{(t+1)} - w^*\|^2 - \|w^{(t+\frac{1}{2})} - w^*\|^2 + \|w^{(t+\frac{1}{2})} - w^*\|^2 - \|w^{(t)} - w^*\|^2 \quad (18)$$

$$\leq \|w^{(t+\frac{1}{2})} - w^*\|^2 - \|w^{(t)} - w^*\|^2 \quad (19)$$

Переменный размер шага

- чем ближе к оптимуму, тем меньше «должен» быть шаг η
- шаг η можно выбирать отдельно для каждой координаты
- множество других вариаций

SGD with momentum



Выбор ответа

- последний $w^{(T)}$
- $\sum_{t=\lceil \alpha T \rceil}^T w^{(t)}$
- лучший по эмпирическому риску

SGD для сильно-выпуклых функций

Сильная выпуклость

Функция f называется λ -сильно выпуклой, если для всех u, w и $\alpha \in [0, 1]$ выполняется

$$f(\alpha w + (1 - \alpha)u) \leq \alpha f(w) + (1 - \alpha)f(u) - \frac{\lambda}{2}\alpha(1 - \alpha)\|w - u\|^2$$

Для λ -сильно выпуклых функций выполняется, что $\forall w, u$ и $\forall v \in \partial f(w)$ выполняется:

$$\langle w - u, v \rangle \geq f(w) - f(v) + \lambda\|w - u\|^2$$

SGD для сильно-выпуклых функций

Алгоритм 2 Stochastic gradient descent для минимизации λ -сильно выпуклой $f(w)$

Вход: $T > 0$

- 1: $w^{(1)} = 0$
 - 2: **for** $t = 1, \dots, T$ **do**
 - 3: v_t — случайный вектор, т.ч. $\mathbb{E}[v_t | w^{(t)}] \in \partial f(w^{(t)})$
 - 4: $\eta_t = 1/(\lambda t)$
 - 5: $w^{(t+\frac{1}{2})} = w^{(t)} - \eta_t v_t$
 - 6: $w^{(t+1)} = \operatorname{argmin}_{w \in H} \|w - w^{(t+\frac{1}{2})}\|$
 - 7: **end for**
 - 8: **return** $\bar{w} = \sum_{t=1}^T w^{(t)}$
-

Лемма о SGD для сильно выпуклых функций

Лемма о SGD для сильно выпуклых функций

Пусть f λ -строго выпуклая и $\mathbb{E}[\|v_t\|^2] < \rho^2$. Тогда для $w^* \in H$ выполняется:

$$\mathbb{E}[f(\bar{w})] - f(w^*) \leq \frac{\rho^2}{2\lambda T} (1 + \log(T))$$

Общие замечания

$$L_D(w) = \mathbb{E}_{z \sim D} [l(w, z)]$$

- раньше минимизировали $L_S(w)$
- SGD позволяет минимизировать $L_D(w)$
- не знаем $\nabla L_D(w)$, зато легко построим оценку

Случай дифференцируемой /

- если I дифференцируема, то L_D — тоже
- выберем $z \sim D$
- $v_t = \nabla I(w, z)$ по w в точке $w^{(t)}$

$$\mathbb{E}[v_t | w^{(t)}] = \mathbb{E}_{z \sim D} [\nabla I(w^{(t)}, z)] = \nabla \mathbb{E}_{z \sim D} [I(w^{(t)}, z)] = \nabla L_D(w^{(t)})$$

Случай недифференцируемой /

Вместо градиента возьмём субградиент. Тогда

$$l(u, z) - l(w^{(t)}, z) \geq \langle u - w^{(t)}, v_t \rangle$$

Тогда:

$$L_D(u) - L_D(w^{(t)}) = \mathbb{E}[l(u, z) - l(w^{(t)}, z) | w^{(t)}] \quad (20)$$

$$\geq \mathbb{E}[\langle u - w^{(t)}, v_t \rangle | w^{(t)}] \quad (21)$$

$$= \langle u - w^{(t)}, \mathbb{E}[v_t | w^{(t)}] \rangle \quad (22)$$

Поэтому $\mathbb{E}[v_t | w^{(t)}]$ — субградиент $L_D(w)$ и $w^{(t)}$

SGD для минимизации true risk

Алгоритм 3 Stochastic gradient descent для минимизации $L_D(w)$

Вход: $\eta > 0$, $T > 0$

- 1: $w^{(1)} = 0$
 - 2: **for** $t = 1, \dots, T$ **do**
 - 3: выбор $z \sim D$
 - 4: выбор $v_t \in \partial l(w^{(t)}, z)$
 - 5: $w^{(t+1)} = w^{(t)} - \eta v_t$
 - 6: **end for**
 - 7: **return** $\bar{w} = \sum_{t=1}^T w^{(t)}$
-

SGD для выпукло-липшицево-ограниченных задач

SGD для выпукло-липшицево-ограниченных задач

Рассмотрим выпукло-липшицево-ограниченную задачу с параметрами ρ , B . Тогда для любого $\epsilon > 0$, то при запуске SGD на T примерах:

$$T \geq \frac{B^2 \rho^2}{\epsilon^2}$$

с $\eta = \sqrt{\frac{B^2}{\rho^2 T}}$, то результат SGD удовлетворяет:

$$\mathbb{E}[L_D(\bar{w})] \leq \min_{w \in H} L_D(w) + \epsilon$$

SGD для RLM

$$\min_w \left(\frac{\lambda}{2} \|w\|^2 + L_S(w) \right)$$

- SGD имеет выборочную сложность не хуже RLM
- RLM может выигрывать у SGD на некоторых D
- $f(w) = \frac{\lambda}{2} \|w\|^2 + L_S(w)$ — λ -строго выпуклая функция
- $v_t \in \partial l(w^{(t)}, z) \Rightarrow \lambda w^{(t)} + v_t$ — субградиент для f

Содержание

- 1 Gradient descent
 - Идея алгоритма
 - GD для выпукло-липшицевых функций
- 2 Субдифференциал
- 3 Stochastic gradient descent
 - Общая идея
 - Варианты SGD
 - SGD для задач машинного обучения

Итоги

- рассмотрели GD, SGD и понятие субградиента
- вывели вариант SGD для регуляризованного риска
- получили способ решать
выпукло-липшицево-ограниченные задачи

Литература

- Shai Shalev-Shwartz and Shai Ben-David — Understanding Machine Learning: From theory to algorithms (глава 14)