Методы стохастической оптимизации

Марат Саидов

НИУ ВШЭ

11.10.2019

Какие задачи называют стохастическими?

Это задачи оптимизации, в которых присутствует случайность.
Она проявлется как:

- Неточность в измерении функции, которую оптимизируем;
- Неточность в определении границ полиэдра;
- Случайность в самом алгоритме поиска экстремума;

Постановка задачи

Хотим минимизировать эмпирическую функцию потерь:

$$\omega^* = \arg\min_{\omega} \left(F_{\xi}(\omega) \right)$$

Считаем, что $\omega \in \Omega \subseteq \mathbb{R}^d$, и далее будем рассматривать пространство $\mathbb{R}^d$.
Ω определяется набором ограничений.

Виды стохастических задач

Стохастические задачи делятся на:

- **Оперативные** – можно следить за итерациями (пример: у больного проверяют его текущее состояние, прежде чем продолжить лечение);
- **Перспективные** – значение ω изначально фиксируется (пример: расчет оптимальной траектории полета неуправляемого объекта);

Виды ограничений

- **Детерминированные**;
- **Вероятностные** – невязка (величина ошибки) в i-м неравенстве не должна превышать заданное ε_i с вероятностью α_i;
- **Статистические** – все случайные величины в них заменяются своим ожиданием;

Знания о функции потерь

- **Полные** – для любого (неизвестного нам) параметра ξ и для любого ω можем посчитать значение $F_\xi(\omega)$ (например, функция правдоподобия);
- **Неполные** – функция задана лишь для некоторых ω или производная определена не на всем Ω;

Стохастический градиентный спуск

Напоминание. Эмпирическая функция потерь имеет конкретный вид:

$$F_\xi(\omega) = \frac{1}{\ell} \sum_{i=1}^{\ell} f_i(\omega)$$

Можем считать градиент только на подпоследовательности $\{i_t\}_{t=1}^{\infty}$:

$$\omega^{(t)} = \omega^{(t-1)} - \eta_t \cdot \nabla f_{i_t}(\omega^{(t-1)})$$

Стохастический градиентный спуск

- (+): Расходуется меньше памяти, чем в обычном градиентном спуске;
- (-): Сублинейная скорость сходимости. $F_{\xi}(\omega^{(k)}) - F_{\xi}(\omega^{*}) = O(\frac{1}{\sqrt{k}})$, тогда как у ГС сходимость линейная.

Адаптивная скорость сходимости

Рассмотрим непрерывную функцию $h : \mathbb{R} \to \mathbb{R}$.
Выбираем шаг с помощью нее: $\eta_{t+1} = h(\eta_t)$.

$h(t)$ выбирается двумя способами:

- **Метод первого порядка**;
- **Метод второго порядка (метод Ньютона)**;

Для удобства положим $\mathrm{grad}(t) = \nabla F_\xi(\omega^{(t)}))$

Метод первого порядка

Идея: рассмотрим функцию $g : \mathbb{R}^d \to \mathbb{R}$, такую, что $\eta \mapsto F_\xi(\omega^{(t)} - \eta \cdot \mathrm{grad}(t))$

Функция g показывает, насколько эффективно взять шаг η.

$$\begin{cases} \omega^{(t+1)} = \omega^{(t)} - \eta_t \cdot \mathrm{grad}(t) \\ \eta_{t+1} = \eta_t - \alpha \cdot g'(\eta_t) \end{cases}$$

Упр.
$g'(\eta_t) = -\langle \mathrm{grad}(t), \nabla F_\xi(\omega^{(t)} - \eta_t \cdot \omega^{(t)})) \rangle = -\langle \mathrm{grad}(t), \mathrm{grad}(t+1) \rangle$

Идея: Когда мы двигаемся в нужном направлении, шаги увеличиваются, а иначе уменьшаются.

Метод второго порядка (метод Ньютона)

Избавимся от гиперпараметра α в смене шага:

$$\begin{cases} \omega^{(t+1)} = \omega^{(t)} - \eta_t \cdot \text{grad}(t) \\ \eta_{t+1} = \eta_t - \frac{g'(\eta_t)}{g''(\eta_t)} \end{cases}$$

Аналогично: $g''(\eta) = -H_{F_\xi}(w^{(t)} - \eta \cdot \text{grad}(t)) \cdot \text{grad}(t)$
H – гессиан эмпирической функции потерь F_ξ.

Какой явный недостаток у данного метода?

Идея: Если при небольшом увеличении шага функция потерь немного уменьшится, то числитель положительный, и нам стоит увеличить шаг.

Метод инерции (Momentum)

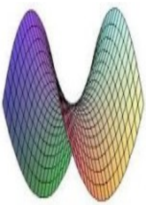

Рис. 1: Функция потерь

Функция потерь
– вытянутый по вертикали гиперболический параболоид. Если мы находимся между подъемами, как оценить поведение SGD? Использование SGD приведет к осцилляции.

Метод инерции (Momentum)

Идея: По тем измерениям, где градиент указывает в одну сторону, двигаемся быстрее. В противном случае замедляемся.

$$
\begin{cases}
\mathrm{h}_0 = 0 \\
\mathrm{h}_t = \alpha \cdot h_{t-1} + \eta_t \cdot \nabla F_{\xi}(\omega^{(t-1)}) \\
\omega^{(t)} = \omega^{(t-1)} - h_t
\end{cases}
$$

Проблема: делаем недопустимо большие шаги в сторону убывания функции.

Метод Нестерова (Nesterov accelerated gradient)

Идея: Возьмем за основу метод инерции, но будем вычислять градиент в области следующего шага (за счет $\alpha \cdot h_{t-1}$).

$$\begin{cases} \mathrm{h}_0 = 0 \\ \mathrm{h}_t = \alpha \cdot h_{t-1} + \eta_t \cdot \nabla F_\xi(\omega^{(t-1)} - \alpha \cdot h_{t-1}) \\ \omega^{(t)} = \omega^{(t-1)} - h_t \end{cases}$$

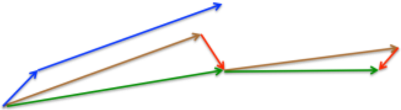

Рис. 2: Метод инерции (синий) и метод Нестерова (зеленый, коричневый, красный)

AdaGrad

Метод учитывает частотность признаков: большая длина шага подбирается для редких признаков, маленькая – для частых.

Поэтому он пригоден для работы с разреженными данными.

Популярен в NLP.
Пример: GloVe (Global Vectors for Word Representation) – построение векторных представлений (embeddings) для слов. AdaGrad использовался для обучения.

AdaGrad

Для каждой координаты будет выбирать свой шаг:

$$g_{t,i} = \nabla F_{\xi}(\omega_i^{(t-1)})$$

Обновим параметры:

$$\omega_{t+1,i} = \omega_{t,i} - \frac{\eta}{\sqrt{G_{t,ii} + \varepsilon}} \cdot g_{t,i}, \varepsilon \approx 10^{-8}$$

G_t – диагональная матрица, $G_{t,ii}$ – сумма квадратов градиентов для $\omega_{t,i}$ за все время до t.

Проблема: для больших t выражение $G_{t,ii}$ становится большим,и скорость сходимости становится бесконечно малой.

AdaDelta

Пусть $E[g^2]_t$ – средний квадрат градиента за t шагов. Предположим, что в момент t скользящее среднее – $E[g^2]_t$, хотим экспоненциальную сходимость:

$$E[g^2]_t = \varrho \cdot E[g^2]_{t-1} + (1 - \varrho) \cdot g_t^2$$

Аналогично рассмотрим среднее для изменение параметра:

$$E[\Delta\omega^2]_t = \gamma E[\Delta\omega^2]_{t-1} + (1 - \gamma)\Delta\omega^2$$

$$\rho, \gamma \approx 0.9$$

AdaDelta

Из последнего получается RMSE:

$$RMSE = RMS[\Delta\omega]_t = \sqrt{E[\Delta\omega^2]_t + \varepsilon}$$

$$\begin{cases} \mathrm{h}_t = -\frac{RMS[\Delta\omega]_{t-1}}{RMS[g]_t} \cdot g_t \\ \omega^{(t+1)} = \omega(t) + h_t \end{cases}$$

RMSprop

Частный случай AdaDelta, эмпирически хорошее значение для $\eta = 10^{-4}, \gamma = 0.9$:

$$\begin{cases} \mathrm{E}[\mathrm{g}^2]_t = 0.9E[g^2]_{t-1} + 0.1g_t^2 \\ \omega^{(t+1)} = \omega^{(t)} + \frac{\eta}{\sqrt{E[g^2]_t + \varepsilon}} g_t \end{cases}$$

Утверждается, что заданные гиперпараметры – баланс между агрессивной скоростью сходимости и затуханием.

Adam

Моменты тоже обновляются взвешенно:

$$\begin{cases} \mathrm{m}_t = \beta_1 m_{t-1} + (1-\beta_1) g_t \\ \mathrm{v}_t = \beta_2 v_{t-1} + (1-\beta_2) g_t^2 \end{cases}$$

Рассмотрим оценки на эти величины:

$$\hat{m}_t = \frac{m_t}{1-\beta_1^t}$$

$$\hat{v}_t = \frac{v_t}{1-\beta_2^t}$$

Обновление параметра происходит с использованием оценок:

$$\omega^{(t+1)} = \omega^{(t)} - \frac{\eta}{\sqrt{\hat{v}_t + \varepsilon}} \hat{m}_t$$

AdaMax

Возможное улучшение Adam: v_t обратно пропорционален ℓ_2-норме от предыдущих v. Перейдем к ℓ_∞:

$$u_t = \beta_2^\infty v_{t-1} + (1 - \beta_2^\infty)|g_t|^\infty = \max(\beta_2 \cdot v_{t-1}, |g_t|)$$

Перепишем обновление параметра:

$$\omega^{(t+1)} = \omega^{(t)} - \frac{\eta}{u_t}\hat{m}_t$$

Дополнительные эвристики

Методы стохастической оптимизации могут быть улучшены за счет:

- Если упорядочить исходные данные правильным образом, то это может дать лучшую сходимость – Curriculum Learning;
- Нормализация батчей как устойчивость к вычислительным погрешностям;

Дополнительные эвристики

Методы стохастической оптимизации могут быть улучшены за счет:

- Быстрая остановка: нужно понять, когда функция потерь перестает сильно изменяться;
- Добавление шума к градиенту: $g_{t,i} = g_{t,i} + \mathcal{N}(0, \sigma_t^2)$. Это позволяет избежать попадание в локальный экстремум;

Контрольные вопросы

- Почему методы с адаптивной скоростью сходимости лучше SGD?
- Почему для разреженных признаков нужна высокая скорость сходимости, а для плотных – низкая?
- Как AdaDelta и RMSprop решают проблему агрессивной скорости сходимости AdaGrad?

Ссылки

1 . S. Ruder, 'An overview of gradient descent optimization algorithms', Insight Centre for Data Analytics, NUI Galway, Aylien Ltd., Dublin.

2 . L.A. Hannah (2014), 'Stochastic optimization'.

3 . W.B. Powell, 'A unified framework for stochastic optimization', Department of Operations Research and Financial Engineering, Princeton University, Sherrerd Hall, Princeton, NJ 08544, United States