Вычислительные аспекты оптимизации. Гладкие функционалы и пр. Метод стохастического градиента как метод оптимизации

Бакшинская Екатерина
Зенкова Наталья
Балагуров Владимир

Санкт-Петербург
2019г.

Постановка задачи

Обучающая выборка: $X^n = (x_i, y_i)_{i=1}^n, \quad x_i \in \mathbb{R}^p$

В регрессии:

- Целевая переменная $y_i \in \mathbb{R}$.
- Задача:

$$Q(w) = \sum_{i=1}^{n} (a(x_i, w) - y_i)^2 \ \to \min_w$$

В классификации:

- Целевая переменная $y_i \in \{-1, +1\}$.
- Задача:

$$Q(w) = \sum_{i=1}^{n} [a(x_i, w) y_i < 0] \leq \sum_{i=1}^{n} \mathscr{L}(M_i(w)) \to \min_w,$$

где $M_i(w) = \langle x_i, w \rangle y_i$ — отступ (margin) объекта x_i

Проблемы оптимизации в ML

- Оптимизация часто условная. Решение проблемы — сведение условной оптимизации к безусловной.
- Оптимизируемая функция бывает негладкой (или вообще не непрерывной). Решение — аппроксимация гладкой.

Для классификаци вспомним постановку задачи в SVM:

$$\sum_{i=1}^{n}(1 - M_i(w))_+ + \frac{1}{2C}||w||^2 \to \min_w,$$

здесь $M_i(w) = y_i(\langle w, x_i \rangle - w_0)$.

Эквивалентная задача безусловной минимизации:

$$C\sum_{i=1}^{n}(1 - M_i(w))_+ + \frac{1}{2}||w||^2 \to \min_w,$$

Градиентный метод численной оптимизации

Минимизация эмпирического риска (регрессия, классификация):

$$Q(w) = \sum_{i=1}^{n} \mathscr{L}_i(w) \to \min_w,$$

Численная минимизация методом градиентного спуска:

Вход: $X^n = (x_i, y_i)_{i=1}^n,\ h$
Выход: w

1. Инициализация $w^{(0)}$
2. **Повторять**
   1. Вычислять: $\nabla Q(w^t)$
   2. Градиентный шаг: $w^{t+1} = w^t - h\nabla Q(w^t)$

пока $||w^t - w^{t-1}|| > \varepsilon$

Варианты выбора градиентного шага и сходимость

- сходимость гарантируется (для выпуклых функций) при

$$h_t \to 0, \quad \sum_{t=1}^{\infty} h_t = \infty, \quad \sum_{t=1}^{\infty} h_t^2 < \infty$$

  В частности можно положить $h_t = 1/t$

- метод скорейшего градиентного спуска:

$$Q(w - h\nabla Q(w)) \to \min_h$$

  позволяет найти адаптивный шаг h^*

- периодически можно делать пробные случайные шаги для “выхода” из локальных минимумов

и др.

Если функционал $Q(w)$ выпуклый, гладкий и имеет минимум w^*, то имеет место следующая оценка сходимости:

$$Q(w^t) - Q(w^*) = O(\frac{1}{t})$$

Недостаток обычного метода градиентного спуска

Проблема — минимизируемая функция представляет собой сумму слагаемых $\mathscr{L}_i(w)$, количество которых равно объему выборки.

Вычисление $Q(w)$ и $\nabla Q(w)$ становится трудоемким.

Метод стохастического градиента: вычисляем не точное значение градиента, а его (случайную и желательно несмещенную) оценку.

Оценивание градиента

Оценить градиент суммы можно градиентом одного случайно взяого слагаемого:

$$\nabla \bar{Q}(w) \approx \nabla \mathscr{L}_i(w),$$

где i — равномерно распределены на $1, ..., n$. Градиентный шаг переписывается в виде:

$$w^{t+1} = w^t - h\nabla \mathscr{L}_i(w)$$

После чего вычисляем оценку функционала $\bar{Q}(w) = \frac{1}{n}Q(w) = \frac{1}{n}\sum_{i=1}^{n} \mathscr{L}_i$, которая является несмещенной т.к. $\mathbb{E}\nabla \mathscr{L}_i = \nabla \bar{Q}(w)$

Для выпуклого и гладкого функционала может быть получена следующая оценка:

$$\mathbb{E}[Q(w^t) - Q(w^*)] = O(\frac{1}{\sqrt{t}})$$

Стохастический градиентный спуск

Вход: $X^n = (x_i, y_i)_{i=1}^n,\ h,\ \lambda$
Выход: w

1. Инициализация $w_j,\ j = 0, \dots, p$
2. Инициализация $\bar{Q}(w) = \frac{1}{n} Q(w) = \frac{1}{n} \sum_{i=1}^n \mathscr{L}_i(w)$
3. **Повторять**
    1. Выбор x_i из X^n случайным образом
    2. Вычисление значения функции потери: $\varepsilon_i = \mathscr{L}_i(w)$
    3. Градиентный шаг: $w^{t+1} = w^t - h \nabla \mathscr{L}_i(w)$
    4. Оценка функционала: $\bar{Q}(w) = (1 - \lambda) \bar{Q}(w) + \lambda \varepsilon_i$

пока значение $\bar{Q}(w)$ и/или веса w не сойдутся

Возникновение такой оценки функционала

Проблема: после каждого шага w по одному объекту x_i не хотим оценивать Q по всей выборке $x_1, ..., x_n$

Решение: использование рекуррентной формулы

Среднее арифметическое $\bar{Q}_m = \frac{1}{m}\sum_{i=1}^{m}\varepsilon_i$:

$$\bar{Q}_m = (1 - \frac{1}{m})\bar{Q}_{m-1} + \frac{1}{m}\varepsilon_m$$

Экспоненциальное скользящее среднее:

$$\bar{Q}_m = (1 - \lambda)\bar{Q}_{m-1} + \lambda\varepsilon_m$$

$$\bar{Q}_m = \lambda\varepsilon_m + \lambda(1 - \lambda)\varepsilon_{m-1} + \lambda(1 - \lambda)^2\varepsilon_{m-2} + \dots$$

Параметр $\lambda \approx \frac{1}{m}$ — *темп забывания*

Например, можно повысить точность оценки градиента, используя несколько слагаемых вместо одного (**mini-batch gradient descent**):

$$\nabla Q(w) \approx \frac{1}{m}\sum_{j=1}^{m} \nabla \mathscr{L}_{itj}(w),$$

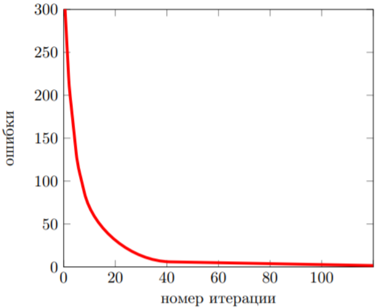
где i_j — случайно выбранные номера слагаемых из функционала, а m — параметр метода.

Stochastic average gradient:

$$\nabla Q(w) \approx \frac{1}{n}\sum_{i=1}^{n} z_i^{(t)},$$

где $z_i^{(t)} = \begin{cases} \nabla \mathscr{L}_i(w^{t-1}) & \text{если } i = i_t \\ z_i^{t-1} & \text{иначе} \end{cases}$ и $z_i^{(0)} = \nabla \mathscr{L}_i(w^{(0)})$

Сходимость метода стохастического градиента



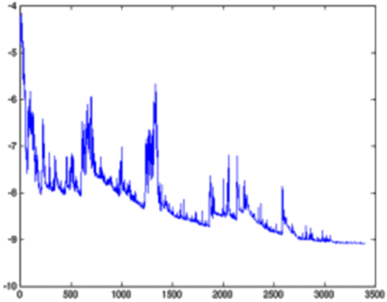


Рис. 1: Ошибка в зависимости от номера итерации

Квадратичная регуляризация

Чтобы ограничить рост абсолютных значений весов, к минимизируемому функционалу $Q(w)$ добавляется штрафное слагаемое:

$$Q_\tau(w) = Q(w) + \frac{\tau}{2}||w||^2 = Q(w) + \frac{\tau}{2}\sum_{j=1}^{n} w_j^2 \to \min_w$$

Аддитивная поправка в градиенте:

$$\nabla Q_\tau(w) = \nabla Q(w) + \tau w$$

Правило обновления весов принимает вид:

$$w^t = w^{t-1}(1 - h\tau) - h\nabla Q(w^{t-1})$$

Недостаток: параметр τ приходится подбирать в режиме скользящего контроля, что связано с большими вычислительными затратами.

Варианты инициализации весов

- $w_j^{(0)} = 0$ для всех $j = 0, ..., n$
- небольшие случайные значения: $w_j^{(0)} = random(-\frac{1}{2n}, \frac{1}{2n})$
- для регрессии: $w_j^{(0)} = \frac{\langle y, x_j \rangle}{\langle x_j, x_j \rangle}$
- для классификации: $w_j^{(0)} = \ln \frac{\sum_i [y_i = +1] x_{ij} \sum_i [y_i = -1]}{\sum_i [y_i = -1] x_{ij} \sum_i [y_i = +1]}$
- оценки $w_j^{(0)}$ по небольшой случайной подвыборке объектов
- мультистарт: многократные запуски из разных случайных приближений и выбор лучшего решения

Варианты порядка предъявления объектов

Несколько вариантов выбора наблюдения (специфично для классификации), помимо выборки из выборочного распределения:

- перетасовка объектов (shuffling): попеременно брать объекты из разных классов.
- чаще брать те объекты, на которых была допущена бо́льшая ошибка:
  (чем меньше M_i, тем больше вероятность взять объект)
  (чем меньше $|M_i|$, тем больше вероятность взять объект)
- вообще не брать объекты у которых $M_i > \mu_+$
- вообще не брать объекты у которых $M_i < \mu_-$

параметры μ_+, μ_- придется подбирать

Пеимущества и недостатки метода стохастического градиента

Пеимущества:

- Легко реализуется.
- Функция потерь и семейство алгоритмов могут быть любыми.
- Легко добавить регуляризацию.
- Метод подходит для динамического обучения, когда обучающие объекты поступают потоком, и вектор весов обновляется при появлении каждого объекта.
- Подходит для задач с большими данными, иногда можно получить решение даже не обработав всю выборку.

Недостатки:

- Подбор эвристик является искусством (не забыть про переобучение, застревание, расходимость)