Семинар по градиентным методам

Талгат Даулбаев

10 февраля 2016 г.

1 Введение

Методы оптимизации и машинное обучение тесно связаны.

Так оптимизация постоянно используется в машинном обучении: мы берём некоторую модель алгоритмов, и нам нужно обучить её по данным — решить задачу оптимизации функционала качества.

Машинное обучение в свою очередь постоянно бросает новые вызовы разработчикам методов оптимизации, заставляя их придумывать алгоритмы для задач машинного обучения.

Оптимизационные задачи делятся на *условные* и *безусловные*:

- Задача безусловной оптимизации заключается в нахождении минимума[1] некоторой функции $f : \mathbb{R}^n \to \mathbb{R}$ на всём пространстве $\mathbb{R}^n$:

$$f(x) \to \min_{x \in \mathbb{R}^n}.$$

- Задача условной оптимизации заключается в нахождении минимума f уже не на всём пространстве $\mathbb{R}^n$, а на некотором его подмножестве. Примером такого подмножества может служить замкнутый шар $\{x \in \mathbb{R}^n : \|x\|_2 \leq 1\}$.

На этом семинаре мы разберём простейшие методы решения задачи безусловной оптимизации.

2 Одномерный случай

При $n = 1$ подобные задачи решают даже в школе. Вспомним, как нужно действовать.

Пусть f — некоторая одномерная дважды дифференцируемая функция. Тогда необходимым условием экстремума в некоторой точке x^* является равенство нулю её производной при $x = x^*$.

Поэтому нужно найти все корни уравнения

$$f'(x) = 0.$$

А затем следует проверить знак второй производной в точках, подозрительных на экстремум:

[1] Задача максимизации функции f сводится к минимизации функции $-f$.

- если $f''(x^*) > 0$, то x^* — точка минимума;
- если $f''(x^*) < 0$, то x^* — точка максимума;
- если $f''(x^*) = 0$, то x^* — точка перегиба;

Однако на практике не всё так просто. Очень часто получить аналитическое решение уравнения $f'(x) = 0$ невозможно. Например, знаменитая теорема Абеля — Руффини гласит, что при $n \geq 5$ нельзя указать формулу, которая выражала бы корни любого многочлена степени n через его коэффициенты при помощи радикалов. А это значит, что если $f(x)$ — полином хотя бы шестой степени, уравение $f'(x) = 0$ не получится решить аналитически.

Здесь на помощь приходят численные методы одномерной минимизации, о которых мы сейчас говорить не будем.

3 Многомерный случай

§3.1 Градиент функции

Как правило, для функций многих переменных задача

$$f(x) \to \min_{\mathbb{R}^n}$$

решается с помощью так называемых градиентных методов (или же методов, использующих как градиент, так и информацию о производных более высокого порядка).

Градиентом функции $f : \mathbb{R}^n \to \mathbb{R}$ называется вектор его частных производных

$$\nabla f(x) = \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}.$$

Естественно, определён он лишь в том случае, когда существуют частные производные.

Заметим, что градиент — это не константный вектор, а функция $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$, поэтому говорить о значении градиента можно лишь в конкретной точке.

Необходимым условием экстремума в некоторой точке x^* является равенство градиента в этой точке нулевому вектору:

$$\nabla f(x) = 0.$$

Как мы помним, в одномерном случае подозрительная на экстремум точка могла быть только минимумом, максимум или точкой перегиба. В многомерном же случае одна точка может быть точкой максимума по одной координате, точкой минимума по другой, точкой максимума по третьей и так далее.

Но для определения, является ли подозрительная на минимум точка, точкой минимума, нужно ввести понятие *гессиана*, или *матрицы Гессе*, функции f — матрицы из вторых частных производных:

$$\nabla^2 f(x) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f}{\partial x_2 \partial x_1} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \frac{\partial^2 f}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix}.$$

Заметим, что гессиан — это тоже не константная матрица, а функция, действующая из $\mathbb{R}^n$ в $\mathbb{R}^{n \times n}$.

Необходимым условием уже минимума в точке x^* является неотрицательная определённость гессиана, то есть

$$\forall d \in \mathbb{R}^n \quad d^T \nabla^2 f(x^*) d \geq 0, \quad \text{причём} \quad d^T \nabla^2 f(x^*) d = 0 \;\Leftrightarrow\; d = 0.$$

Обозначается это как

$$\nabla^2 f \succeq 0.$$

Достаточным условием минимума является положительная определённость гессиана, то есть

$$\forall d \in \mathbb{R}^n \quad d^T \nabla^2 f(x^*) d > 0,$$

что обозначается как

$$\nabla^2 f \succ 0.$$

3.1.1 Ключевое свойства градиента

Градиент является направлением наискорейшего роста функции, а антиградиент (то есть $-f$) — направлением наискорейшего убывания.

Докажем это утверждение. Пусть $v \in \mathbb{R}^n$ — произвольный вектор, лежащий на единичной сфере: $\|v\| = 1$, а $\hat{x} \in \mathbb{R}^n$ — фиксированная точка пространства. Скорость роста функции в точке $\hat{x}$ вдоль вектора v характеризуется производной по направлению $\frac{\partial f}{\partial v}$:

$$\frac{\partial f}{\partial v} = \frac{d}{dt} f(\hat{x}_1 + tv_1, \hat{x}_2 + tv_2, \ldots, \hat{x}_n + tv_n)\big|_{t=0}.$$

Эту производную можно посчитать по правилу дифференцирования сложной функции многих переменных:

$$\frac{\partial f}{\partial v} = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} \frac{d}{dt}(\hat{x}_i + tv_i) = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i} v_i = \langle \nabla f, v \rangle.$$

Распишем скалярное произведение:

$$\langle \nabla f, v \rangle = \|\nabla f\| \|v\| \cos\varphi = \|\nabla f\| \cos\varphi,$$

где φ — угол между градиентом и вектором v. Таким образом, производная по направлению будет максимальной, если угол между градиентом и направлением равен нулю, и минимальной, если угол равен 180 градусам. Иными словами, производная по направлению максимальна вдоль градиента и минимальна вдоль антиградиента.

3.1.2 Векторное дифференцирование

При аналитическом вычислении градиента крайне полезны формулы векторного дифференцирования. Выведем простейшие из них.

Задача 3.1. *Покажите, что*

$$\nabla_x \langle a, x \rangle = a,$$

где $\langle ., . \rangle$ *— это обычное скалярное произведение.*

Решение. Найдем производную по j-й координате:

$$\frac{\partial}{\partial x_j} \langle a, x \rangle = \frac{\partial}{\partial x_j} \sum_{k=1}^{n} a_k x_k = a_j.$$

Значит, градиент равен a.

■

Задача 3.2. *Покажите, что*

$$\nabla_x \|x\|_2^2 = 2x.$$

Решение. Найдем производную по j-й координате:

$$\frac{\partial}{\partial x_j} \|x\|_2^2 = \frac{\partial}{\partial x_j} \sum_{k=1}^{n} x_k^2 = 2x_j.$$

Значит, градиент равен $2x$.

■

Задача 3.3. *Покажите, что*

$$\nabla_x \langle Ax, x \rangle = (A + A^T)x,$$

где $A \in \mathbb{R}^{n \times n}$.

Решение. Распишем интересующую нас функцию:

$$\langle Ax, x\rangle = \sum_{j=1}^{n}(Ax)_j x_j = \sum_{j=1}^{n}\left(\sum_{k=1}^{n} a_{jk}x_k\right)x_j = \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}x_j x_k = \sum_{j=1}^{n} a_{jj}x_j^2 + \sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq j}}^{n} a_{jk}x_j x_k. \quad (3.1)$$

Во втором равенстве в 3.1 мы просто расписали j-ую компоненту вектора Ax, в последнем — разделили слагаемые, содержащие квадраты, от остальных для удобства дальнейшего дифференцирования.

Найдем частную производную по i-й координате:

$$\frac{\partial}{\partial x_i}\langle Ax, x\rangle = 2a_{ii}x_i + \sum_{\substack{j=1\\j\neq i}}^{n} a_{ij}x_j + \sum_{\substack{k=1\\k\neq i}}^{n} a_{ki}x_k = \sum_{j=1}^{n} a_{ij}x_j + \sum_{k=1}^{n} a_{ki}x_k = (Ax)_i + (A^Tx)_i. \quad (3.2)$$

В предпоследнем равенстве 3.2 мы занесли по $a_{ii}x_i$ в каждую из сумм.

Итого

$$\nabla_x\langle Ax, x\rangle = Ax + A^Tx = (A + A^T)x.$$

■

Задача 3.4. *Покажите, что*

$$\nabla_x\|Ax + b\|_2^2 = 2A^T(Ax + b).$$

Здесь $x \in \mathbb{R}^n,\ A \in \mathbb{R}^{m\times n},\ b \in \mathbb{R}^m$.

Решение. Распишем норму:

$$\|Ax+b\|_2^2 = \langle Ax+b, Ax+b\rangle = \langle Ax, Ax\rangle + 2\langle Ax, b\rangle + \langle b, b\rangle = \langle A^TAx, x\rangle + 2\langle x, A^Tb\rangle + \langle b, b\rangle.$$

Здесь мы использовали свойство, что $\langle Ax, y\rangle = \langle x, A^Ty\rangle$, и симметричность скалярного произведения: $\langle x, y\rangle = \langle y, x\rangle$.

Воспользуемся уже полученными нами формулами векторного дифференцирования:

$$\nabla_x\|Ax + b\|_2^2 = \nabla_x\langle A^TAx, x\rangle + 2\nabla_x\langle x, A^Tb\rangle + \nabla_x\langle b, b\rangle = (A^TA + A^TA)x + 2A^Tb = 2A^TAx + 2A^Tb = 2A^T(Ax + b)$$

■

4 Градиентный спуск

Наиболее простым итерационным способом решения задач оптимизации вида

$$f(x) \to \min_{x \in \mathbb{R}^n}$$

является *метод градиентного спуска* (gradient descent, GD), в котором выбирается некоторое начальное приближение x^0, и затем до сходимости делаются шаги по антиградиенту:

$$x^k = x^{k-1} - \eta_k \nabla f(x^k).$$

Известно, что если f — выпуклая, гладкая функция, достигающая минимума в точке x^*, то имеет место следующая оценка сходимости:

$$f(x^k) - f(x^*) = O(1/k).$$

Однако оптимизируемая (целевая) функция в задачах машинного обучения обычно имеет вид суммы по всем объектам обучающей выборки от большого числа функций:

$$Q(w) = \sum_{i=1}^{l} q_i(w) \to \min_{w \in \mathbb{R}^d}$$

Так, например, при построении линейной регрессии решается задача

$$\sum_{i=1}^{l} (\langle w, x_i \rangle - y_i)^2 \to \min_{w \in \mathbb{R}^d},$$

где l — мощность обучающей выборки, а d — количество признаков.

И если l велико, то градиентный спуск будет очень трудоёмким. Однако для функции такого особого вида можно воспользоваться методом *стохастического градиента* (stochastic gradient descent, SGD). Его суть заключается в том, что шаг градиентного спуска делается не по всей выборке, а по небольшой случайной подвыборке. Например, шаг можно делать даже по одному объекту:

$$w^{k+1} = w^k - \eta_k \nabla q_{i_k}(w),$$

где i_k — это случайно выбранный объект.

Для выпуклой и гладкой функции с точкой минимума w^* может быть получена следующая оценка:

$$\mathbb{E}[Q(w^k) - Q(w^*)] = O\left(1/\sqrt{k}\right).$$

Недавно был предложен метод среднего стохастического градиента (stochastic average gradient), который сочетает в себе быстроту итераций стохастического градиента и высокую скорость сходимости полного градиента. Перед началом итераций

в нём выбирается начальное приближение w^0 и инициализируются вспомогательные переменные y_i^0, соответствующие градиентам слагаемых функционала:

$$y_i^0 = \nabla q_i(w^0), \quad i = 1, \ldots, n.$$

На k-й итерации выбирается случайное слагаемое i_k и обновляются вспомогательные переменные:

$$y_i^k = \begin{cases} \nabla q_i(w^{k-1}), & \text{если } i = i_k \\ y_i^{k-1}, & \text{иначе} \end{cases}$$

Иными словами, пересчитывается один из градиентов слагаемых. Наконец, делается градиентный шаг:

$$w^k = w^{k-1} - \eta_k \sum_{i=1}^{l} y_i^k.$$

Данный метод имеет такой же порядок сходимости для выпуклых и гладких функций, как и обычный градиентный спуск:

$$\mathbb{E}[Q(w^k) - Q(w^*)] = O\left(1/k\right).$$

Список литературы

[1] *Евгений Соколов* Семинары по линейным классификаторам

[2] *Д. А. Кропотов* Курс лекций «Методы оптимизации в машинном обучении»