Вычислительные аспекты оптимизации

Метод стохастического градиента

Пимахов Кирилл, 622 гр.

2017

Содержание

1 Постановка задачи

Пусть имеется матрица данных $\mathbf{X} = [x_1, \ldots, x_p]$, $x_j \in \mathbb{R}^n, j = 1 \ldots p$, где p – количество переменных, n – количество наблюдений. Целевая переменная $y \in \mathbb{R}^n$ в случае регрессии, $y_i \in \{-1, +1\}$, $i = 1 \ldots n$ в случае классификации.

Задача в случае регрессии: найти параметры w семейства функций $f(x, w)$, минимизирующие

$$\frac{1}{n}\sum_{i=1}^{n}\mathscr{L}(f(x_i, w) - y_i) = \frac{1}{n}Q(w), \tag{1}$$

где $\mathscr{L}$ – функция потерь.

Например, в этих обозначениях записывается линейная регрессия: $w = (\beta_0, \beta_1, \ldots, \beta_p)$, $f(x_i, w) = \sum_{j=1}^{p} x_{ij} * \beta_j + \beta_0$, $\mathscr{L}(f(x_i, w) - y) = (f(x_i, w) - y)^2$. Так же к минимизации (1) сводится и любая параметрическая регрессия.

В случае классификации так же задача заключается нахлждении параметров w семейства функций $g(x, w)$, минимизирующие следующую целевую функцию:

$$\frac{1}{n}\sum_{i=1}^{n}\mathscr{L}(g(x_i, w)y_i) = \frac{1}{n}\sum_{i=1}^{n}[g(x_i, w)y_i < 0] = \frac{1}{n}Q(w). \tag{2}$$

Обозначим $M_i(w) = g(x_i, w)y_i$ – отступ.

Вместо $\mathscr{L}(g(x_i, w)y_i)$ можно брать гладкую (или непрерыаную) оценку сверху, в том числе за счет оценки $[M_i(w) < 0]$. Вспомним постановку задачи в SVM:

$$\sum_{i=1}^{n}(1 - M_i(w))_+ + \frac{1}{2C}\|w\|^2 \to \min_w, \tag{3}$$

где $M_i = y_i\left(\langle w, x_i\rangle - w_0\right)$. Замена целевой функции на мажоранту меняет смысл задачи. В (3) штрафуются близкие к разделяющей гиперплоскости наблюдения, а также введена регуляризация.

2 Метод стохастического градиента

Один из простейших методов решения задачи минимизации $Q(w)$, определенной в (1) или в (2) – градиентный спуск.

Алгоритм .1. Градиентный спуск

1: Инициализация параметров $w^{(0)}$ и выбор скорости обучения h.

2: **Повторять**

3: Вычисление градиента Q при текущем векторе параметров.

$$\nabla Q(w^{(t)}) = \left(\frac{\partial Q(w^{(t)})}{\partial w_1}, \ldots, \frac{\partial Q(w^{(t)})}{\partial w_p} \right)$$

.

4: Обновление вектора параметров $w^{(t+1)} = w^{(t)} - h\nabla Q(w^{(t)})$.

5: **Пока выполняется** $|Q(w^{(t+1)}) - Q(w^{(t)})| > \epsilon$ или $\|w^{(t+1)} - w^{(t)}\| > \epsilon$ или $\|\nabla Q(w^{(t)})\| > \epsilon$.

Условие остановки алгоритма .1 в 5 пункте выбирается в зависимости от постановки задачи.

Проблемой применения алгоритма .1 для задач (1) и (2) является то, что минимизируемая функция представляет собой сумму слагаемых $\mathcal{L}$, количество которых равно объему выборки. При больших объемах выборки вычисление $Q(w)$ и $\nabla Q(w)$ становится трудоемким.

Метод стохастического градиента решает эту проблему. Идея заключается в использовании оценок целевой функции и ее градиента вместо их прямого вычисления

Алгоритм .2. Метод стохастического градиента

1: Инициализация вектора параметров $w^{(0)}$, выбор скорости обучения h, темп забывания λ, вычисление начального значения функционала

$$\overline{Q}(w_0) = \frac{1}{n}Q(w_0) = \frac{1}{n}\sum_{i=1}^{n}\mathscr{L}(g(x_i, w_0), y_i).$$

2: **Повторять**

3: Случайный выбор элемента из выборки x_i и вычисление функции потерь

$$\epsilon_i = \mathscr{L}(g(x_i, w), y_i)$$

.

4: Обновление вектора параметров $w^{(t+1)} = w^{(t)} - h\nabla\mathscr{L}(g(x_i, w^{(t)}), y_i)$.

5: Оценка функционала $Q(w) = (1-\lambda)Q(w) + \lambda\epsilon_i$.

6: **Пока выполняется** $|Q(w^{(t+1)}) - Q(w^{(t)})| > \epsilon$ или $\|w^{(t+1)} - w^{(t)}\| > \epsilon$ или $\|\nabla Q(w^{(t)})\| > \epsilon$.

3 Сходимость метода стохастического градиента

Что касается сходимости алгоритма .2, следует учитывать, что оценки параметров w – случайные величины, поэтому под сходимостью понимается сходимость почти всюду. В (Stochastic learning, Léon Bottou, 2003) приводятся условия, при которых стохастический градиент сходится почти всюду при

$$h_t \to 0, \sum_{t=1}^{\infty} h_t = \infty, \sum_{t=1}^{\infty} h_t^2 < \infty. \tag{4}$$

Более сильным, но простым является условие строгой выпуклости целевой функции. Выпуклой, например, является целевая функция для линейной регрессии: гессиан в этом случае равен $\mathbf{X}^{\mathrm{T}}\mathbf{X}$, и если он положительно определен, то задача строго выпукла.

Также в (Stochastic learning, Léon Bottou, 2003) по поводу скорости сходимости утверждается, что $1/\|w_t - w^*\|^2$ растет линейно по t. Использование второй производной (гессиана) целевой функции ускоряет сходимость. Обозначим $\Phi^{(t)}$ – оценку обратной матрицы к гессиану на шаге t, тогда обновление параметров:

$$w^{(t+1)} = w^{(t)} - h\Phi^{(t)}\nabla\mathscr{L}(g(x_i, w^{(t)}), y_i). \tag{5}$$

Вместо $\nabla\mathscr{L}(g(x_i, w), y_i)$ можно использовать более точную оценку градиента:

$$1/M \sum_{j=1}^{M} \mathscr{L}(g(x_j, w), y_i).$$

В случае достаточно точной оценки градиента, имеет смысл выбирать h, обеспечивающую скорейший спуск:

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

4 Инициализация параметров

Несколько вариантов инициализации $w^{(0)}$.

- $w_j^{(0)} = 0$, $j = 1 \dots p$.
- Небольшие случайные значения $w_j^{(0)} \in U(-1/(2n), 1/(2n))$.
- Для регрессии: $w_j^{(0)} = \langle y, x_j \rangle / \langle x_j, x_j \rangle$.
- Для классификации: $w_j^{(0)} = \log\left(\sum_i [y_i = 1] x_{ij} / \sum_i [y_i = -1] x_{ij}\right)$.
- Обучение по небольшой случайной подвыборке. Как и в предыдущих 2 пунктах ожидается, что начальные параметры окажутся достаточно близко к оптимальным.
- Мультистарт. Многократный запуск на случайных начальных значениях $w_j^{(0)}$ и выбор наилучшего. Помогает решить проблему выбора параметров из локального минимума.

Скорость обучения можно инициализировать, например, $h_t = 1/t$, это удовлетворяет условию сходимости метода (4). В качестве темпа забывания λ можно взять $1/n$, где n– объем выборки.

5 Предъявление объектов

Несколько вариантов предъявления объектов, помимо выборки из выборочного распределения:

- Попеременно брать наблюдения из разных классов (shuffling).

- Чаще брать наблюдения с большей ошибкой (отступом), например, вероятность предъявления задать пропорционально величине ошибки. Не стоит использовать, если в выборке есть выбросы.
- Не брать «хорошие» наблюдения большим положительным отступом M_i, то есть такие, что $M_i > \mu_+ > 0$.
- Не брать выбросы, наблюдения с большим отрицательным отступом: $M_i < \mu_- < 0$.

6 Проблема переобучения

Несмотря на то, что проблема переобучения – это, в основном, проблема неправильного выбора модели (например слишком много параметров), недостаточного объема выборки или избыточного числа переменных, с этим можно бороться на этапе оптимизации. При избыточном количестве параметров или при малом объеме выборки, оценки этих параметров неустойчивы, а значит могут принимать большие значения. С помощью регуляризации можно напрямую бороться с этой проблемой. Введем дополнительное слагаемое L_2 норму вектора параметров:

$$Q_\tau(w) = Q(w) + \frac{\tau}{2}\|w\|_2^2,$$

где τ – параметр регуляризации.

Для градиентного метода это значит, что на каждом шаге мы двигаем вектор параметров к нулю:

$$w^{(t+1)} = (1 - h\tau)w^{(t)} - h\nabla\mathscr{L}(g(x_i, w^{(t)}), y_i).$$

7 Некоторые модификации метода стохастического градиента

- Momentum: Изменение вектора параметров $\Delta w^{(t+1)} = -h\nabla Q(w^{(t)}) + \alpha\Delta w^{(t)}$. То есть параметры двигаются как бы с инерцией, чем больше α – тем больше инерция.
- Nesterov Accelerated Gradient: $\Delta w^{(t+1)} = -h\nabla Q(w^{(t)} + \alpha\Delta w^{(t)}) + \alpha\Delta w^{(t)}$. Здесь мы считаем градиент в той точке, куда бы мы двинулись по инерции, то есть заглядываем

вперед. Если целевая функция впереди уменьшается быстрее, то и параметр будет двигаться к минимуму быстрее.

- AdaGrad: Идея: сильнее менять параметры, которые меняются редко. На каждом шаге пересчитывается вектор накопленных изменений параметров $G^{(t)} = G^{(t-1)} + \left(\nabla Q(w^{(t-1)})\right)^2$, квадрат вектора поэлементный. Изменение вектора параметров $\Delta w^{(t+1)} = -h/\sqrt{G^{(t)} + \epsilon}\,\nabla Q(w^{(t)})$, здесь операции над векторами поэлементные. Параметр ϵ для отделения знаменателя от нуля.