Методы оптимизации в машинном обучении
Практическое задание #1

Рожин Андрей

НИУ ВШЭ — 15 мая 2022 г.

Вступление

Решение задач оптимизации, является неотъемлемой частью машинного обучения. В этом документе обсуждаются базовые подходы к решению задач этого типа, а также проведение и анализ экспериментов и аналитический вывод формулы логистической регрессии в матричном виде.

$$L(w) = -\frac{1}{m}\sum_{i=1}^{m}\left[y_i log\left(\frac{1}{1+e^{-w^T x_i}}\right) + (1-y_i) log\left(1-\frac{1}{1+e^{-w^T x_i}}\right)\right] + \frac{\lambda}{2}\|w\|^2 \tag{1}$$

Рекомендуется ознакомиться с выкладкой ниже.

Информация:
Документ оформлен согласно этому заданию.
Краткое содержание задания:

I. Двухклассовая логистическая регрессия

Прежде чем начать работу, следует ввести некоторые обозначения, которые будут использоваться в дальнейшем. Все они должны быть привычными, используемыми постоянно в теоретических выкладках.

w - вектор весов объекта
x - матрица значений признаков объектов
y - истинная целевая переменная
m - кол-во объектов
$L(w)$ - функционал ошибки

i. Градиент

Так как мы решаем задачу бинарной классификации, то множество значений, которые принимает целевая переменная y состоит из 2 цифр — {1, 0}. Это очень важный элемент, которые мы будем использовать в дальнейшем, поэтому стоит его запомнить. В формуле (1) заметим сигмоидную функцию, которая дает вероятность класса. Введем функцию.

$$g(z) = \frac{1}{z + e^{-z}} \tag{2}$$

$$h_w(x) = g(w^T x) = \frac{1}{z + e^{-z}}$$

Функция (2) обладает следующими свойствами, которые нетрудно доказать.

$$g\prime(z) = g(z)(1 - g(z))$$
$$g(-z) = 1 - g(z)$$

Перепишем функцию (1), используя новые обозначения

$$L(w) = \frac{1}{m}\sum_{i=1}^{m}\Big[y_i log(g(w^T x)) + (1 - y_i) log(1 - g(w^T x))\Big] + \frac{\lambda}{2}\|w\|^2 \tag{3}$$

Используя эти обозначения и свойства сигмоидной функции, приступим к нахождению производной. Предположим, что у нас есть только один объект с вектором признаков x_i и одно значение целевой переменной y_i. Продифференцируем функцию (3) по j-тому значению вектора весов w. Дифференцировать будем без члена регуляризации, допишем его позднее.

$$\frac{\partial}{\partial w_j} L(w) = -\left[\frac{y_i}{g(w^T x)} - (1 - y_i)\left(\frac{1}{1 - g(w^T x)}\right)\right]\frac{\partial g(w^T x)}{\partial w_j} =$$

$$= -\left[\frac{y_i}{g(w^T x)} - (1 - y_i)\left(\frac{1}{1 - g(w^T x)}\right)\right] g(w^T x)(1 - g(w^T x))\frac{\partial w^T x}{\partial w_j} =$$

$$= -\big[y_i - y_i g(w^T x) - g(w^T x) + y_i g(w^T x)\big] x_j =$$

$$= \big[h_w(x) - y_i\big] x_j$$

Для случая из m объектов получим:

$$\frac{\partial}{\partial w_j} L(w) = \frac{1}{m}\sum_{i=1}^{m}\big[h_w(x_i) - y_i\big] x_{ij} \tag{4}$$

Введем новые обозначения в терминах векторов и матриц.

$X \in \mathbb{R}^{m \times n}$ — матрица объекты-признаки
$y \in \mathbb{R}^{m \times 1}$ — вектор целевых переменных
$w \in \mathbb{R}^{1 \times n}$ — вектор весов
Используя эти обозначения введем матричное представление функции (4)

$$\nabla L(w) = \frac{1}{m} X^T \big[g(Xw) - y\big]$$

Добавим член регуляризации и получим формулу (5).

$$\nabla L(w) = \frac{1}{m} X^T \big[g(Xw) - y\big] + \lambda w \tag{5}$$

Сам функционал логистической регрессии, формула (1), можно представить в такой матричной форме.

$$L(w) = -\frac{1}{m}(1, 1, ..., 1) log\big(-(2y - 1) \circ g(Xw)\big) + \frac{\lambda}{2}\|w\|^2 \tag{6}$$

ii. Гессиан

Выше мы получили функцию (4). Теперь снова продифференцируем ее, в предположении, что у нас один объект в выборке и нет регуляризации.

$$\frac{\partial}{\partial w_j \partial w_j^T} L(w) = \frac{\partial}{\partial w_j^T} \big[g(w^T x_i) - y_i\big] x_i = x_i x_i^T g(w^T x_i)(1 - g(w^T x_i)) \tag{7}$$

Для m объектов и регуляризации формула(7) выглядит так:

$$\nabla^2 L(w) = \frac{1}{m} \sum_{i=1}^{m} \Big[x_i x_i^T g(w^T x_i)(1 - g(w^T x_i)) \Big] - \lambda I \tag{8}$$

Заметим, что матрицу $g(w^T x_i)(1-g(w^T x_i))$ можно заменить диагональной матрицей, на главной диагонали которой, будут располагаться элементы $g(w^T x_i)(1-g(w^T x_i))$. Назовем эту диагональную матрицу буквой Z.

Таким образом, мы получаем матричную форму функции (8).

$$\nabla^2 L(w) = \frac{1}{m} X^T Z X - \lambda I \tag{9}$$

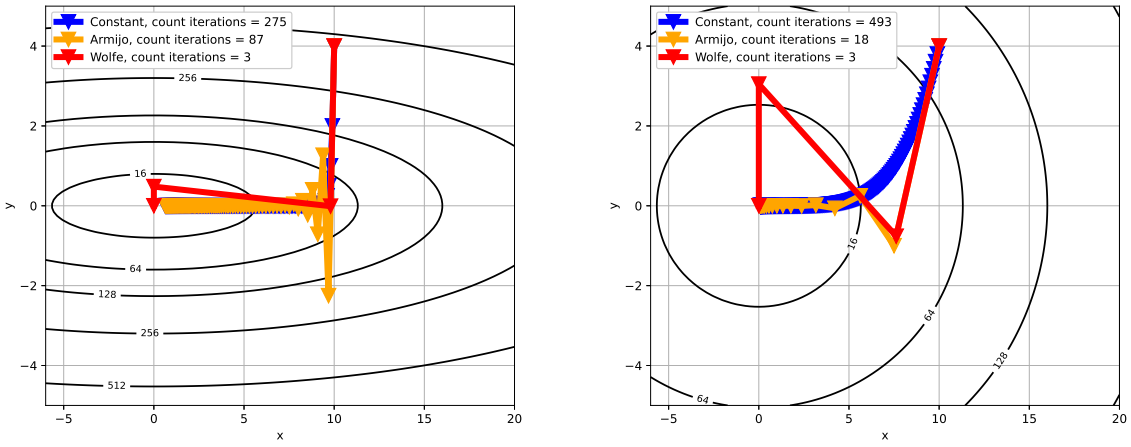
II. Эксперименты

В этой главе мы приступим к оценке реализованных алгоритмов градиентного спуска. Начнем с анализа вариантов линейного поиска шага и закончим логистической регрессией.

i. Траектория градиентного спуска на квадратичной функции

Начнем с *двумерной* квадратичной функции $f(x) = \frac{1}{2}\langle Xw, w\rangle - \langle y, w\rangle$, где $X \in \mathbb{S}^n_{++}$, $b \in \mathbb{R}^n$. Придумаем 4 функции, поведения спуска на которых, будет отличаться.

Обусловленность функции. Проанализируем график обусловленности матрицы ниже.

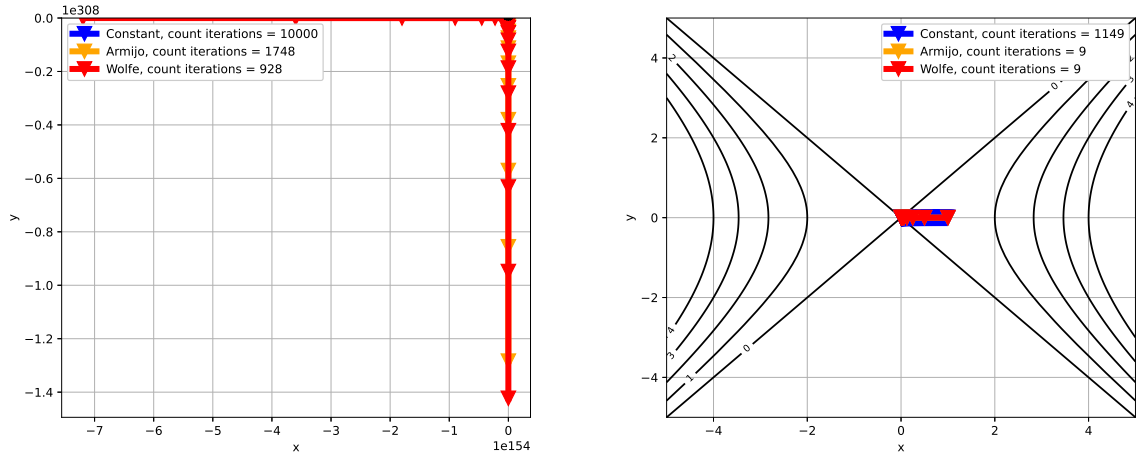


(a) $X = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 50 \end{smallmatrix}\right)$, $w_0 = (10.0, 4.0)$

(b) $X = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 5 \end{smallmatrix}\right)$, $w_0 = (10.0, 4.0)$

Рис. 1: Обусловленность матрицы

Можно заметить, что, чем ниже обусловленность матрицы, тем больше итераций делает спуск с постоянной длиной шага, а спуску методом Армихо, наоборот, требуется в 3 раза меньше операций. Также траектория спуска с постоянным шагом стала более плавной, на Рис.1 (a), мы видим, как направление спуска резко сменилось на $90°$, а на том же рисунке с литерой (b), траектория меняется более плавно. Траектория Армихо в среднем не изменилась, также видны резкие изменения траектории, но ближе к точке минимума траектория становится более плавной и совпадает со спуском с постоянным шагом. Траектория Вульфа стала более криволинейной, каждый шаг делается практически с разворотом на $90°$, но благодаря подбору шага этому методу требуется меньше итераций, чем двум другим.



(a) $X = \left(\begin{smallmatrix} 0.5 & 0 \\ 0 & -0.5 \end{smallmatrix}\right)$, $w_0 = (3.0, -3.0)$

(b) $X = \left(\begin{smallmatrix} 0.5 & 0 \\ 0 & -0.5 \end{smallmatrix}\right)$, $w_0 = (1.0, 0.0)$

Рис. 2: Выбор начальной точки

На Рис.2 ищется локальный минимум функции $f(x) = \frac{1}{2}x^2 - \frac{1}{2}y^2$ — так называемая, седловая функция. Ее особенность заключается в том, что найти ее минимум можно тогда и только тогда, когда начинаем с точки $w_{init} = (1.0, 0.0)$, это можно легко доказать теоретически. Также заметим, что на Рис.2 (a) константный метод дошел до предела итераций, но так и не достиг $-\infty$ по значению функции. Армихо и Вульф достигли этого нижнего предела, сработал критерий останова, причем метод Вульфа снова обогнал другие методы спуска. На Рис. 2 (b) все 3 метода сходятся в одну точку, но спуску с постоянным шагом требуется куда больше итераций для этого.