

Методы оптимизации в машинном обучении

Практическое задание #1

Рожин Андрей

НИУ ВШЭ — 14 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.

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

Рекомендуется ознакомиться с выкладкой ниже.



Информация:

Документ оформлен согласно [этому заданию](#).

Краткое содержание задания:

1. Алгоритм спуска.
 - 1.1 Общая концепция
 - 1.2 Критерий останова
 - 1.3 Линейный поиск - условие Армихо, сильное условие Вульфа.
 - 1.4 Градиентный спуск.
 - 1.5 Метод Ньютона
 - 1.6 Оптимизация вычислений
2. Модели
 - 2.1 Двухклассовая логистическая регрессия.
 - 2.2. Разностная проверка градиента и гессиана
3. Эксперименты.
 - 3.1 Оценка реализованных алгоритмов.

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.

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

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

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

$$\begin{aligned} g'(z) &= g(z)(1 - g(z)) \\ g(-z) &= 1 - g(z) \end{aligned}$$

Перепишем функцию (1), используя новые обозначения

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

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.

$$\begin{aligned} \frac{\partial}{\partial w_j} L(w) &= - \left[\frac{y_i}{g(w^T x)} - (1 - y_i) \left(\frac{1}{1 - g(w^T x)} \right) \right] \frac{\partial g(w^T x)}{\partial w_j} = \\ &= - \left[\frac{y_i}{g(w^T x)} - (1 - y_i) \left(\frac{1}{1 - g(w^T x)} \right) \right] g(w^T x)(1 - g(w^T x)) \frac{\partial w^T x}{\partial w_j} = \\ &= - [y_i - y_i g(w^T x) - g(w^T x) + y_i g(w^T x)] x_j = \\ &= [h_w(x) - y_i] x_j \end{aligned}$$

Для случая из m объектов получим:

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

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)

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

Добавим член регуляризации и получим формулу (5).

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

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

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

ii. Гессиан

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

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

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

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

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).

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