Линейные модели и методы оптимиизации

1 Регрессия

$$\mathcal{D}^l = \{x_i, y_i\}_{i=1}^l, \quad x_i \in \mathbb{R},\ y_i \in \mathbb{R}. \tag{1.1}$$

Пусть имеется некоторая обучающая выборка $\mathcal{D}^l$ размера l по которой мы хотим построить некоторую модель.

Определение 1.1: Линейной моделью регрессии назовем функцию $\mathbf{a}(x, \mathbf{w})$ из (1.2), которая зависит от некоторого неизвестного параметра $\mathbf{w} \in \mathbb{R}^n$.

$$\mathbf{a}(x, \mathbf{w}) = \sum_{j=1}^{n} w_j f_j(x), \tag{1.2}$$

где $f_j(x)$ — это функция которая по заданному объекту x выдает j-й признак этого объекта.

Заметим, что вектор параметров $\mathbf{w}$ является неизвестным и его нужно найти по заданной выборке $\mathcal{D}^l$

Определение 1.2: Введем понятия функции потерь модели $\mathbf{a}$ на некотором объекте $(x, y) \in \mathcal{D}^l$ следующиим образом:

$$\mathcal{L}(\mathbf{w}, (x, y)) = (\mathbf{a}(x, \mathbf{w}) - y)^2. \tag{1.3}$$

Определение 1.3: Введем понятия функции потерь модели регрессии $\mathbf{a}$ на выборке $\mathcal{D}^l$ следующим образом:

$$\mathcal{Q}(\mathbf{w}) = \sum_{j=1}^{l} \mathcal{L}(\mathbf{w}, (x_j, y_j)). \tag{1.4}$$

Теперь, мы можем сформулировать задачу машинного обучения, как поиск $\mathbf{w}$, такого что $\mathcal{Q}(\mathbf{w})$ является минимальным. Формальная запись этого факта, это:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^n} \mathcal{Q}(\mathbf{w}). \tag{1.5}$$

Тогда после нахождения такого $\hat{\mathbf{w}}$, мы получаем обученную линейную модель.

2 Классификация

$$\mathcal{D}^l = \{x_i, y_i\}_{i=1}^l, \quad x_i \in \mathbb{R},\ y_i \in \{-1, +1\}. \tag{2.1}$$

Пусть имеется некоторая обучающая выборка $\mathcal{D}^l$ размера l по которой мы хотим построить некоторую модель.

Определение 2.1: Линейной моделью классификации назовем функцию $\mathbf{a}(x, \mathbf{w})$ из (2.2), которая зависит от некоторого неизвестного параметра $\mathbf{w} \in \mathbb{R}^n$.

$$\mathbf{a}(x, \mathbf{w}) = \operatorname{sign} \sum_{j=1}^{n} w_j f_j(x), \tag{2.2}$$

где $f_j(x)$ — это функция которая по заданному объекту x выдает j-й признак этого объекта.

Определение 2.2: Введем понятия функции потерь модели классификации $\mathbf{a}$ на некотором объекте $(x, y) \in \mathcal{D}^l$ следующим образом:

$$\mathcal{L}(\mathbf{w}, (x, y)) = -\mathbf{a}(x, \mathbf{w}) \cdot y. \tag{2.3}$$

Определение 2.3: Введем понятия функции потерь модели регрессии $\mathbf{a}$ на выборке $\mathcal{D}^l$ следующим образом:

$$\mathcal{Q}(\mathbf{w}) = \sum_{j=1}^{l} \mathcal{L}(\mathbf{w}, (x_j, y_j)). \tag{2.4}$$

Теперь аналогично задачи регрессии сформулируем оптимизационную задачу:

$$\hat{\mathbf{w}} = \arg\min_{\mathbf{w} \in \mathbb{R}^n} \mathcal{Q}(\mathbf{w}). \tag{2.5}$$

3 Решение оптимизационной задачи

3.1 Производная

$$f'(x) = \lim_{\Delta x \to 0} \frac{\Delta f}{\Delta x}. \tag{3.1.1}$$

Будем использовать свойство знака производной. Знак производной указывает на то растет функция в этой точке или убывает, это свойство прямо следует из определения производной. Покажем этот факт:

$$f'(x_0) = \lim_{\Delta x \to 0,\ \Delta x > 0} \frac{f(x_0 + \Delta x) - f(x_0)}{\Delta x}, \tag{3.1.2}$$

Из уравнения (3.2) видно, что знак производной в точке x_0 равен знаку $f(x_0 + \Delta x) - f(x_0)$, что и требовалось показать.

Для примера из рис. 1 $y'(6) = 13$. Тогда с этого следует, что функция в этой точке растет при увеличения x, тогда для того, чтобы найти минимальное значение y, нужно уменьшить x.

Вот мы пришли к выводу, что если у нас есть некоторая функция одного переменного, то для того, чтобы найти ее минимум нужно менять x в противоположном направлении к знаку производной.

На этом базируется следующий итеративный подход к нахождению минимума функции.

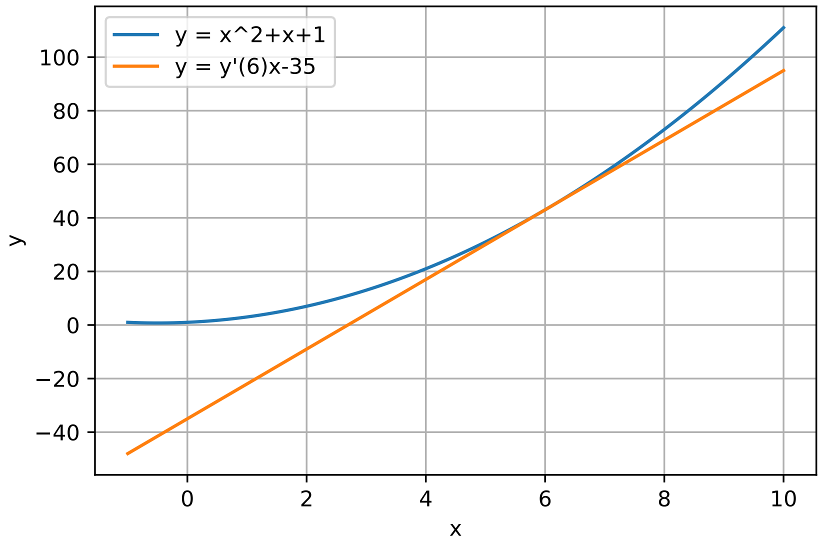

Рис. 1: График функции и касательная в точке**a**

Определение 3.1.1: Итеративный процесс обозначает, что мы делаем что-то шаг за шагом.

Рассмотрим следующий итеративный процесс, для нахождения минимума одномерной функции $f(x)$ с областью определения D_f.

1. Пусть имеется $x_0 \in D_f$ — некоторая точка из области определения функции. 2. Пересчитывать новую точку будем по формуле:

$$x_{n+1} = x_n - \alpha \cdot f'(x_n), \tag{3.1.3}$$

где α некоторое значение — шаг который мы делаем, он может быть как и постоянным так и переменным. Мы будем пока считать его постоянным числом, например 0.0001.

Мы научились находить минимум функции от скаляра. Но что же делать, для функции от вектора, которой является $\mathcal{Q}(\mathbf{w})$.

3.2 Градиент

Определение 3.2.1: Частной производной функции многих переменных $f(\mathbf{x})$ по x_j назовем производную функции $f'(x_j)$ считая все остальными переменные константой. Частная производная обозначается следующим образом:

$$\frac{\partial f}{\partial x_j} = f'_j(x_j), \tag{3.2.1}$$

где $f'_j(x)$ — это функция одной переменной, где все переменные кроме j-й фиксированы.

Определение 3.2.2: Градиентом функции $f(\mathbf{x})$ называется вектор $\nabla f(\mathbf{x})$ элементы которого, это частные производные функции $f(\mathbf{x})$.

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{x_1} \\ \frac{\partial f}{x_2} \\ \cdots \\ \frac{\partial f}{x_n} \end{bmatrix}, \tag{3.2.2}$$

где $x_1,\ x_2, \cdots,\ x_n$ — компоненты вектора $\mathbf{x}$.

По аналогии с функцией одного переменного можно определить итеративный процесс нахождения минимума функции многих переменных.

1. Пусть имеется $\mathbf{x}^0 \in D_f$ — некоторая точка из области определения функции $f(\mathbf{x})$. 2. Пересчитывать новую точку будем по формуле:

$$x^{n+1} = x^n - \alpha \cdot \nabla f(\mathbf{x}^n), \tag{3.2.3}$$

где α некоторое значение — шаг который мы делаем, он может быть как и постоянным так и переменным. Мы будем пока считать его постоянным числом, например 0.0001.

Как видно все изменения в итеративной формуле это производная на градиент.

3.3 Пример вычисления градиентов:

$$f(x_1, x_2) = x_1^2 + x_2^2 \Rightarrow \nabla f(x_1, x_2) = \begin{bmatrix} \frac{\partial f}{x_1} \\ \frac{\partial f}{x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 2x_2 \end{bmatrix}. \tag{3.3.1}$$

$$f(x_1, x_2) = e^{x_1} + e^{x_2} + x_1 x_2 \Rightarrow \nabla f(x_1, x_2) = \begin{bmatrix} \frac{\partial f}{x_1} \\ \frac{\partial f}{x_2} \end{bmatrix} = \begin{bmatrix} e^{x_1} + x_2 \\ e^{x_2} + x_1 \end{bmatrix}. \tag{3.3.2}$$

Список литературы

[1] *Воронцов К. В.* Машинное обучение // Годовой курс кафедры «Интеллектуальные системы» Москва, 2018. http://www.machinelearning.ru/wiki/index.php?title=Vokov