

Машинное обучение, ФКН ВШЭ

Семинар №3

1 Линейные модели

В этом семинаре речь пойдёт о таком базовом инструменте анализа данных, как линейная регрессия. Несмотря на простоту, материал важен для изучения, так как содержит интересные приёмы анализа данных.

В качестве разминки предлагается самостоятельно вывести формулу линейной регрессии для одномерного случая. Задача ставится следующим образом:

Задача 1.1. Дана одномерная обучающая выборка $\{(x_i, y_i)\}_{i=1}^{\ell}$, $x_i, y_i \in \mathbb{R}, i = \overline{1, \ell}$. Найдите параметры k, b , минимизирующие MSE на выборке для модели $a(x) = kx + b$.

$$RMSE = \sqrt{\frac{1}{\ell} \sum_{i=1}^{\ell} (y_i - a(x_i))^2} \rightarrow \min_{k, b}$$

Для нахождения решения в многомерном случае нам потребуется изучить некоторые приёмы векторного дифференцирования.

§1.1 Векторное дифференцирование

Иногда при взятии производных по вектору или от вектор-функций удобно оперировать матричными операциями. Это сокращает запись и упрощает вывод формул. Введём следующие определения:

- При отображении вектора в число $f(x) : \mathbb{R}^n \rightarrow \mathbb{R}$

$$\nabla_x f(x) = \left[\frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right]^T.$$

- При отображении матрицы в число $f(A) : \mathbb{R}^{n \times m} \rightarrow \mathbb{R}$

$$\nabla_A f(A) = \left(\frac{\partial f}{\partial A_{ij}} \right)_{i,j=1}^{n,m}.$$

Мы хотим оценить, как функция изменяется по каждому из аргументов по отдельности. Поэтому производной функции по вектору будет вектор, по матрице — матрица. Теперь поупражняемся в дифференцировании:

Задача 1.2. Пусть $a \in \mathbb{R}^n$ — вектор параметров, а $x \in \mathbb{R}^n$ — вектор переменных. Необходимо найти производную их скалярного произведения по вектору переменных $\nabla_x a^T x$.

Решение.

$$\frac{\partial}{\partial x_i} a^T x = \frac{\partial}{\partial x_i} \sum_j a_j x_j = a_i,$$

поэтому $\nabla_x a^T x = a$.

Заметим, что $a^T x$ — это число, поэтому $a^T x = x^T a$, следовательно,

$$\nabla_x x^T a = a.$$

■

Задача 1.3. Пусть теперь $A \in \mathbb{R}^{n \times n}$. Необходимо найти $\nabla_x x^T A x$.

Решение.

$$\begin{aligned} \frac{\partial}{\partial x_i} x^T A x &= \frac{\partial}{\partial x_i} \sum_j x_j (A x)_j = \frac{\partial}{\partial x_i} \sum_j x_j \left(\sum_k a_{jk} x_k \right) = \frac{\partial}{\partial x_i} \sum_{j,k} a_{jk} x_j x_k = \\ &= \sum_{j \neq i} a_{ji} x_j + \sum_{k \neq i} a_{ik} x_k + 2a_{ii} x_i = \sum_j a_{ji} x_j + \sum_k a_{ik} x_k = \sum_j (a_{ji} + a_{ij}) x_j. \end{aligned}$$

Поэтому $\nabla_x x^T A x = (A + A^T)x$.

■

Задача 1.4. Пусть $A \in \mathbb{R}^{n \times n}$. Необходимо найти $\nabla_A \det A$.

Решение. Воспользуемся теоремой Лапласа о разложении определителя по строке:

$$\frac{\partial}{\partial A_{ij}} \det A = \frac{\partial}{\partial A_{ij}} \left[\sum_k (-1)^{i+k} A_{ik} M_{ik} \right] = (-1)^{i+j} M_{ij},$$

где M_{ik} — дополнительный минор матрицы A . Также вспомним формулу для элементов обратной матрицы

$$(A^{-1})_{ij} = \frac{1}{\det A} (-1)^{i+j} M_{ji}.$$

Подставляя выражение для дополнительного минора, получаем ответ $\nabla_A \det A = (\det A) A^{-T}$.

■

Задача 1.5. Пусть $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times n}$. Необходимо найти $\nabla_A \text{tr}(AB)$.

Решение.

$$\frac{\partial}{\partial A_{ij}} \text{tr}(AB) = \frac{\partial}{\partial A_{ij}} \sum_k (AB)_{kk} = \frac{\partial}{\partial A_{ij}} \sum_{k,l} A_{kl} B_{lk} = B_{ji}.$$

То есть, $\nabla_A \text{tr}(AB) = B^T$.

■

Задача 1.6. Пусть $x \in \mathbb{R}^n$, $A \in \mathbb{R}^{n \times m}$, $y \in \mathbb{R}^m$. Необходимо найти $\nabla_A x^T A y$.

Решение. Воспользовавшись циклическим свойством следа матрицы (для матриц подходящего размера):

$$\text{tr}(ABC) = \text{tr}(BCA) = \text{tr}(CAB)$$

и результатом предыдущей задачи, получаем

$$\nabla_A x^T A y = \nabla_A \text{tr}(x^T A y) = \nabla_A \text{tr}(A y x^T) = x y^T.$$

■

§1.2 Решение задачи регрессии для многомерного случая

Вспомним, зачем мы хотели научиться дифференцировать. В общем случае мы имеем выборку $\{(x_i, y_i)\}_{i=1}^{\ell}$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ $i = \overline{1, \ell}$, и хотим найти наилучшие параметры модели $a(x) = \langle w, x \rangle$ с точки зрения минимизации функции ошибки

$$Q(w) = (y - Xw)^T (y - Xw).$$

Здесь $X \in \mathbb{R}^{\ell \times d}$ — матрица «объекты-признаки» для обучающей выборки, $y \in \mathbb{R}^{\ell}$ — вектор значений целевой переменной на обучающей выборке, $w \in \mathbb{R}^d$ — вектор параметров. Выпишем градиент функции ошибки по w :

$$\begin{aligned} \nabla_w Q(w) &= \nabla_w [y^T y - y^T X w - w^T X^T y + w^T X^T X w] = \\ &= 0 - X^T y - X^T y + (X^T X + X^T X) w = 0. \end{aligned}$$

Таким образом, искомый вектор параметров выражается как

$$w = (X^T X)^{-1} X^T y.$$

Заметим, что это общая формула, и нет необходимости выводить формулу для регрессии вида $a(x) = Xw + w_0$, т.к. мы всегда можем добавить признак (столбец матрицы X), который всегда будет равен 1, и по уже выведенной формуле найдём параметр w_0 .

Покажем, почему найденная точка — точка минимума, если матрица $X^T X$ обратима. Из курса математического анализа мы знаем, что если матрица Гессе функции положительно определена в точке, градиент которой равен нулю, то эта точка является локальным минимумом.

$$\nabla^2 Q(w) = 2X^T X.$$

Необходимо понять, является ли матрица $X^T X$ положительно определённой. Запишем определение положительной определённости матрицы $X^T X$:

$$z^T X^T X z > 0, \forall z \in \mathbb{R}^d, z \neq 0.$$

Видим, что тут записан квадрат нормы вектора Xz , то есть это выражение будет не меньше нуля. В случае, если матрица X имеет «книжную» ориентацию (строк не меньше, чем столбцов) и имеет полный ранг (нет линейно зависимых столбцов), то вектор Xz не может быть нулевым, а значит выполняется

$$z^T X^T X z = \|Xz\|^2 > 0, \forall z \in \mathbb{R}^d, z \neq 0.$$

То есть $X^T X$ является положительно определённой матрицей. Также, по критерию Сильвестра, все главные миноры (в том числе и определитель) положительно определённой матрицы положительны, а, следовательно, матрица $X^T X$ обратима, и решение существует. Если же строк оказывается меньше, чем столбцов, или X не является полноранговой, то $X^T X$ необратима и решение w определено неоднозначно.

2 Градиентный спуск

Ситуации, когда нам удаётся найти решение оптимизационной задачи в явном виде, — большая удача. В общем случае оптимизационные задачи можно решать итерационно с помощью градиентных методов (или же методов, использующих как градиент, так и информацию о производных более высокого порядка). Для понимания работы этих методов давайте ознакомимся со свойствами градиента.

§2.1 Градиент и его свойства

Антиградиент $(-\nabla f)$ является направлением наискорейшего убывания функции в заданной точке. Это ключевое свойство градиента, обосновывающее его использование в методах оптимизации. Докажем эквивалентное утверждение.

Утв. 1. Градиент является направлением наискорейшего роста функции.

Доказательство.

Пусть $v \in \mathbb{R}^d$ — произвольный вектор, лежащий на единичной сфере: $\|v\| = 1$. Пусть $x_0 \in \mathbb{R}^d$ — фиксированная точка пространства. Скорость роста функции в точке x_0 вдоль вектора v характеризуется производной по направлению $\frac{\partial f}{\partial v}$:

$$\frac{\partial f}{\partial v}(x_0) = \frac{d}{dt} f(x_{0,1} + tv_1, \dots, x_{0,d} + tv_d) \Big|_{t=0}.$$

Из курса математического анализа известно, что данную производную сложной функции можно переписать следующим образом:

$$\frac{\partial f}{\partial v}(x_0) = \sum_{j=1}^d \frac{\partial f}{\partial x_j}(x_0) \frac{d}{dt}(x_{0,j} + tv_j) = \sum_{j=1}^d \frac{\partial f}{\partial x_j}(x_0) v_j = \langle \nabla f(x_0), v \rangle.$$

Распишем скалярное произведение:

$$\langle \nabla f(x_0), v \rangle = \|\nabla f(x_0)\| \|v\| \cos \varphi = \|\nabla f(x_0)\| \|v\| \cos \varphi,$$

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. ■

Напоследок докажем ещё одно фундаментальное свойство градиента.

Утв. 2. Градиент ортогонален линиям уровня.

Доказательство.

Пусть x_0 — некоторая точка, $S(x_0) = \{x \in \mathbb{R}^d \mid f(x) = f(x_0)\}$ — соответствующая линия уровня. Разложим функцию в ряд Тейлора на этой линии в окрестности x_0 :

$$f(x_0 + \varepsilon) = f(x_0) + \langle \nabla f(x_0), \varepsilon \rangle + o(\|\varepsilon\|),$$

где $x_0 + \varepsilon \in S(x_0)$. Поскольку $f(x_0 + \varepsilon) = f(x_0)$ (как-никак, это линия уровня), получим

$$\langle \nabla f(x_0), \varepsilon \rangle = o(\|\varepsilon\|).$$

Поделим обе части на $\|\varepsilon\|$:

$$\left\langle \nabla f(x_0), \frac{\varepsilon}{\|\varepsilon\|} \right\rangle = o(1).$$

Устремим $\|\varepsilon\|$ к нулю. При этом вектор $\frac{\varepsilon}{\|\varepsilon\|}$ будет стремиться к касательной к линии уровня в точке x_0 . В пределе получим, что градиент ортогонален этой касательной. ■