Машинное обучение, ФКН ВШЭ
Семинар №3

1 Устойчивость среднеабсолютной ошибки к выбросам

На лекциях были рассмотрены различные функционалы ошибки, которые могут быть использованы в задаче регрессии, — в частности, были рассмотрены средняя абсолютная и квадратичная ошибки, а также отмечался тот факт, что первая более устойчива к выбросам по сравнению со второй. Для демонстрации этого факта давайте вычислим, чему равно оптимальное (с точки зрения каждого из функционалов на некоторой выборке X) константное предсказание $a(x) = C$.

Задача 1.1. *Найдите C, минимизирующий среднеквадратичную ошибку.*

Решение.

$$MSE(C) = \frac{1}{\ell}\sum_{i=1}^{\ell}(C - y_i)^2$$

$$\frac{\partial MSE(C)}{\partial C} = \frac{1}{\ell}\sum_{i=1}^{\ell}2(C - y_i) = 2C - \frac{1}{\ell}\sum_{i=1}^{\ell}2y_i = 0$$

$$C = \frac{1}{\ell}\sum_{i=1}^{\ell}y_i$$

.

■

Задача 1.2. *Найдите C, минимизирующий среднюю абсолютную ошибку.*

Решение.

$$MAE(C) = \frac{1}{\ell}\sum_{i=1}^{\ell}|C - y_i|$$

Покажем, что минимум MAE достигается при $C = median(y_1, \dots, y_\ell) = m$. Рассмотрим $C < m$.

$$|y_i - C| - |y_i - m| = \begin{cases} C - m, y_i < m \\ -(C + m - 2y_i), C \leqslant y_i \leqslant m \\ -(C - m), y_i > m \end{cases}$$

$$|y_i - C| - |y_i - m| \geqslant -(C - m) + 2(C - m)[y_i \leqslant m]$$

Суммируем по i:

$$\ell MAE(C) - \ell MAE(m) \geqslant -\ell(C - m) + 2(C - m)\sum_{i=1}^{\ell}[y_i \leqslant m]$$

Так как m — медиана, $\sum_{i=1}^{\ell}[y_i \leqslant m] \geqslant \frac{\ell}{2}$. Тогда

$$\ell MAE(C) - \ell MAE(m) \geqslant -\ell(C - m) + 2(C - m)\frac{\ell}{2} = 0.$$

Итак, для $C < m$ выполняется $MAE(C) \geqslant MAE(m)$. Аналогично показывается, что при для $C > m$ выполняется $MAE(C) \geqslant MAE(m)$.

■

Можем видеть, что оптимальной константной для MAE является медиана, которая является более устойчивой к выбросам по сравнению со средним арифметическим. Несмотря на это, MSE обладает своими достоинствами — в частности, для нее можно выписать аналитическое решение в случае линейной регрессии, а также её можно оптимизировать напрямую при помощи градиентного спуска, в отличие от MAE, которая не является дифференцируемой по w (подробнее про способы оптимизации таких функционалов будет рассказано в курсе «Методы оптимизации»).

2 Решение задачи линейной регрессии

Ранее мы упоминали, что для задачи линейной регрессии с функционалом MSE оптимальное значение вектора весов w^* можно выписать в явном виде.

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

Задача 2.1. *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* $a(x) = kx + b$.

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

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

§2.1 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 $f(x) : \mathbb{R}^n \to \mathbb{R}$

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

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

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

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 2.2. *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 2.3. *Пусть теперь $A \in \mathbb{R}^{n\times n}$. Необходимо найти $\nabla_x x^T Ax$.*

Решение.

$$\frac{\partial}{\partial x_i} x^T Ax = \frac{\partial}{\partial x_i} \sum_j x_j (Ax)_j = \frac{\partial}{\partial x_i} \sum_j x_j \left(\sum_k a_{jk} x_k\right) = \frac{\partial}{\partial x_i} \sum_{j,k} a_{jk} x_j x_k =$$

$$= \sum_{j\neq i} a_{ji} x_j + \sum_{k\neq i} a_{ik} x_k + 2a_{ii} x_i = \sum_j a_{ji} x_j + \sum_k a_{ik} x_k = \sum_j (a_{ji} + a_{ij}) x_j.$$

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

■

Задача 2.4. *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 2.5. *Пусть $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times n}$. Необходимо найти $\nabla_A tr(AB)$.*

Решение.

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

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

∎

Задача 2.6. *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):

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

и результатом предыдущей задачи, получаем

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

∎

§2.2 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 определёна 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.

3 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.

§3.1 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 $\frac{\partial f}{\partial v}$:

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

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

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

Распишем скалярное произведение:

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

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, $S(x_0) = \{x \in \mathbb{R}^d \,|\, f(x) = f(x_0)\}$ — 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 стремится к касательной к линии уровня в точке x_0. В пределе получим, что градиент ортогонален этой касательной.

■