Линейная регрессия

Минимум квадратичной функции

Производные, градиентный спуск и СЛАУ

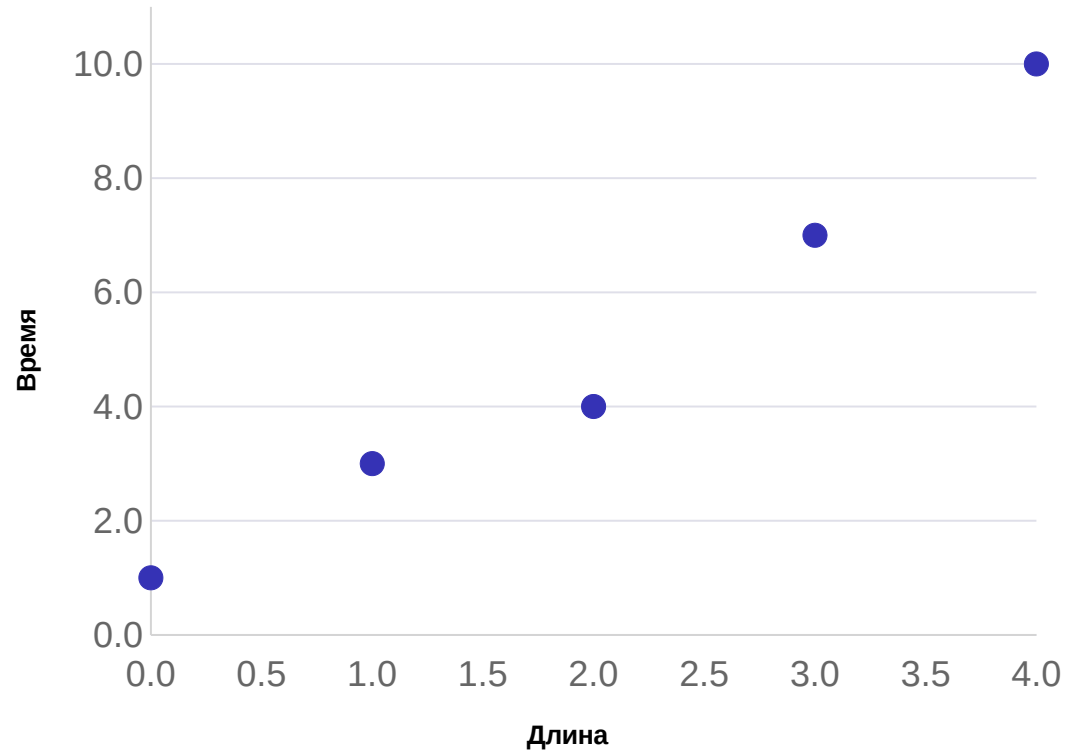
Банцевич Максим

Данные и предсказание

По длине маршрута хотим предсказать время в пути.

Один маршрут — объект. Его длина — признак, время — целевая переменная.

Выборка содержит известные пары входов и ответов. На новых маршрутах ответ ещё неизвестен.



Регрессия предсказывает числовую величину. На протяжении занятия используем одни и те же пять синтетических точек.

Уравнение прямой

Выберем семейство моделей, в котором прогноз зависит от одного признака:

$$\hat{y} = wx + b$$

Наклон w

При увеличении признака на единицу прогноз меняется на w.

Свободный коэффициент b

Задаёт прогноз при нулевом значении признака и сдвигает прямую вверх или вниз.

При обучении данные фиксированы. Мы подбираем параметры w и b, чтобы прямая хорошо описывала выборку.

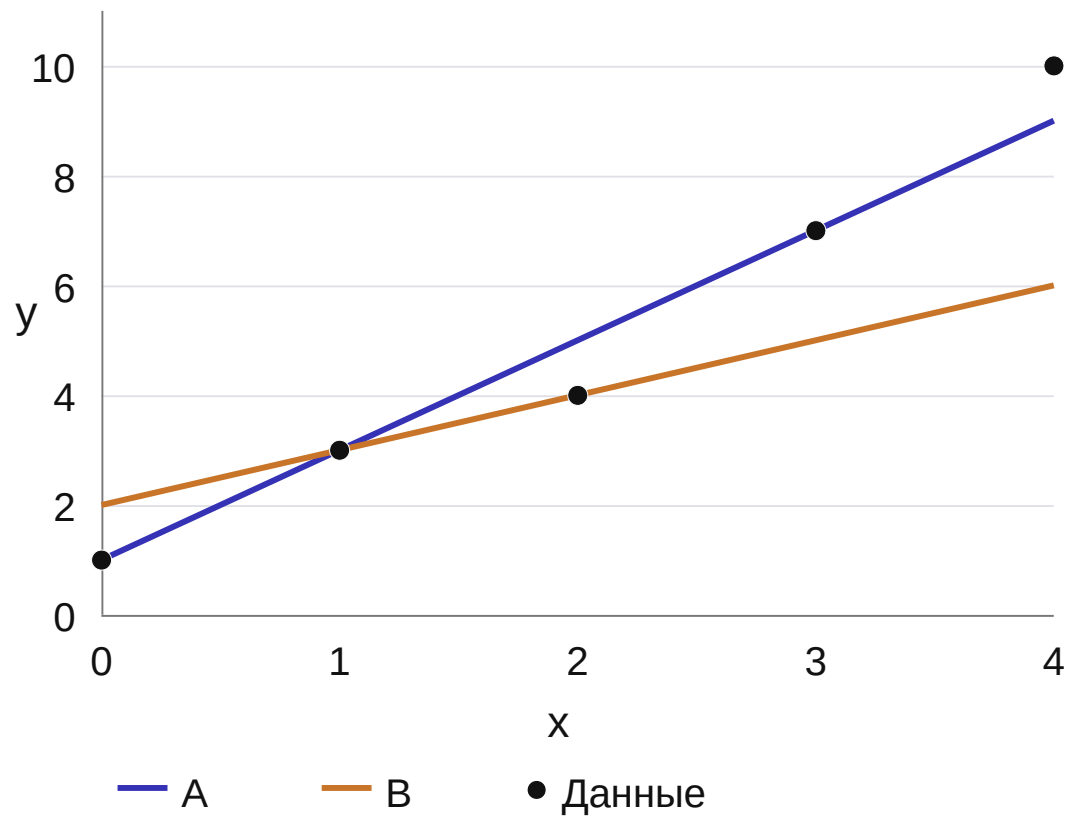
Одна модель, разные параметры

$$\text{A:} \quad \hat{y} = 2x + 1$$

$$\text{B:} \quad \hat{y} = x + 2$$

Обе модели — прямые. Изменяются только наклон и сдвиг.

Как численно сравнить их качество?



Визуального сравнения недостаточно: одна модель может лучше подходить одним точкам, другая — остальным. Нужен общий критерий ошибки.

Ошибка одного предсказания

Остаток — разность между прогнозом и известным ответом:

$$e_i = \hat{y}_i - y_i$$

Завышенный прогноз

На третьей точке модель A предсказывает 5 вместо 4. Ошибка равна +1.

Заниженный прогноз

На пятой точке модель A предсказывает 9 вместо 10. Ошибка равна −1.

Эти два промаха в сумме дают ноль. Поэтому сумма ошибок со знаком плохо отражает качество модели.

Средний квадрат ошибки

Возведём каждый остаток в квадрат и усредним по всем объектам:

$$L(w, b) = \frac{1}{n} \sum_{i=1}^{n} (wx_i + b - y_i)^2$$

Модель	Ошибки на пяти точках	MSE
A	0, 0, 1, 0, −1	0,4
B	1, 0, 0, −2, −4	4,2

Квадраты убирают знак и сильнее штрафуют большие промахи. Обучение означает поиск параметров с минимальным значением MSE.

Геометрический смысл MSE

Соберём истинные ответы и прогнозы в два вектора длины n. Их разность — вектор ошибок.

$$L(w, b) = \frac{\|\hat{\mathbf{y}} - \mathbf{y}\|_2^2}{n}$$

$$\|\hat{\mathbf{y}} - \mathbf{y}\|_2^2 = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

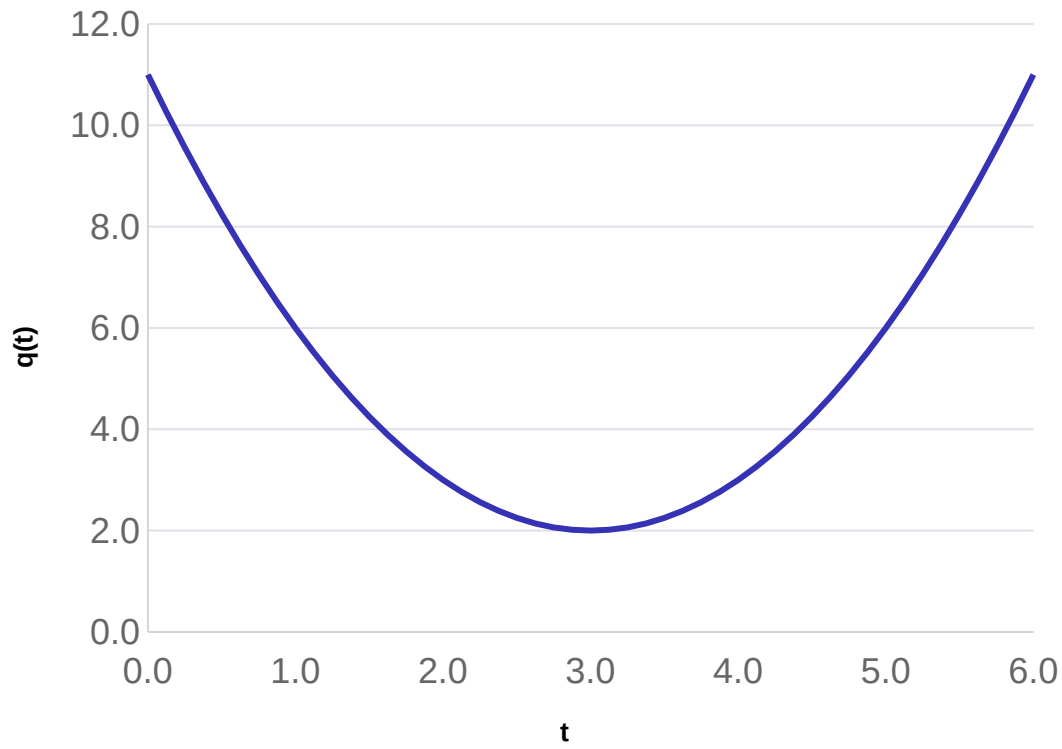
Евклидово расстояние, его квадрат и MSE достигают минимума при одних и тех же параметрах. Численные значения и градиенты этих функций различаются.

Минимум квадратичной функции

$$q(t) = t^2 - 6t + 11$$

$$q(t) = (t-3)^2 + 2$$

Квадрат неотрицателен. Наименьшее значение равно 2 и достигается при t = 3.



Выделение полного квадрата позволяет сразу найти минимум параболы с ветвями вверх.

Корни и минимум функции

Решение уравнения

Ищем точки, в которых значение функции равно нулю. У нашей параболы действительных корней нет.

Минимизация

Ищем наименьшее значение функции. У нашей параболы оно существует и равно 2.

$$q(t) = (t-3)^2 + 2$$

При обучении минимизируем ошибку по параметрам. Она может оставаться положительной из-за шума или ограниченного класса моделей.

Почему ошибка квадратична

Зафиксируем свободный коэффициент b = 1 и подставим пять известных точек в MSE:

$$L(w,1) = \frac{1}{5}\Big[(w-2)^2 + (2w-3)^2 + (3w-6)^2 + (4w-9)^2\Big]$$

$$L(w,1) = 6w^2 - 24{,}8w + 26$$

Каждый квадрат линейного выражения даёт квадрат веса. Сумма таких слагаемых остаётся квадратичной функцией.

При полном обучении меняются оба параметра. Тогда MSE — квадратичная функция точки (w, b).

График модели и график ошибки

График модели

Вход объекта и прогноз при фиксированных параметрах.

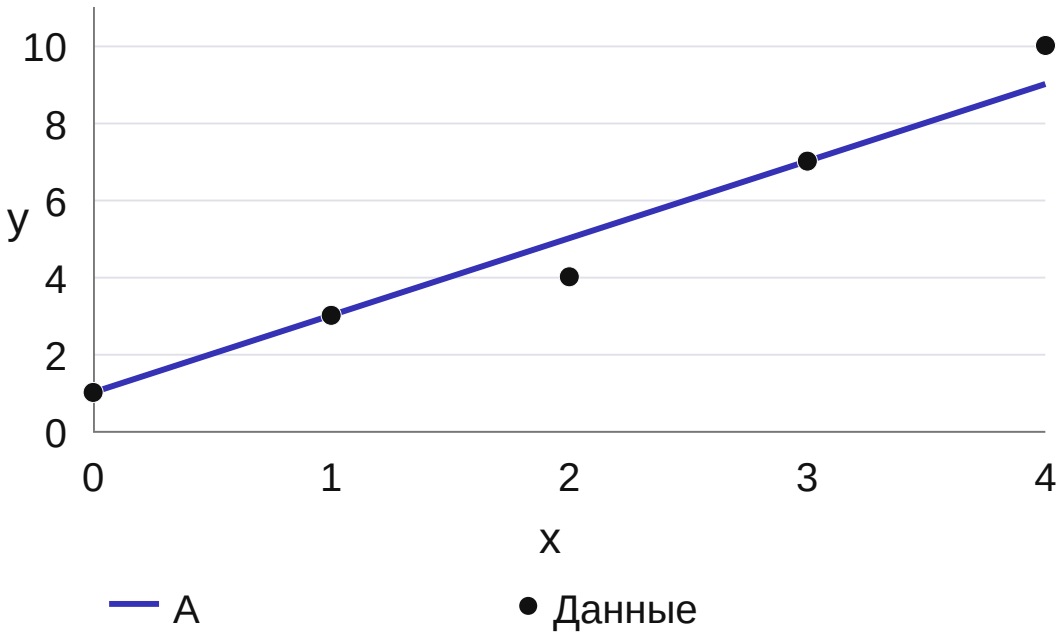
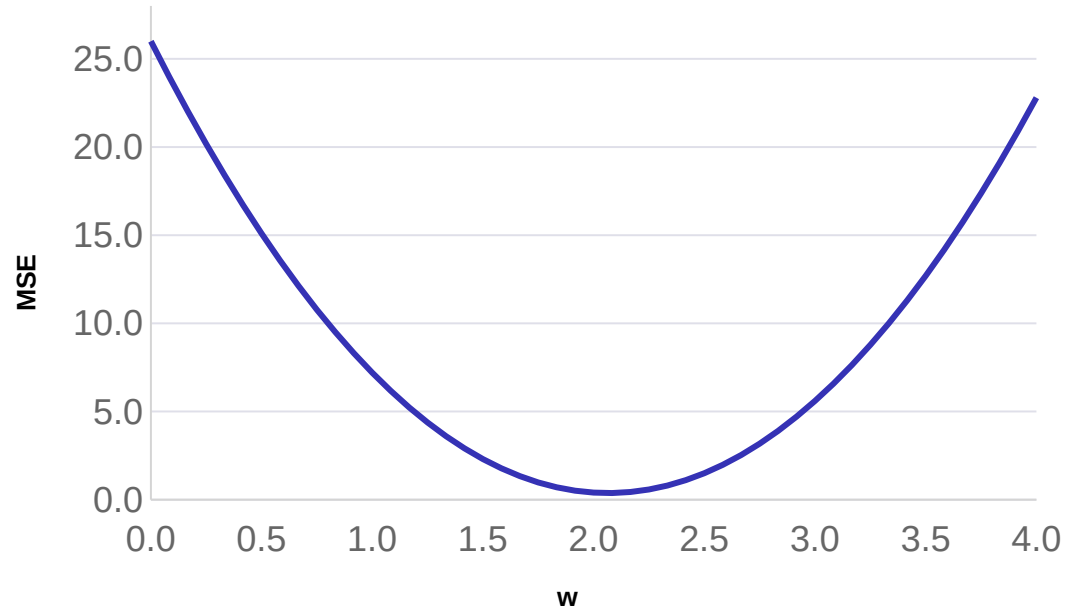


График ошибки

Параметр модели и MSE на неизменной выборке.



Слева выбираем объект и получаем прогноз. Справа выбираем вес и оцениваем ошибку всей модели. Прямая по признаку может давать параболу ошибки по весу.

Производная как локальный наклон

Производная показывает, как быстро меняется функция при малом изменении аргумента:

$$q'(t) = \lim_{h \to 0} \frac{q(t+h) - q(t)}{h}$$

Для нашей параболы раскроем квадраты, разделим прирост на h и упростим:

$$\frac{q(t+h) - q(t)}{h} = 2(t-3) + h$$

При стремлении h к нулю последнее слагаемое исчезает.

$$q'(t) = 2(t-3)$$

Знак производной и минимум

$$q'(t) = 2(t - 3)$$

Положение	**Производная**	**Как уменьшать функцию**
Слева от 3	Отрицательная	Увеличивать аргумент
В точке 3	Ноль	Минимум параболы
Справа от 3	Положительная	Уменьшать аргумент

Для параболы с ветвями вверх нулевая производная даёт минимум. Для произвольной функции одного этого условия недостаточно.

Итерационный поиск минимума

Начнём с некоторого приближения и будем многократно выполнять одно правило:

$$t_{k+1} = t_k - \eta q'(t_k)$$

Текущее значение задаёт точку вычисления производной. Положительный шаг η определяет величину изменения.

Вычитаем производную: при положительном наклоне движемся влево, при отрицательном — вправо. При достаточно малом шаге функция убывает.

Несколько шагов вручную

$$q(t) = (t-3)^2 + 2, \qquad \eta = 0{,}25, \qquad t_0 = 0$$

k	Текущее значение	Производная	Следующее значение
0	0	−6	1,5
1	1,5	−3	2,25
2	2,25	−1,5	?

Заполните пропуск и вычислите ещё одну итерацию. Сравните расстояния до точки минимума на соседних шагах.

Как шаг влияет на сходимость

Для нашей параболы расстояние до минимума меняется по простому правилу:

$$t_{k+1} - 3 = (1 - 2\eta)(t_k - 3)$$

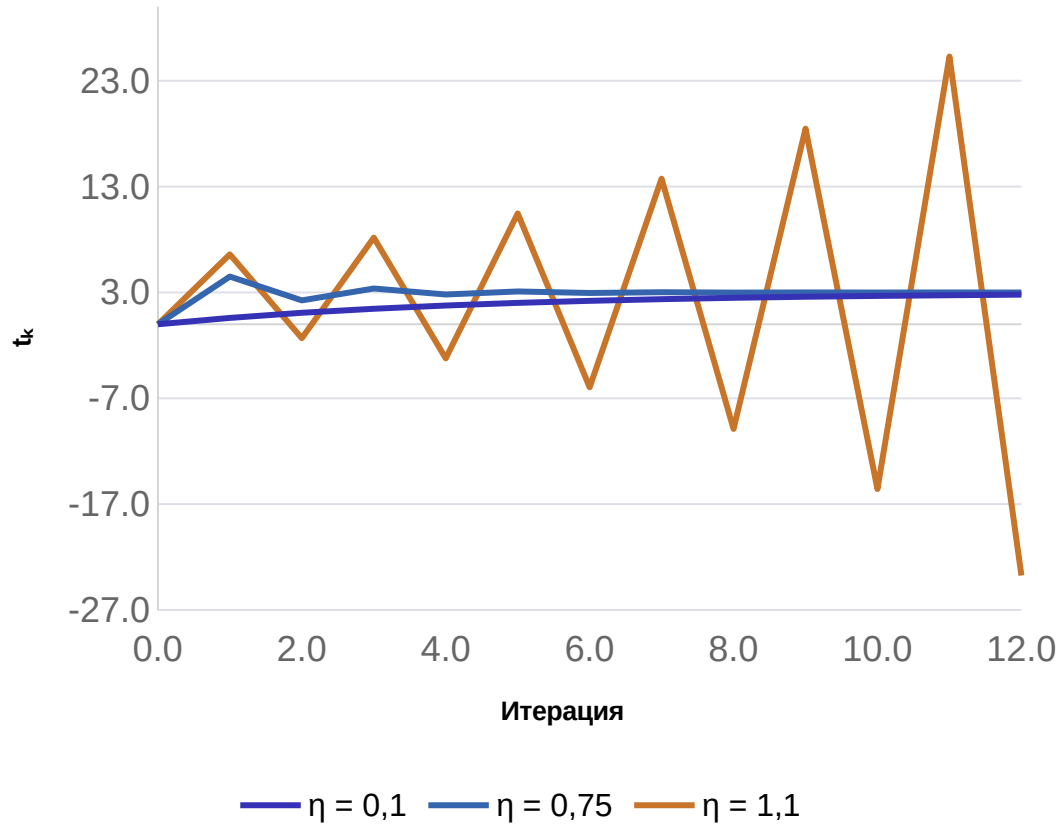
Шаг η	Множитель	Поведение из нуля
0,1	0,8	Медленная сходимость
0,75	−0,5	Затухающие колебания
1	−1	Постоянные колебания
1,1	−1,2	Расходимость

Для сходимости из любой начальной точки модуль множителя должен быть меньше единицы. Здесь это означает $0 < \eta < 1$.

Траектории при разных шагах

Одна функция и один старт. Меняется только шаг.

Колебания могут затухать или нарастать. Важен их размер, а не сам факт смены направления.



Маленький шаг даёт устойчивое, но медленное движение. Слишком большой шаг уводит от минимума.

Частные производные и градиент

Ошибка зависит от двух параметров. Чтобы изучить изменение по одному из них, удерживаем другой постоянным.

$$\frac{\partial L}{\partial w}$$

Изменяем только наклон прямой.

$$\frac{\partial L}{\partial b}$$

Изменяем только свободный коэффициент.

$$\nabla L = \begin{pmatrix} \frac{\partial L}{\partial w} \\ \frac{\partial L}{\partial b} \end{pmatrix}$$

Градиент объединяет частные производные в вектор. Он указывает направление наиболее быстрого локального роста функции.

Производная квадрата ошибки

$$e_i = wx_i + b - y_i$$

Изменение веса на h увеличивает остаток на произведение h и признака.
Изменение сдвига на h увеличивает остаток на h.

По параметру w

$$\frac{(e_i + hx_i)^2 - e_i^2}{h} = 2e_ix_i + hx_i^2 \longrightarrow 2e_ix_i$$

$$\frac{\partial(e_i^2)}{\partial w} = 2e_ix_i$$

По параметру b

$$\frac{(e_i + h)^2 - e_i^2}{h} = 2e_i + h \longrightarrow 2e_i$$

$$\frac{\partial(e_i^2)}{\partial b} = 2e_i$$

Так объясняется множитель признака в производной по весу.

Градиент ошибки модели

Производную суммы можно вычислять по слагаемым. Множитель усреднения сохраняется:

$$g_w = \frac{\partial L}{\partial w} = \frac{2}{n}\sum_{i=1}^{n} x_i(wx_i + b - y_i)$$

$$g_b = \frac{\partial L}{\partial b} = \frac{2}{n}\sum_{i=1}^{n} (wx_i + b - y_i)$$

Обе величины вычисляем при одних и тех же текущих параметрах. Затем одновременно обновляем вес и сдвиг.

Один шаг обучения вручную

$$(x_1, y_1) = (0, 1), \quad (x_2, y_2) = (1, 3), \quad w_0 = b_0 = 0, \quad \eta = 0{,}1$$

На двух точках начальные ошибки равны −1 и −3. Для этой выборки множитель 2/n равен единице.

$$g_w = 0 \cdot (-1) + 1 \cdot (-3) = -3$$

$$g_b = -1 - 3 = -4$$

$$g_w = -3, \quad g_b = -4, \quad w_1 = 0{,}3, \quad b_1 = 0{,}4$$

После шага MSE уменьшается с 5 до 2,825. Это отдельный маленький пример для ручного счёта.

Цикл обучения

1. Получить прогнозы и ошибки на обучающей выборке.

2. По тем же ошибкам вычислить обе частные производные.

3. Одновременно обновить параметры:

$$w_{k+1} = w_k - \eta g_w, \qquad b_{k+1} = b_k - \eta g_b$$

4. Повторять до лимита итераций или малости градиента.

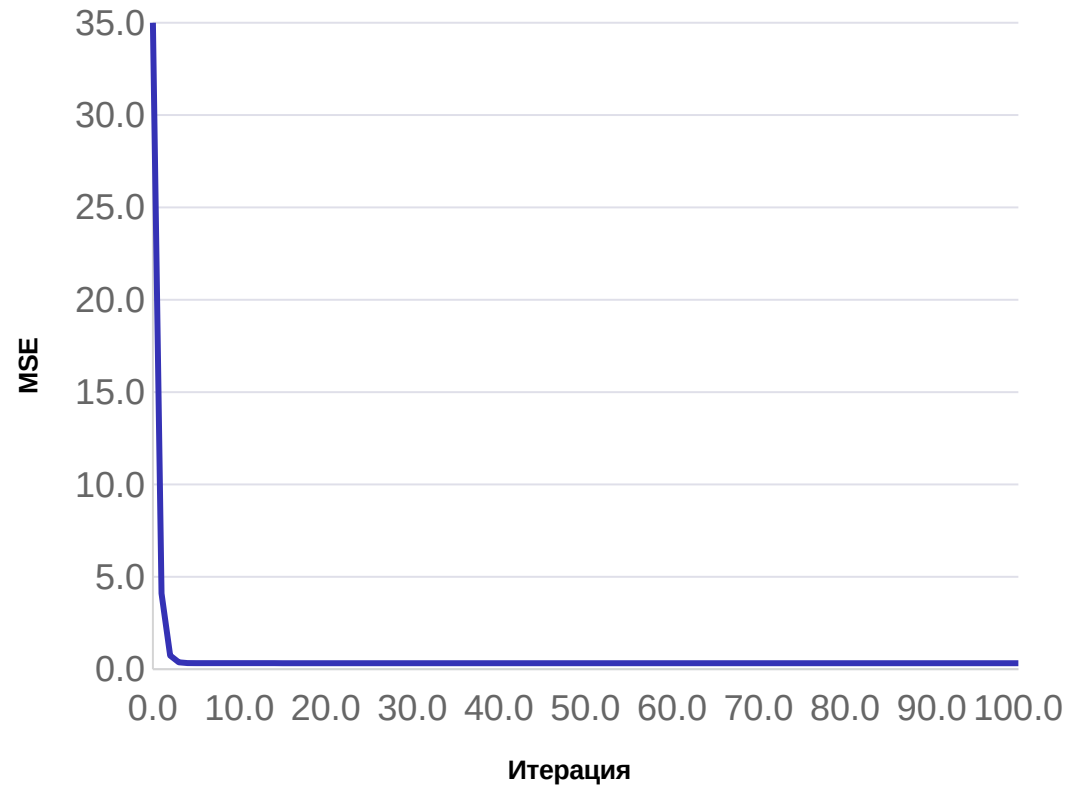
История MSE помогает проверить процесс. Не нужно ждать нулевой ошибки: при шумных данных она обычно недостижима.

Ошибка по ходу обучения

$$w_0 = b_0 = 0, \qquad \eta = 0{,}05$$

По мере обновления параметров MSE снижается.

На наших данных минимальная MSE равна 0,32. Исходные точки не лежат на одной прямой.



Резкий рост ошибки требует проверки шага, масштаба признаков и формул.
Плато выше нуля само по себе не означает ошибку реализации.

Проверка на новых объектах

На обучающей выборке подбираем параметры. На отложенных объектах проверяем, насколько хорошо работает найденное правило.

Простой ориентир — всегда предсказывать средний ответ на train:

$$\hat{y}_{\text{base}} = \frac{1}{n_{\text{train}}} \sum_{i \in \text{train}} y_i$$

Если выбираем настройки по validation, отдельный test оставляем для финальной оценки. Низкая ошибка на train ещё не доказывает полезность модели.

Масштаб признаков и шаг

Смена единиц измерения изменяет численные значения признака, градиенты и подходящий размер шага.

$$z = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}}$$

Среднее и стандартное отклонение считаем только на train. При обработке новых объектов используем эти же статистики.

Стандартизация упрощает подбор шага. Если признак постоянный, его стандартное отклонение равно нулю: делить на него нельзя.

Что такое СЛАУ

СЛАУ — система линейных алгебраических уравнений.

Во всех уравнениях одни и те же неизвестные. Они входят только в первой степени: неизвестные не перемножаются друг с другом.

$$\begin{cases} 2u + v = 5, \\ u - v = 1. \end{cases}$$

Решение системы — значения неизвестных, которые одновременно удовлетворяют каждому уравнению.

$$u = 2, \qquad v = 1$$

Метод исключения неизвестной

Чтобы решить систему, можно исключить одну неизвестную. В нашем примере складываем уравнения: слагаемые с v сокращаются.

$$\begin{cases} 2u + v = 5, \\ u - v = 1. \end{cases}$$

$$(2u + v) + (u - v) = 5 + 1 \quad \Longrightarrow \quad 3u = 6 \quad \Longrightarrow \quad u = 2$$

Подставляем найденное значение в одно из исходных уравнений:

$$v = u - 1 = 1$$

Проверяем результат в обоих исходных равенствах:

$$2 \cdot 2 + 1 = 5, \qquad 2 - 1 = 1$$

Сколько решений может быть

Одно решение

$$\begin{cases} 2u + v = 5, \\ u - v = 1. \end{cases}$$

Два независимых условия определяют одну пару значений.

Решений нет

$$\begin{cases} u + v = 2, \\ u + v = 3. \end{cases}$$

Одна и та же сумма не может равняться двум разным числам.

Бесконечно много

$$\begin{cases} u + v = 2, \\ 2u + 2v = 4. \end{cases}$$

Второе равенство повторяет первое после умножения на 2.

Число уравнений само по себе не гарантирует ни существования, ни единственности решения. Важна связь между условиями.

Матричная запись системы

Соберём коэффициенты в матрицу, неизвестные — в столбец, правые части — в другой столбец:

$$\underbrace{\begin{pmatrix} 2 & 1 \\ 1 & -1 \end{pmatrix}}_{A} \underbrace{\begin{pmatrix} u \\ v \end{pmatrix}}_{\boldsymbol{\theta}} = \underbrace{\begin{pmatrix} 5 \\ 1 \end{pmatrix}}_{\mathbf{c}}$$

Каждая строка матрицы соответствует уравнению. Каждый столбец соответствует одной неизвестной.

$$A \in \mathbb{R}^{m \times p}, \qquad \boldsymbol{\theta} \in \mathbb{R}^{p}, \qquad \mathbf{c} \in \mathbb{R}^{m}$$

В общем случае число уравнений и число неизвестных могут различаться. Матричная запись позволяет одинаково описывать маленькие и большие системы.

Минимум MSE линейной модели

В стационарной точке обе частные производные равны нулю. Для остатков в этой точке выполняются условия:

$$\sum_{i=1}^{n} e_i^* = 0, \qquad \sum_{i=1}^{n} x_i e_i^* = 0$$

Изменим параметры на произвольные величины и раскроем квадраты:

$$L(w^* + \Delta w, b^* + \Delta b) - L(w^*, b^*) = \frac{1}{n}\sum_{i=1}^{n}(\Delta w\, x_i + \Delta b)^2 \geq 0.$$

Смешанные слагаемые исчезают благодаря двум равенствам сверху. Остаётся сумма квадратов, поэтому стационарная точка даёт глобальный минимум.

Вывод нормальных уравнений

Приравняем производную по весу к нулю и уберём постоянный множитель усреднения:

$$\sum_{i=1}^{n} x_i(wx_i + b - y_i) = 0$$

$$w\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i - \sum_{i=1}^{n} x_i y_i = 0$$

Параметры не зависят от индекса объекта, поэтому их можно вынести за знак суммы. Из производной по сдвигу аналогично получаем:

$$\sum_{i=1}^{n}(wx_i + b - y_i) = 0$$

Минимум ошибки как СЛАУ

После раскрытия сумм и переноса известных слагаемых получаем систему для параметров прямой:

$$\begin{cases} w\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i y_i, \\ w\sum_{i=1}^{n} x_i + nb = \sum_{i=1}^{n} y_i. \end{cases}$$

Коэффициенты — суммы по исходной выборке. Неизвестных два: вес и свободный коэффициент.

Мы не требуем прохождения прямой через каждую точку. Система выражает условие минимума общей квадратичной ошибки.

Точное решение на пяти точках

Суммы по данным

$$\sum_i x_i = 10, \qquad \sum_i y_i = 25$$

$$\sum_i x_i^2 = 30, \qquad \sum_i x_i y_i = 72$$

$$\begin{cases} 30w + 10b = 72, \\ 10w + 5b = 25. \end{cases}$$

Исключение и результат

Умножим второе уравнение на 2 и вычтем его из первого.

$$10w = 22, \qquad w = 2{,}2, \qquad b = 0{,}6$$

Минимальная MSE равна 0,32.

Это независимый способ проверить градиентный спуск: при подходящем шаге его параметры должны приближаться к тому же решению.

Итерационное решение той же СЛАУ

$$A\boldsymbol{\theta} = \mathbf{c}, \qquad \boldsymbol{\theta} = \begin{pmatrix} w \\ b \end{pmatrix}$$

$$A = \begin{pmatrix} 30 & 10 \\ 10 & 5 \end{pmatrix}, \qquad \mathbf{c} = \begin{pmatrix} 72 \\ 25 \end{pmatrix}$$

Вектор невязки показывает, насколько текущие параметры не удовлетворяют нормальным уравнениям. Для нашей задачи обновление имеет вид:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \frac{2\eta}{5}(A\boldsymbol{\theta}_k - \mathbf{c})$$

Это то же правило градиентного спуска, записанное через матрицу системы. Подходящий шаг всё ещё необходим.

Когда решение не единственно

Если признак одинаков у всех объектов, разные веса и сдвиги могут давать одинаковые прогнозы.

$$x_i = 2, \quad y = (3, 5)^\mathsf{T}, \qquad 2w + b = 4$$

$$(w, b) \in \{(0, 4), (1, 2), (2, 0)\}$$

Все эти пары дают постоянный прогноз 4 и минимальную MSE = 1.
Единственного набора параметров нет.

В терминах СЛАУ условия зависимы: система не может отдельно определить вес и свободный коэффициент.

Несколько признаков

С одним объектом теперь связан вектор из d признаков:

$$\hat{y} = \sum_{j=1}^{d} w_j x_j + b$$

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{pmatrix}, \quad \mathbf{w} = \begin{pmatrix} w_1 \\ \vdots \\ w_d \end{pmatrix}$$

$$\hat{\mathbf{y}} = X\mathbf{w} + b\mathbf{1}$$

$$\nabla_{\mathbf{w}} L = \frac{2}{n} X^{\mathsf{T}}(X\mathbf{w} + b\mathbf{1} - \mathbf{y})$$

После добавления столбца единиц включаем свободный коэффициент в общий вектор параметров. Нормальные уравнения сохраняют ту же форму:

$$\widetilde{X}^{\mathsf{T}}\widetilde{X}\boldsymbol{\theta} = \widetilde{X}^{\mathsf{T}}\mathbf{y}$$

Формулы в NumPy

В одномерной реализации входы и ошибки — массивы одинаковой длины. Формулы переводятся в поэлементные операции и усреднение:

```
err = w * x + b - y

gw = 2 * np.mean(x * err)
gb = 2 * np.mean(err)

w_new = w - learning_rate * gw
b_new = b - learning_rate * gb
```

Сначала вычисляем обе производные на старых параметрах, затем присваиваем оба новых значения. Задания и проверки находятся в отдельном ноутбуке.