Производная и градиент

Как небольшое изменение параметра помогает понять, куда двигаться для уменьшения ошибки?

У модели есть настраиваемое число. Немного изменив его, мы меняем ошибку. Как выбрать направление изменения, не перебирая все возможные значения? Начнём с площади квадрата и дойдём до одного шага обучения.

1. Насколько меняется результат

Пусть площадь квадрата зависит от длины стороны: $f(x) = x^2$. При стороне три площадь равна девяти. Если увеличить сторону до четырёх, площадь станет шестнадцатью. Изменение площади равно семи, изменение стороны единице. Их отношение равно семи и описывает среднюю скорость изменения на этом промежутке.

Для меньшего изменения стороны результат другой. При переходе от 3 до 3,1 площадь меняется с 9 до 9,61. Отношение приращений равно 6,1. При переходе от 3 до 3,01 оно равно 6,01. Чем короче промежуток около трёх, тем ближе оценка к шести.

Производная описывает предельную скорость изменения в точке. Геометрически это наклон касательной к графику. У квадрата в точке три производная равна шести. Это не площадь и не новое значение функции: число говорит, насколько примерно изменится площадь на единицу небольшого изменения стороны.

У производной тоже есть единицы. Если функция описывает пройденный путь в метрах, а вход время в секундах, производная измеряется в метрах в секунду. Такая связь объясняет, почему производную часто называют мгновенной скоростью изменения. Слово «скорость» здесь относится к любому изменяющемуся результату, не только к движению тела.

Пример

При длине стороны три небольшая прибавка 0,01 даёт прибавку площади примерно `6 * 0,01 = 0,06`. Точная прибавка 0,0601 чуть больше: приближение хорошо для малого шага.

> Функция y = 3x + 2 получила приращение входа 4. На сколько изменился выход?

2. Производная квадрата из разности

Чтобы увидеть правило для квадрата, возьмём вход x и небольшую ненулевую прибавку h. Разность площадей равна $(x + h)^2 - x^2$. Раскрытие скобок даёт $2xh + h^2$. Деление на h оставляет $2x + h$. Когда прибавка стремится к нулю, остаётся $2x$.

Поэтому производная x^2 равна $2x$. При положительном входе она положительна, при отрицательном отрицательна, в нуле равна нулю. График квадрата слева от нуля опускается при движении вправо, справа поднимается. Знак производной передаёт именно это поведение.

У линейной функции $ax + b$ производная равна a: одинаковое приращение входа везде даёт одинаковое приращение выхода. У постоянной функции производная ноль. При сложении функций складываются их производные, а постоянный множитель сохраняется. Поэтому у $3x^2 + 2x + 5$ производная равна $6x + 2$.

Численную оценку можно получить и без вывода формулы: посчитать значения в двух близких точках и разделить разность на расстояние между ними. Слишком большой промежуток плохо описывает локальный наклон. Слишком маленький может усилить ошибки округления при вычитании близких машинных чисел. Выбор шага для численного расчёта требует проверки.

> Функция f(x) = x². Чему равна её производная при x = 4?

Профи. Численная производная: шаг вперёд и шаг в обе стороны

Оценка наклона по точкам x и $x + h$ — это **разность вперёд**: $\dfrac{f(x+h) - f(x)}{h}$. Для квадрата она равна $2x + h$: ошибка ровно h, и при $h = 0{,}1$ в точке 3 получаем $6{,}1$ вместо 6. Уменьшили h в 10 раз — ошибка уменьшилась в 10 раз.

Возьмём точки по обе стороны: **центральная разность** $\dfrac{f(x+h) - f(x-h)}{2h}$. Для квадрата числитель равен $(x+h)^2 - (x-h)^2 = 4xh$, после деления остаётся ровно $2x$ при любом h. Для других функций ошибка не исчезает, но становится порядка h^2: для x^3 центральная разность равна $3x^2 + h^2$, и при $h = 0{,}1$ ошибка всего $0{,}01$ вместо $0{,}61$ у разности вперёд.

Почему так: в разности $f(x+h) - f(x-h)$ слагаемые с чётными степенями h сокращаются, первая выжившая ошибка идёт от h^3, а после деления на $2h$ — от h^2.

Типичная ошибка — делить на h, а не на $2h$: ответ удваивается. Проверка: на прямой $y = 3x$ любая разность обязана дать ровно 3.

> Для $f(x) = x^3$ при $x = 4$ и шаге $h = 2$ вычислите центральную разность $\dfrac{f(x+h) - f(x-h)}{2h}$.

Обозначения

Запись	Смысл
h	небольшое ненулевое изменение входа
$f'(x)$	производная в точке x
Δf	изменение значения функции

Сравните две разности

Цель: Получите центральную разность для x³ в точке 2 с шагом 0,1; точное значение производной 12.

```
def f(x):
    return x ** 3

x, h = 2.0, 0.1
forward = (f(x + h) - f(x)) / h
central = forward  # TODO: разность по обе стороны от x
print(f"вперёд: {forward:.4f}  центральная: {central:.4f}  точно: {3 * x
** 2}")
```

3. Уменьшить ошибку одним шагом

Представьте модель с одним настраиваемым числом w. Пусть её ошибка равна $L(w) = (w - 3)^2$. Это учебная задача: минимум заранее понятен, при $w = 3$ ошибка ноль. Она нужна, чтобы увидеть механику обучения, прежде чем минимум станет неизвестен.

Производная ошибки равна $2(w - 3)$. В точке ноль она равна минус шести. Отрицательный наклон означает: небольшое увеличение входа уменьшит ошибку. Для движения вниз вычтем из текущего параметра производную, умноженную на положительный размер шага η. Правило записывается как $w_{new} = w - \eta L\prime(w)$.

При $w = 0$ и шаге 0,1 получаем новое значение 0,6. Ошибка уменьшается с девяти до 5,76. Если повторять расчёт, каждый раз производную нужно вычислять в новой точке: прежний наклон уже не описывает текущее положение.

В опыте меняется размер одного шага из нуля. Сначала попробуйте 0,1, затем 0,5 и 1,1. При 0,5 шаг попадает точно в минимум. При 1,1 параметр становится 6,6 и ошибка возрастает до 12,96. Верное направление движения в начале ещё не гарантирует, что большой шаг закончится ближе к минимуму.

Проверьте свой прогноз

Заметка

В этой задаче шаг ровно один отражает параметр относительно минимума и сохраняет ошибку. Это особенность конкретной квадратичной функции, а не универсальный рецепт выбора шага.



L(w) = (w − 3)². Из w = 0 идём в w = 0,6. Ошибка: 9 → 5,76. Первый столбик — до шага, второй — после.

Один шаг из w = 0 для ошибки (w − 3)². Сравните маленький и слишком большой шаг.

Интерактивный эксперимент — в электронной версии урока.

> Для ошибки (w − 3)² в точке w = 1 производная равна −4. Выполните шаг w_new = w − 0,1 × производная. Чему равно новое w?

4. Несколько входов и градиент

Теперь ошибка зависит от двух параметров: $L(x, y) = x^2 + 2y^2$. Чтобы понять влияние первого, временно удержим второй постоянным. Производная по x равна $2x$. Аналогично по y получаем $4y$. Их называют **частными производными**, потому что каждая описывает изменение по одной координате.

Градиент собирает частные производные в вектор: $(2x, 4y)$. В точке (1; 2) он равен (2; 8). В обычной евклидовой геометрии ненулевой градиент направлен в сторону самого быстрого локального роста. Чтобы уменьшать ошибку, идём в противоположную сторону.

При шаге 0,1 новая точка равна (0,8; 1,2): из первой координаты вычитаем 0,2, из второй 0,8. Прежняя ошибка равна девяти, новая 3,52. Оба обновления используют градиент в исходной точке. Если после изменения первой координаты пересчитать вторую производную, вы получите уже другой алгоритм.

Длины шагов по координатам могут различаться, даже если общий коэффициент один. Кроме того, масштаб самих координат влияет на численные производные. Это одна из причин, по которым подготовка и масштабирование признаков помогают обучению. Градиент полезно сначала проверить на простой функции, где ожидаемое направление видно без программы.

Ссылки

MIT: дифференциальное исчисление развивает производную и её геометрический смысл. Для этого урока достаточно приведённого вывода квадрата.

Профи. **Почему градиент — направление самого быстрого роста**

Сдвинемся из точки (x, y) на маленький шаг t вдоль направления (u, v) единичной длины. Каждая координата вносит вклад через свою частную производную:

$$L(x + tu,\, y + tv) \approx L(x, y) + t\,(L_x u + L_y v).$$

Скобка — скалярное произведение градиента и направления. Оно равно $|\nabla L| \cos\theta$, где θ — угол между ними, и наибольшее при $\theta = 0$: вдоль самого градиента. Против градиента $\cos\theta = -1$ — самое быстрое убывание. Перпендикулярно градиенту изменение в первом порядке нулевое: это направление вдоль линии уровня.

Посчитаем для $L = x^2 + 2y^2$ в точке $(1; 2)$ с градиентом $(2; 8)$. Вдоль оси x скорость роста 2, вдоль оси y — 8, вдоль единичного вектора по градиенту $\sqrt{4 + 64} \approx 8{,}25$. Шаг $0{,}1$ против градиента сдвигает y в четыре раза сильнее, чем x: длина шага по координате пропорциональна частной производной, а не одинакова.

Типичная ошибка — сравнивать направления по неединичным векторам: $(4; 16)$ даст «скорость» 68, хотя направление то же. Поймать её можно, разделив вектор на его длину перед подстановкой.

> Градиент ошибки в текущей точке равен $(5;\ 5)$. Найдите скорость изменения ошибки вдоль единичного направления $(0{,}6;\ 0{,}8)$.

5. Когда нулевой наклон не спасает

Нулевая производная сама по себе не доказывает минимум. У функции $-x^2$ в нуле производная тоже ноль, но это максимум. У x^3 производная в нуле равна нулю, однако функция продолжает возрастать по обе стороны. Чтобы понять тип точки, нужно исследовать соседние значения или другие свойства функции.

В нескольких измерениях возможна **седловая точка**: по одному направлению значение растёт, по другому уменьшается. Например, у $x^2 - y^2$ в начале координат градиент нулевой. По оси x соседние значения положительны, по оси y отрицательны. Алгоритм, который смотрит только на текущий градиент, может остановиться и здесь.

В лаборатории исправьте знак обновления. Стартовый код идёт по градиенту и увеличивает ошибку. После исправления один шаг из нуля должен дать 0,6. Проверьте также шаг из точки четыре: производная там положительна, поэтому параметр должен уменьшиться. Одно правило должно работать по обе стороны минимума.

Пойти вниз по ошибке

Цель: Выполните шаг против градиента, а не вдоль него.

```
w = 0
rate = 0.1
grad = 2*(w-3)
new_w = w + rate*grad  # TODO: исправьте направление
print(f"Параметр после шага: {new_w}")
print(f"Новая ошибка: {(new_w-3)**2:.2f}")
```

Практическое обучение требует следить за ошибкой после обновления, а не только доверять знаку производной. Если значение растёт, уменьшите шаг и проверьте формулу градиента. Если перестало меняться, выясните, действительно ли достигнут минимум, не слишком ли мал шаг и не мешает ли округление.

Эпик. Какие шаги уменьшают эту ошибку?

Для $L(w) = (w - 3)^2$ найдите все положительные постоянные шаги, при которых ошибка уменьшается из любой точки, кроме минимума.

▾*Решение*

После обновления отклонение от минимума равно $(1 - 2\eta)(w - 3)$. Ошибка умножается на $(1 - 2\eta)^2$. Для строгого уменьшения нужен модуль множителя меньше единицы: $0 < \eta < 1$. При $\eta = 1$ ошибка сохраняется, при $\eta > 1$ растёт.

История

Метод наискорейшего спуска связывает локальное изменение функции с последовательностью улучшений. Курс Stanford по оптимизации рассматривает строгие условия сходимости. Одного наблюдения «ошибка упала» для доказательства сходимости недостаточно.

Подумай

Проверьте ошибку после шага из минимума: производная нулевая, поэтому параметр не изменится. Затем рассмотрите $-x^2$ в нуле: тот же нулевой шаг уже не свидетельствует о минимуме.

Крутизна меняет границу. Для $L(w) = a(w-3)^2$ с $a > 0$ производная равна $2a(w-3)$, множитель отклонения — $1 - 2a\eta$, и условие превращается в $0 < \eta < 1/a$. Чем круче парабола, тем меньше допустимый шаг.

Худший случай — две разные крутизны. Пусть $L(x, y) = x^2 + 50y^2$. По y шаг обязан быть меньше $1/50 = 0{,}02$, иначе эта координата разлетится. Безопасный шаг $\eta = 0{,}01$ сжимает отклонение по x всего на множитель $0{,}98$ за шаг. Чтобы уменьшить $|x|$ в сто раз, нужно n с $0{,}98^n = 0{,}01$, то есть $n = \ln 100 / \ln(1/0{,}98) \approx 4{,}6 / 0{,}0202 \approx 230$ шагов. Один общий шаг для координат с разной крутизной либо расходится, либо ползёт — поэтому признаки масштабируют перед обучением.

Контрпример. Для $L(w) = |w - 3|$ производная равна $+1$ справа от минимума и -1 слева: её величина не убывает при приближении. Постоянный шаг η из точки $3 + \eta/2$ переносит параметр в $3 - \eta/2$ и обратно — ошибка не уменьшается ни при каком $\eta > 0$. Правило «при малом шаге ошибка падает из любой точки» опирается на то, что производная стремится к нулю у минимума; без этого нужен убывающий шаг.

> Ошибка равна $L(w) = 3(w-3)^2$. Найдите наибольший размер постоянного шага η, такой что любой шаг градиентного спуска $w_{new} = w - \eta L'(w)$ с размером меньше него уменьшает ошибку из любой точки, кроме минимума. Ответ с точностью до сотых.

> Для $L(x, y) = x^2 + 50y^2$ градиентный спуск идёт с постоянным шагом $\eta = 0{,}01$, безопасным для координаты y. Примерно сколько шагов нужно, чтобы отклонение по x уменьшилось в 100 раз?
>
> ○ Около 230
> ○ Около 10 000
> ○ Около 20
> ○ Около 5

Где это встретится дальше

Производная описывает локальное изменение. Шаг против градиента выбирает направление уменьшения, но его размер нужно проверять.

Оценить наклон по двум значениям и найти производную квадрата.

Вычислить градиент простой функции двух переменных.

Выполнить шаг уменьшения ошибки и проверить результат.

Домашняя работа

1. Для f(x) = 3x² + 2x найдите производную при x = 2.

2. Для функции ошибки $L(w) = (w - 3)^2$ из $w = 5$ сделайте шаг градиентного спуска: $w_{new} = w - 0{,}1L'(w)$. Найдите новое значение параметра и сравните ошибку до и после шага.
3. ★ Почему нулевой градиент $x^2 - y^2$ в начале координат не доказывает минимум?

Источники

- MIT: производные (https://ocw.mit.edu/courses/18-01sc-single-variable-calculus-fall-2010/).
- Stanford: выпуклая оптимизация (https://web.stanford.edu/~boyd/cvxbook/).
- Python: математические функции (https://docs.python.org/3/library/math.html).