Как учится модель

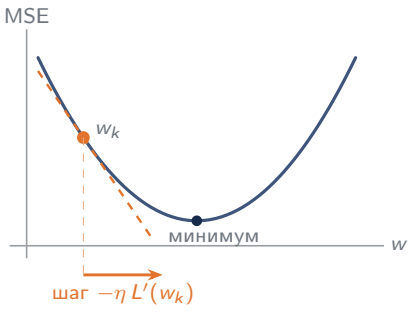
Данные, проверка качества, производные и градиентный спуск

Алексей Чернуха

24 октября 2026

Повторение

Что было на прошлой неделе



Линейная регрессия

Модель $\hat{y} = wx + b$, ошибка $\mathrm{MSE} = \frac{1}{n}\sum(\hat{y}_i - y_i)^2$.

Градиентный спуск в 1D

$$w_{k+1} = w_k - \eta\, L'(w_k)$$

Производная показывает наклон. Идём в сторону убывания, шаг η выбираем сами: маленький — медленно, большой — разлетаемся.

Сегодня: откуда берутся данные, как честно проверить модель и как считать градиент, когда параметров много.

Откуда берутся датасеты

Откуда берутся данные

Готовые

Kaggle, Hugging Face, UCI. Быстро, но задачу за вас уже кто-то сформулировал.

Собранные

Логи сервиса, парсинг сайтов, датчики, фото на телефон.

Размеченные

Объекты есть, ответов нет. Ответы ставят люди: эксперты или краудсорсинг.

Синтетические

Сгенерированы по формуле или другой моделью.

Чаще всего самое долгое в ML-проекте — не обучение, а сбор и чистка данных.

MNIST: рукописные цифры



Настоящие рукописные цифры 8×8 пикселей из набора UCI — младшего брата MNIST. В MNIST картинки 28×28.

Что это

70 000 цифр: 60 000 для обучения, 10 000 для теста. Первый датасет почти любого курса по ML.

Как собирали

Исходно в NIST одну часть цифр писали сотрудники Бюро переписи США, другую — школьники. Школьники писали заметно небрежнее.

MNIST: в чём была ловушка



Вопрос

В исходном разбиении NIST модель учили на аккуратных цифрах, а проверяли на небрежных. Что пойдёт не так?

MNIST: в чём была ловушка



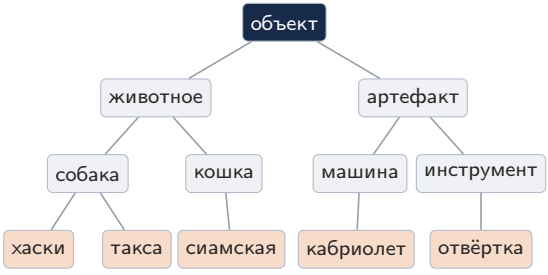
Вопрос

В исходном разбиении NIST модель учили на аккуратных цифрах, а проверяли на небрежных. Что пойдёт не так?

Ответ

Train и test из разных распределений: на тесте ошибок намного больше, и непонятно, плоха модель или данные. В MNIST обе части поделили поровну.

ImageNet: 14 миллионов картинок



Классы образуют дерево понятий (WordNet). Всего около 22 тысяч классов.

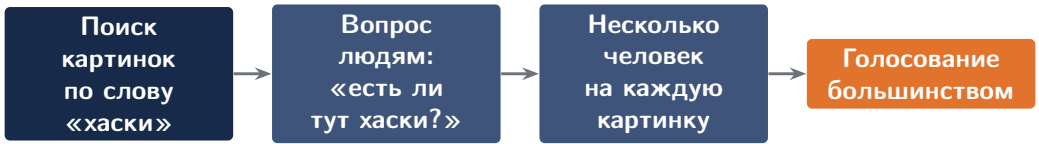
Что это

Проект Фей-Фей Ли (2009). Для соревнования ILSVRC взяли 1000 классов и 1,2 млн картинок для обучения.

Почему важен

На нём в 2012 году нейросеть AlexNet сильно обогнала все остальные методы. С этого начался бум глубокого обучения.

ImageNet: как разметить 14 миллионов картинок



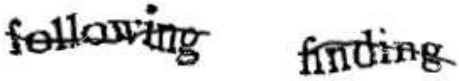
Краудсорсинг

Работу делили на крошечные задания и раздавали тысячам людей через Amazon Mechanical Turk. Одно задание — несколько секунд и пара центов.

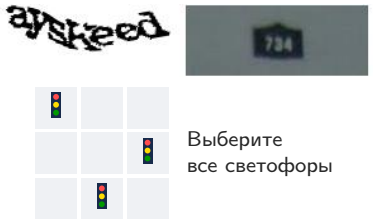
Почему не один человек

Люди ошибаются и устают. Если трое из пяти сказали «да», ответу можно верить больше, чем одному голосу.

reCAPTCHA: вы тоже размечали данные



2007. Два слова из отсканированных книг. Одно система знает — по нему проверяют, что вы человек. Второе не смогла прочитать программа — его читаете вы.



Позже. Номера домов с Google Street View, затем светофоры, машины и переходы — разметка для карт и беспилотников.

Волк или хаски?

Учим классификатор отличать волков от хаски. Вот часть обучающей выборки:


волк


волк


хаски


хаски

Вопрос. Точность на отложенной выборке из таких же фото хорошая. Что может пойти не так?

Волк или хаски: ответ



Ответ модели: волк

Модель смотрела на снег

Все волки в обучении стояли на снегу, все хаски — на траве. Проще всего отличать их по фону. Хаски на снегу — «волк».

Эксперимент (2016)

Исследователи специально собрали такую выборку, чтобы показать: высокая точность ещё не значит, что модель поняла задачу.

Модель выучит любую закономерность, которая помогает на train, даже если она случайна. Такие проблемы надо ловить при сборе данных.

Задание: собираем датасет листьев

Обсудим вместе. Хотим, чтобы модель по фото листа определяла дерево.

1. Какие классы? Сколько фото на класс нужно минимум?
2. Фон, свет, ракурс: одинаковые или разные? Что будет в каждом случае?
3. Как называть файлы и где хранить ответ (породу)?
4. Если один лист сфотографирован 5 раз, можно ли эти фото раскидать в train и test?
5. Как проверить, что другой человек разметил так же?

Ответы: правила сбора

Разнообразие

Разные фоны, освещение и телефоны внутри **каждого** класса. Если все клёны сняты на асфальте, а все дубы на траве, модель выучит асфальт.

Баланс

Примерно одинаковое число фото на класс, иначе accuracy обманет.

Утечка

Фото одного листа — одна группа. Группа целиком идёт либо в train, либо в test. Иначе тест проверяет память, а не умение обобщать.

Единый формат

Инструкция для всех: имя файла, папка класса, таблица с метаданными. Тогда датасеты всех учеников можно склеить в один.

Жизнь модели

Путь модели: от данных до инференса



train

Подбираем параметры w: минимизируем ошибку градиентным спуском. Здесь же считаем среднее и std для нормализации.

validation

Выбираем гиперпараметры: k в KNN, шаг η, число итераций. Смотрим сколько угодно раз.

test

Один раз в конце, чтобы честно сообщить качество. Подсмотрели и поменяли модель — тест испорчен.

Инференс: модель в работе

Что происходит

Приходит новый объект без ответа. Модель только считает $\hat{y} = f(x; w)$. Параметры w заморожены, градиенты не считаем.

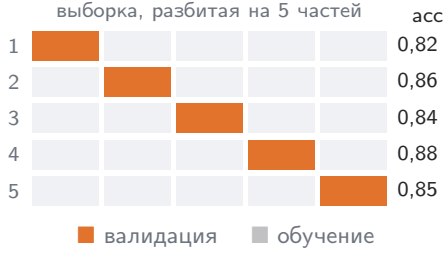
Та же предобработка

Нормализуем новый объект средним и std **с train**, а не по новым данным.

Что важно

- скорость: ответ нужен за миллисекунды
- память: модель должна влезть в телефон или сервер
- стоимость: инференс идёт миллионы раз, обучение — один

Кросс-валидация



k-fold

Делим train на k частей. k раз обучаем модель заново, каждый раз проверяем на новой части.

Что получаем

Не одно число, а пять:

$$\text{acc} = 0{,}85 \pm 0{,}02$$

Видно и среднее качество, и насколько оно скачет.

Минус: обучаем модель k раз. Для маленьких данных это почти всегда того стоит.

Модель B лучше модели A?

Вопрос. На тесте модель A дала accuracy 0,85, а модель B — 0,86. Берём B?

Тест из 100 объектов

Разница — **один** объект. Погрешность accuracy примерно

$$\pm 2\sqrt{\frac{p(1-p)}{n}} = \pm 2\sqrt{\frac{0{,}85 \cdot 0{,}15}{100}} \approx \pm 0{,}07.$$

Разница 0,01 тонет в шуме.

Тест из 10 000 объектов

Та же формула даёт погрешность около $\pm 0{,}007$. Разница 0,01 уже похожа на настоящую.

Правило на практике

Если разница меньше разброса между фолдами, не верим ей.

Откуда формула: каждый объект теста — «монетка», угаданная с вероятностью p. Доля угаданных из n имеет стандартное отклонение $\sqrt{p(1-p)/n}$ (биномиальное распределение), а в ± 2 стандартных отклонения попадает около 95 % случаев.

На каком этапе это происходит?

1. Выбрать число соседей k в KNN
2. Посчитать среднее и std признаков для нормализации
3. Показать пользователю рекомендации в приложении
4. Сообщить итоговую accuracy в отчёте
5. Сделать шаг градиентного спуска

Ответы: этапы

1	Валидация (или кросс-валидация): это гиперпараметр
2	Обучение: только по train, иначе утечка
3	Инференс
4	Тест, один раз
5	Обучение

Производная

Таблица производных и правила

$f(x)$	$f'(x)$
c	0
x^n	$n\,x^{n-1}$
e^x	e^x
$\ln x$	$\frac{1}{x}$
$\sin x$	$\cos x$
$\cos x$	$-\sin x$

Правила

сумма	$(f+g)' = f' + g'$
число	$(c\,f)' = c\,f'$
произведение	$(fg)' = f'g + fg'$

сложная функция

$$\big(f(g(x))\big)' = f'\big(g(x)\big)\cdot g'(x)$$

Цепное правило словами

Производная внешней функции, внутренность не трогаем, умножить на производную внутренней.

Разминка: правила

Самостоятельно. Найдите производные.

1. $f(x) = 4x^3 - 2x + 7$
2. $f(x) = 3\sin x + e^x$

Ответы: разминка

1. $12x^2 - 2$
2. $3\cos x + e^x$

Цепное правило на примерах

Функция	Внешняя и внутренняя	Производная
$(3x-1)^2$	$u^2,\ \ u = 3x-1$	$2(3x-1)\cdot 3 = 6(3x-1)$
$\sin^2 x$	$u^2,\ \ u = \sin x$	$2\sin x \cos x$
$\ln(x^2+1)$	$\ln u,\ \ u = x^2+1$	$\dfrac{2x}{x^2+1}$
e^{-x^2}	$e^u,\ \ u = -x^2$	$-2x\, e^{-x^2}$

Проверить себя можно численно: $f'(x) \approx \frac{f(x+h)-f(x-h)}{2h}$ при маленьком h, например 10^{-5}.

Самостоятельно: цепное правило

1	$f(x) = (2x + 5)^3$
2	$f(x) = x\, e^{2x}$ (произведение и цепное правило)
★	$\sigma(x) = \dfrac{1}{1 + e^{-x}}$. Покажите, что $\sigma'(x) = \sigma(x)\big(1 - \sigma(x)\big)$

Ответы: цепное правило

1	$3(2x+5)^2 \cdot 2 = 6(2x+5)^2$
2	$e^{2x} + 2x\, e^{2x} = (1+2x)\, e^{2x}$
★	$\sigma' = \dfrac{e^{-x}}{(1+e^{-x})^2} = \dfrac{1}{1+e^{-x}} \cdot \dfrac{e^{-x}}{1+e^{-x}} = \sigma\,(1-\sigma)$

Производная по параметру, а не по x

В ML переменная — это параметр модели, а данные x, y — просто числа.

Ошибка на одном объекте

$$\frac{\partial}{\partial w}(wx - y)^2 = 2(wx - y) \cdot x$$

Внешняя функция — квадрат, внутренняя — $wx - y$. Её производная по w равна x.

Два признака

$$\frac{\partial}{\partial w_2}(w_1x_1 + w_2x_2 - y)^2 = 2(w_1x_1 + w_2x_2 - y) \cdot x_2$$

w_1 при этом считаем числом: производная по w_2 от w_1x_1 равна нулю.

Самостоятельно: производная по параметру

1. $L(b) = (3w + b - 7)^2$ по b
2. $L(w) = e^{wx}$ по w

Ответы: производная по параметру

1. $2(3w + b - 7)$
2. $x\, e^{wx}$: здесь x — число

Градиент

Частные производные и градиент

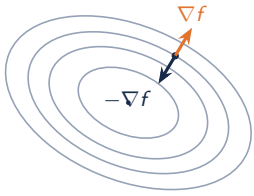
Пример

$$f(x, y) = x^2 + xy + 2y^2$$

Производная по x: y считаем числом. По y: x считаем числом.

$$\frac{\partial f}{\partial x} = 2x + y, \qquad \frac{\partial f}{\partial y} = x + 4y$$

$$\nabla f = \begin{pmatrix} 2x + y \\ x + 4y \end{pmatrix}, \qquad \nabla f(1, 1) = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$$



Линии уровня: f одинакова вдоль каждой. Градиент перпендикулярен им и смотрит в сторону роста.

Откройте в GeoGebra 3D: точку B можно двигать по поверхности, оранжевый вектор — ∇f в этой точке.

Градиент и сдвиг параметров

Определение

Градиент — вектор из всех частных производных:

$$\nabla f(w) = \left(\frac{\partial f}{\partial w_1}, \ldots, \frac{\partial f}{\partial w_k}\right)$$

Сдвиг h

На сколько меняем каждый параметр: $w \to w + h$. Пример: $(1,1) + (-0{,}15;\ -0{,}25) = (0{,}85;\ 0{,}75)$.

Линейное приближение

Если h маленький, функция рядом почти плоская:

$$f(w + h) \approx f(w) + \nabla f(w) \cdot h$$

Справа скалярное произведение:

$$\nabla f \cdot h = \frac{\partial f}{\partial w_1} h_1 + \cdots + \frac{\partial f}{\partial w_k} h_k$$

Каждый параметр сдвинули — функция изменилась примерно на «наклон $\times$ сдвиг».

Почему шаг против градиента уменьшает функцию

Куда шагнуть

Изменение функции:

$$\nabla f \cdot h = \|\nabla f\| \, \|h\| \cos\varphi$$

При шаге фиксированной длины оно самое отрицательное при $\cos\varphi = -1$, то есть когда h смотрит **против** градиента.

Шаг ГД

Берём $h = -\eta \nabla f$:

$$f(w - \eta \nabla f) \approx f(w) - \eta \|\nabla f\|^2 < f(w)$$

Проверим на числах

$f = x^2 + xy + 2y^2$, $w = (1, 1)$, $f = 4$, $\nabla f = (3, 5)$, $\eta = 0{,}05$. Прогноз $4 - 0{,}05 \cdot 34 = 2{,}3$, на деле $f(0{,}85;\ 0{,}75) \approx 2{,}49$.

Много параметров: линейная регрессия

Объект — k признаков, модель $\hat{y}_i = w_1 x_{i1} + \cdots + w_k x_{ik}$. Ошибка $\mathrm{err}_i = \hat{y}_i - y_i$.

Шаг 1. Один объект

Цепное правило:

$$\frac{\partial}{\partial w_j} \mathrm{err}_i^2 = 2\,\mathrm{err}_i \cdot x_{ij}$$

Шаг 2. Среднее по объектам

Производная суммы — сумма производных:

$$g_j = \frac{2}{n} \sum_{i=1}^{n} x_{ij}\,\mathrm{err}_i$$

Это просто цикл: для каждого параметра j пробегаем объекты и копим сумму. Осталось заметить, что такой цикл уже есть в numpy.

Шаг 3. Цикл — это $X^\top \mathrm{err}$

3 объекта, 2 признака:

$$X = \begin{pmatrix} 1 & 2 \\ 3 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathrm{err} = \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix}$$

Циклом:

$$\sum_i x_{i1}\mathrm{err}_i = 1 \cdot 1 + 3 \cdot (-1) + 0 \cdot 2 = -2$$
$$\sum_i x_{i2}\mathrm{err}_i = 2 \cdot 1 + 0 \cdot (-1) + 1 \cdot 2 = 4$$

Матрицей: столбец j матрицы X — это строка j в $X^\top$.

$$X^\top \mathrm{err} = \begin{pmatrix} 1 & 3 & 0 \\ 2 & 0 & 1 \end{pmatrix} \begin{pmatrix} 1 \\ -1 \\ 2 \end{pmatrix} = \begin{pmatrix} -2 \\ 4 \end{pmatrix}$$

Ответы совпали.

$$\nabla \mathrm{MSE} = \frac{2}{n} X^\top (Xw - y)$$

Правило размерностей

Не помните, где транспонировать? Проверьте формы.

Что	Форма	Смысл
X	$n \times k$	объекты $\times$ признаки
w	$k \times 1$	по весу на признак
$Xw - y$	$n \times 1$	по ошибке на объект
∇	$k \times 1$	**как у** w: по числу на параметр

Из X $(n \times k)$ и ошибок $(n \times 1)$ получить $k \times 1$ можно только так: $(k \times n) \cdot (n \times 1)$, то есть $X^\top \mathrm{err}$. Если формы не сходятся — формула неверна.

```
grad = 2 / n * X.T @ (X @ w - y)
```

Градиентный спуск

Градиентный спуск для k параметров

Алгоритм

1. Начальное w: нули или случайные числа.
2. Считаем $\nabla L(w)$ по всем объектам.
3. Обновляем **все** параметры сразу: $w \leftarrow w - \eta \nabla L(w)$
4. Повторяем, пока $\|\nabla L\|$ не станет маленьким или не кончится лимит итераций.

Код

```
w = np.zeros(k)
for step in range(1000):
    err = X @ w - y
    grad = 2 / n * X.T @ err
    if np.linalg.norm(grad) < eps:
        break
    w = w - lr * grad
```

Если функция не парабола, минимумов может быть несколько: куда придём, зависит от стартовой точки.

Стохастический градиентный спуск (SGD)

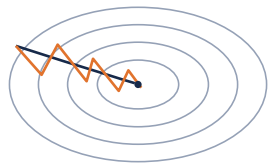
Проблема

Шаг ГД — проход по **всем** n объектам. При $n = 10^7$ это долго.

Идея

Градиент по случайному батчу из B объектов, например $B = 64$:

$$\nabla L \approx \frac{2}{B} X_B^\top (X_B w - y_B)$$



Шаги SGD шумные, но каждый в n/B раз дешевле. Проход по всем данным батчами — **эпоха**. Так учат почти все нейросети.

Зачем нормализовать признаки

200 квартир: площадь 30–120 м^2, комнат 1–4. Цена $\approx 0{,}05 \cdot$ площадь $+ 0{,}8 \cdot$ комнаты.

Без нормализации

Первый градиент из $w = 0$:

$$\nabla L = \begin{pmatrix} -936 \\ -30 \end{pmatrix}$$

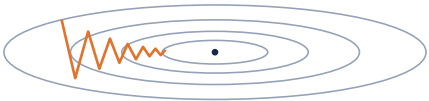
По площади в 30 раз больше: значения признака больше, а в g_j стоит множитель x_{ij}.

Шаг один на всех

Шаг ограничен «площадью». При таком шаге вес комнат еле двигается.

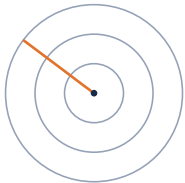
Шаг η	**Итераций**
10^{-3}	расходится
$3 \cdot 10^{-4}$	расходится
10^{-4}	21 404

Овраг и круг



Без нормализации

Линии уровня вытянуты. ГД скачет поперёк оврага и медленно ползёт вдоль.



После стандартизации

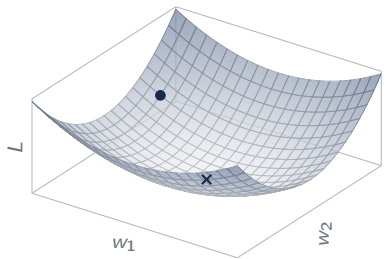
Градиенты одного порядка, можно взять большой шаг: $\eta = 0{,}3$ — **11** итераций вместо 21 404.

Стандартизация: $x' = (x - \text{среднее})/\text{std}$. Среднее и std считаем только на train.

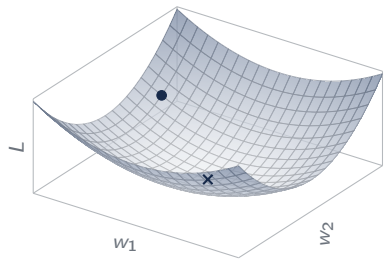
Слишком большой шаг

$L(w_1, w_2) = w_1^2 + 3w_2^2$, старт в одной и той же точке. Листайте: каждый клик — один шаг ГД. Пунктир — часть шага, скрытая за передней стенкой чаши.

$\eta = 0{,}1$: спускаемся



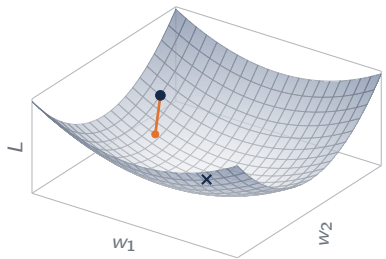
$\eta = 0{,}35$: перепрыгиваем



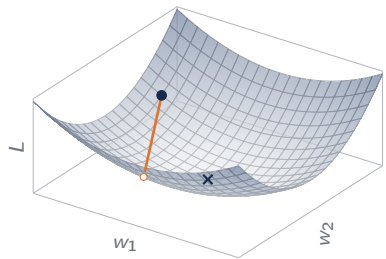
Слишком большой шаг

$L(w_1, w_2) = w_1^2 + 3w_2^2$, старт в одной и той же точке. Листайте: каждый клик — один шаг ГД. Пунктир — часть шага, скрытая за передней стенкой чаши.

$\eta = 0{,}1$: спускаемся



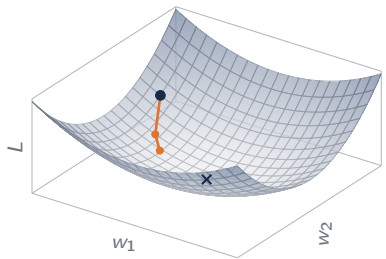
$\eta = 0{,}35$: перепрыгиваем



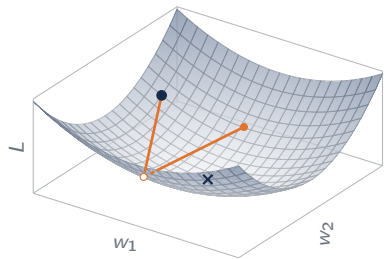
Слишком большой шаг

$L(w_1, w_2) = w_1^2 + 3w_2^2$, старт в одной и той же точке. Листайте: каждый клик — один шаг ГД. Пунктир — часть шага, скрытая за передней стенкой чаши.

$\eta = 0{,}1$: спускаемся



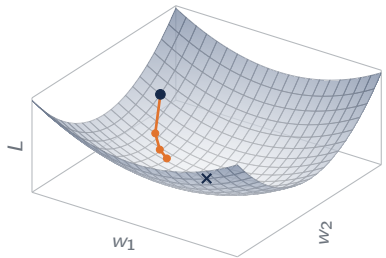
$\eta = 0{,}35$: перепрыгиваем



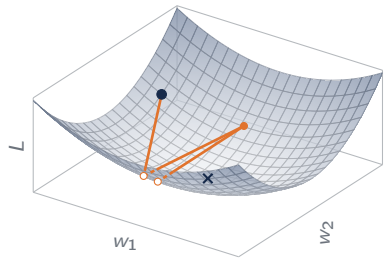
Слишком большой шаг

$L(w_1, w_2) = w_1^2 + 3w_2^2$, старт в одной и той же точке. Листайте: каждый клик — один шаг ГД. Пунктир — часть шага, скрытая за передней стенкой чаши.

$\eta = 0{,}1$: спускаемся



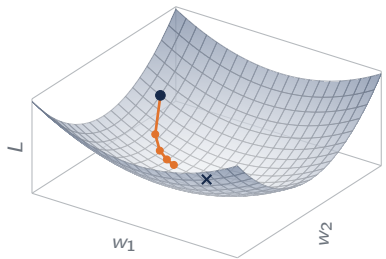
$\eta = 0{,}35$: перепрыгиваем



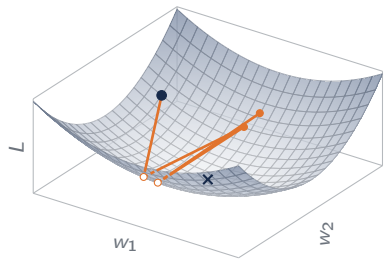
Слишком большой шаг

$L(w_1, w_2) = w_1^2 + 3w_2^2$, старт в одной и той же точке. Листайте: каждый клик — один шаг ГД. Пунктир — часть шага, скрытая за передней стенкой чаши.

$\eta = 0{,}1$: спускаемся



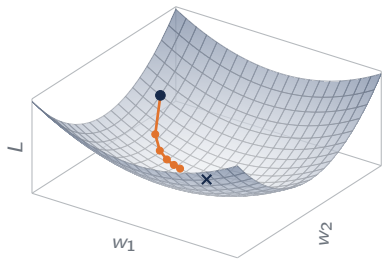
$\eta = 0{,}35$: перепрыгиваем



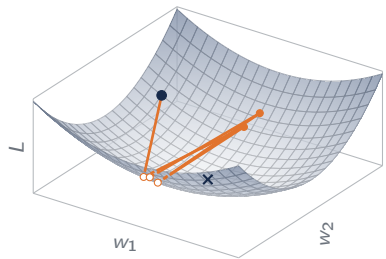
Слишком большой шаг

$L(w_1, w_2) = w_1^2 + 3w_2^2$, старт в одной и той же точке. Листайте: каждый клик — один шаг ГД. Пунктир — часть шага, скрытая за передней стенкой чаши.

$\eta = 0{,}1$: спускаемся



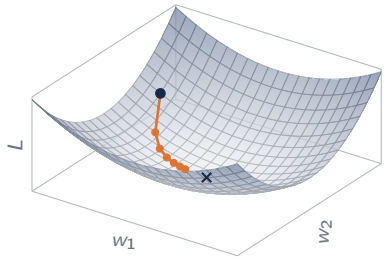
$\eta = 0{,}35$: перепрыгиваем



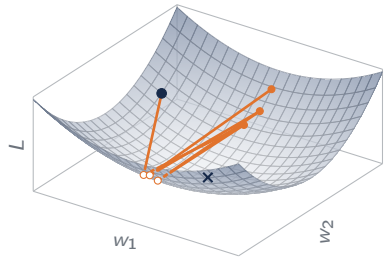
Слишком большой шаг

$L(w_1, w_2) = w_1^2 + 3w_2^2$, старт в одной и той же точке. Листайте: каждый клик — один шаг ГД. Пунктир — часть шага, скрытая за передней стенкой чаши.

$\eta = 0{,}1$: спускаемся



$\eta = 0{,}35$: перепрыгиваем



Справа шаг перелетает дно оврага, и каждый раз мы оказываемся всё выше: ГД расходится.

Расписание шага

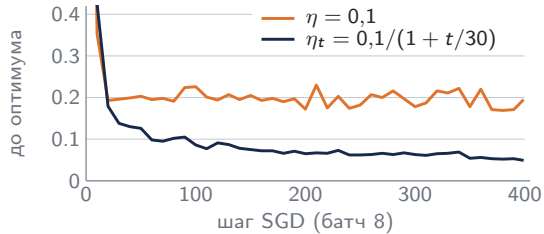
Проблема

Большой шаг быстро подводит к минимуму, но дальше прыгает вокруг него. Маленький шаг медленный с самого начала.

Решение

Начинаем с большого шага и уменьшаем его:

$$\eta_t = \frac{\eta_0}{1 + t/T}$$



Линейная регрессия на синтетике, 1000 объектов, среднее по 20 запускам. С постоянным шагом застреваем около 0,2, с убывающим доходим до 0,05. Другие расписания: делить шаг пополам каждые T шагов, умножать на $\gamma < 1$ на каждом шаге.

Итоги

Проверь себя

1. Почему фото одного листа нельзя раскидать по train и test?
2. Кросс-валидация дала $0{,}84 \pm 0{,}03$ и $0{,}85 \pm 0{,}03$. Какая модель лучше?
3. X имеет форму 500×4. Какой формы ∇L?
4. Чем шаг SGD отличается от шага ГД?
5. ГД расходится уже при маленьком шаге. Что проверить первым делом?

Ответы

1	Тест будет проверять память, а не обобщение: утечка
2	Сказать нельзя: разница меньше разброса
3	4×1, как у w
4	Градиент по случайному батчу, а не по всем объектам
5	Масштаб признаков и формулу градиента

На практике: восстановление коэффициентов

Даны точки $(x, f(x))$, где

$$f(x) = \big((a+\varepsilon_a)\sin x + (b+\varepsilon_b)\ln x\big)^2 + (c+\varepsilon_c)\,x^2, \qquad \varepsilon \in [-0{,}001;\ 0{,}001].$$

Нужно найти a, b, c с точностью 10^{-2}.

Подумайте

- Что здесь данные, а что параметры?
- Чему равна $\partial f/\partial a$? А $\partial f/\partial c$?

Задача: `https://coderun.yandex.ru/problem/coefficients-restoration`

На практике: восстановление коэффициентов

Даны точки $(x, f(x))$, где

$$f(x) = \big((a+\varepsilon_a)\sin x + (b+\varepsilon_b)\ln x\big)^2 + (c+\varepsilon_c)\,x^2, \qquad \varepsilon \in [-0{,}001;\ 0{,}001].$$

Нужно найти a, b, c с точностью 10^{-2}.

Подумайте

- Что здесь данные, а что параметры?
- Чему равна $\partial f/\partial a$? А $\partial f/\partial c$?

Подсказка

$\sin x$, $\ln x$, x^2 не зависят от a, b, c — это просто столбцы признаков:

x	$\sin x$	$\ln x$	x^2
0,5	0,479	−0,693	0,25
1	0,841	0,000	1
2	0,909	0,693	4

Задача: `https://coderun.yandex.ru/problem/coefficients-restoration`

Extra: матричное дифференцирование

Что такое производная по вектору и по матрице

Правило одно: **производная — это таблица частных производных, по одной на каждую пару «что меняется — по чему меняем».**

Что по чему	Результат	Форма	Пример
число по вектору w	градиент: $(\nabla f)_j = \partial f / \partial w_j$	как w	$w_1^2 + 3w_2 \;\to\; (2w_1;\ 3)$
число по матрице W	элемент (i,j): $\partial f / \partial W_{ij}$	как W	веса слоя нейросети
вектор g по вектору w	матрица Якоби: $J_{ij} = \partial g_i / \partial w_j$	$m \times k$	$Xw \;\to\; J = X$

Смысл везде тот же, что у обычной производной: при маленьком сдвиге h

$$f(w+h) \approx f(w) + \nabla f(w)^\top h, \qquad g(w+h) \approx g(w) + J\,h.$$

Цепное правило для векторов

Правило

Если $u = g(w)$, а $L = f(u)$ — число, то

$$\nabla_w L = J^\top \nabla_u f,$$

где J — матрица Якоби внутренней функции g. Как в школе: производная внешней, умноженная на производную внутренней.

Пример: MSE

$L = \|u\|^2, \;\; u = Xw - y.$

- внешняя: $\nabla_u \|u\|^2 = 2u$
- внутренняя: $J = X$

$$\nabla_w L = X^\top \cdot 2u = 2X^\top (Xw - y)$$

Вот откуда берётся $X^\top$: транспонированная матрица Якоби. Формы тоже сходятся: $(k \times n) \cdot (n \times 1) = k \times 1$.

Матричные производные: шпаргалка

Те же правила, что в школе, только для векторов. Здесь w, a — векторы длины k.

Одно число	Вектор	Почему
$(aw)' = a$	$\nabla_w (a^\top w) = a$	$a_1 w_1 + \cdots + a_k w_k$ по w_j: a_j
$(w^2)' = 2w$	$\nabla_w (w^\top w) = 2w$	$w_1^2 + \cdots + w_k^2$ по w_j: $2w_j$
$(aw^2)' = 2aw$	$\nabla_w (w^\top A w) = 2Aw$	если $A = A^\top$
$\big((xw - y)^2\big)' = 2x(xw - y)$	$\nabla_w \|Xw - y\|^2 = 2X^\top (Xw - y)$	цепное правило

В последней строке: внешняя функция — $\|u\|^2 = u^\top u$, внутренняя — $u = Xw - y$. Множитель x из одномерного случая превращается в $X^\top$.

Градиент MSE в одну строку

Раскроем квадрат

$$\|Xw - y\|^2 = w^\top X^\top X\, w \;-\; 2\, y^\top X w \;+\; y^\top y$$

По шпаргалке, слагаемое за слагаемым:

- $w^\top (X^\top X)\, w \;\to\; 2X^\top X w$ ($X^\top X$ симметричная)
- $-2\,(X^\top y)^\top w \;\to\; -2X^\top y$
- $y^\top y \;\to\; 0$: не зависит от w

$$\nabla_w \mathrm{MSE} = \frac{1}{n}\big(2X^\top X w - 2X^\top y\big) = \frac{2}{n} X^\top (Xw - y)$$

Сверим

Тот же ответ, что мы получили циклом по объектам. Два способа — одна формула.

Попробуйте

$f(w) = w^\top w + a^\top w$, где $a = (1;\ -2)$. Найдите ∇f и точку, где он равен нулю.

Градиент MSE в одну строку

Раскроем квадрат

$$\|Xw - y\|^2 = w^\top X^\top X\, w \;-\; 2\, y^\top X w \;+\; y^\top y$$

По шпаргалке, слагаемое за слагаемым:

- $w^\top (X^\top X)\, w \;\to\; 2X^\top X w$ ($X^\top X$ симметричная)
- $-2\,(X^\top y)^\top w \;\to\; -2X^\top y$
- $y^\top y \;\to\; 0$: не зависит от w

$$\nabla_w \mathrm{MSE} = \frac{1}{n}\big(2X^\top X w - 2X^\top y\big) = \frac{2}{n} X^\top (Xw - y)$$

Сверим

Тот же ответ, что мы получили циклом по объектам. Два способа — одна формула.

Попробуйте

$f(w) = w^\top w + a^\top w$, где $a = (1;\ -2)$. Найдите ∇f и точку, где он равен нулю.

Ответ: $\nabla f = 2w + a$, ноль при $w = -a/2 = (-0{,}5;\ 1)$.

Несколько слоёв: цепное правило по цепочке

Маленькая нейросеть: $\hat{y} = w_2\,\sigma(w_1 x)$, ошибка $L = (\hat{y} - y)^2$. Каждый слой — одна простая функция.



$$\frac{\partial L}{\partial w_1} = \underbrace{2(\hat{y} - y)}_{\partial L/\partial \hat{y}} \cdot \underbrace{w_2}_{\partial \hat{y}/\partial h} \cdot \underbrace{\sigma'(z)}_{\partial h/\partial z} \cdot \underbrace{x}_{\partial z/\partial w_1}$$

Это и есть backpropagation. Каждый слой знает только свою производную, а произведения справа налево переиспользуются для всех слоёв.

Backpropagation на числах

$x = 1$, $y = 1$, $w_1 = 0{,}5$, $w_2 = 2$, σ — сигмоида, $\sigma'(z) = \sigma(z)\big(1 - \sigma(z)\big)$.

Прямой проход

$z = w_1 x$	$= 0{,}5$
$h = \sigma(z)$	$\approx 0{,}622$
$\hat{y} = w_2 h$	$\approx 1{,}245$
$L = (\hat{y} - y)^2$	$\approx 0{,}060$

Обратный проход

$\partial L/\partial \hat{y} = 2(\hat{y} - y)$	$\approx 0{,}490$
$\partial L/\partial w_2 = 0{,}490 \cdot h$	$\approx 0{,}305$
$\partial L/\partial h = 0{,}490 \cdot w_2$	$\approx 0{,}980$
$\partial L/\partial z = 0{,}980 \cdot \sigma'(z)$	$\approx 0{,}230$
$\partial L/\partial w_1 = 0{,}230 \cdot x$	$\approx 0{,}230$

В настоящей сети вместо чисел векторы и матрицы, а произведения матричные: как $X^\top$ в градиенте MSE. PyTorch делает обратный проход сам: `loss.backward()`.

Extra: когда градиентный спуск сходится

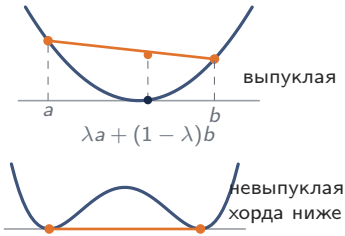
Выпуклая функция

Определение

f выпуклая, если для любых точек a, b и любого $\lambda \in [0, 1]$

$$f\big(\lambda a + (1-\lambda)b\big) \;\le\; \lambda f(a) + (1-\lambda) f(b).$$

Слева — функция в точке между a и b, справа — хорда над ней: **хорда не ниже графика.** Для гладкой функции то же самое: $f''(x) \ge 0$.



Зачем это ГД: у выпуклой функции любой локальный минимум — глобальный. Спускаясь вниз, не застрянем в «не той» яме.

Сильно выпуклая функция

Определение

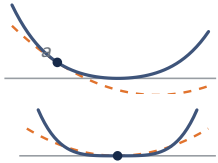
f сильно выпуклая с константой $\mu > 0$, если для любых a, b

$$f(b) \;\geq\; f(a) + f'(a)(b-a) + \frac{\mu}{2}(b-a)^2.$$

Если есть вторая производная: $f''(x) \geq \mu$.

«Растёт не медленнее параболы»

В любой точке a под график можно подложить параболу с кривизной μ. Плоских участков нет: от минимума функция растёт хотя бы как $\frac{\mu}{2}x^2$.



$\frac{1}{2}x^2 + \frac{1}{4}x^4$, $\mu = 1$: парабола под графиком

x^4: у дна ниже любой параболы

Смысл для ГД

Вдали от минимума градиент не крошечный: $|f'(x)| \geq \mu\,|x - x^*|$. В 0,1 от минимума наклон x^2 равен 0,2, у x^4 — 0,004. 100 шагов ГД с $\eta = 0{,}1$: на x^2 до $2 \cdot 10^{-10}$, на x^4 — до 0,11.

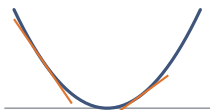
Когда ГД сходится: липшицев градиент

L-гладкость

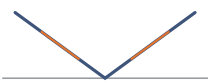
Градиент меняется не слишком быстро: $\|\nabla f(a) - \nabla f(b)\| \le L\,\|a - b\|$. Для одной переменной: $f''(x) \le L$.

Зачем это нужно

Если наклон резко меняется, длинный шаг по старому градиенту уводит не туда. При L-гладкости шаг $\eta < 2/L$ безопасен.



x^2: $f'' = 2$, значит $L = 2$



$|x|$: наклон прыгает с -1 на 1



x^4: $f'' = 12x^2$ растёт

Вспомните прошлую неделю: у параболы $(t-3)^2 + 2$ вторая производная равна 2, то есть $L = 2$ и $2/L = 1$. Отсюда условие сходимости $0 < \eta < 1$.

Что гарантирует теория

Условия	Шаг	Что получаем
L-гладкая и выпуклая	$\eta \le 1/L$	$f(w_k) - f^* \le \dfrac{\|w_0 - w^*\|^2}{2\eta k}$: ошибка падает как $1/k$
L-гладкая и μ-сильно выпуклая	$\eta = 1/L$	$\|w_k - w^*\|^2 \le \left(1 - \frac{\mu}{L}\right)^k \|w_0 - w^*\|^2$: как геометрическая прогрессия
L-гладкая, невыпуклая	$\eta \le 1/L$	$\nabla f(w_k) \to 0$, но это может быть локальный минимум или седло

Число обусловленности $\kappa = L/\mu$. Чем оно больше, тем медленнее множитель $1 - 1/\kappa$. У «оврага» $x^2 + 20y^2$ вторые производные 2 и 40: $\mu = 2$, $L = 40$, $\kappa = 20$. После нормализации линии уровня — круги, $\kappa \approx 1$.

MSE линейной регрессии выпукла всегда, а сильно выпукла, если признаки линейно независимы.