Омский Государственный ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Онлайн-курс

ВВЕДЕНИЕ В ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ

6. Инициализация весов и нормализация данных

Автор: Шевляков Артём Николаевич

Такие дела...

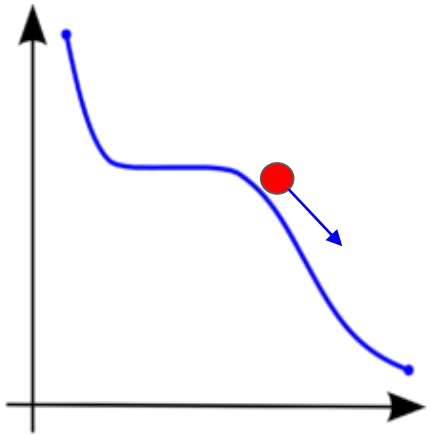
Общая схема тренировки НС

1. Задать архитектуру НС.
2. Выписать функцию потерь $L(w)$ по заданной ТВ.
3. Найти точку минимума функции потерь, которая даст оптимальные значения весов НС.
4. После этого НС может предсказывать значения для новых объектов (которых уже нет в ТВ).

Теоретически все просто! Но есть проблемы...

Проблемы НС

Начальная точка ГС может быть выбрана неудачно.
Мы можем оказаться сразу на плато или попасть в зону бесконечного спуска (см. рисунок).



Для борьбы с этим были придуманы нетривиальные алгоритмы. **Например, с попаданием на плато борются алгоритмы инициализации весов** и **нормализация данных**.

С бесконечным спуском косвенно борется регуляризация.

Проблемы глубоких НС

Если в НС очень много слоев, то у нее возникают дополнительные проблемы.

1. **Затухание сигнала при прохождении через сеть.**
   Методы борьбы: алгоритмы инициализации весов и нормализация по мини-батчам.

1. **Переобучение:** кода НС не пытается извлечь закономерность из данных, а тупо запоминает всю ТВ с её ответами (либо извлекает несуществующую закономерность).
   Метод борьбы: регуляризация и дропаут.

Онлайн-курс

ВВЕДЕНИЕ В ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ

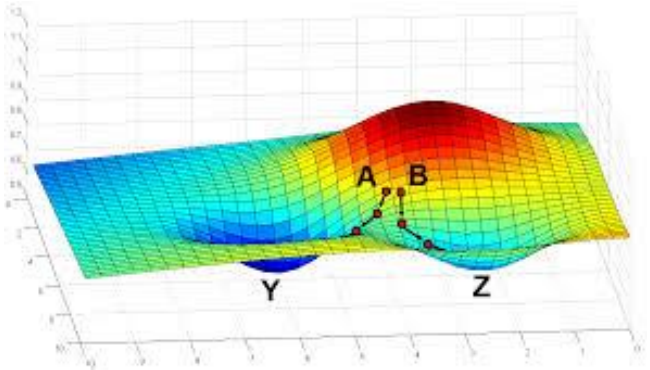
6. Инициализация весов и нормализация данных

Инициализация весов

Начальные значения весов НС

В ГС нужно выбрать начальную точку для спуска, то есть начальные значения весов НС.

Обычно эта точка выбирается случайно. Но от нее существенно зависит успех ГС.



По этой причине веса нужно инициализировать правильно.

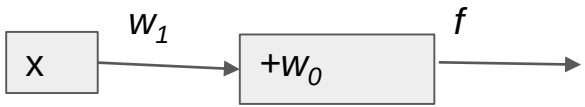
Альтернативный взгляд на проблему:

Казалось бы, задача невозможная: найти такие начальные значения весов, которые лежат вблизи глубокой точки минимума функции потерь, не используя информации о функции потерь.

Пример (поучительный)

От начальных значений весов действительно много что зависит.
Если начальные значения весов подобраны неудачно,
то мы окажемся на плато, откуда ГС никуда не пойдёт.

Рассмотрим простую НС:



где ФА $f(x)=Relu(x)=r(x)$.

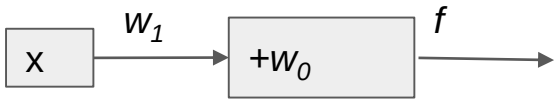
НС вычисляет функцию $F_{NN}(x)=r(xw_1+w_0)$

Пример

Берем такую ТВ:

Объекты	x	Y
A_1	-1	1
A_2	0	0
A_3	1	1
A_4	2	4

и составляем функцию потерь:



$$L(w)=(r(-w_1+w_0)-1)^2+(r(w_0)-0)^2+(r(w_1+w_0)-1)^2+(r(2w_1+w_0)-4)^2$$

Пример

$L(w)=(r(-w_1+w_0)-1)^2+(r(w_0)-0)^2+(r(w_1+w_0)-1)^2+(r(2w_1+w_0)-4)^2$.

ЧП по переменной w_1:

$-2(r(-w_1+w_0)-1)r'(-w_1+w_0)+2(r(w_1+w_0)-1)r'(w_1+w_0)+4(r(2w_1+w_0)-4)r'(2w_1+w_0)$

ЧП по переменной w_0:

$2(r(-w_1+w_0)-1)r'(-w_1+w_0)+2r(w_0)r'(w_0)+2(r(w_1+w_0)-1)r'(w_1+w_0)+2(r(2w_1+w_0)-4)r'(2w_1+w_0)$

Если текущие значения w_1,w_0 таковы, что все выражения $-w_1+w_0$, w_0, w_1+w_0,$2w_1+w_0$ меньше или равны 0, то градиент равен 0 в данной точке (то есть это плато).
А таких значений w_1,w_0 немало.

Пример

Если, например, начальные значения весов равны $w_1=0$, $w_0=0$, то все ЧП равны 0 в точке $a_0=(0,0)$.

$dL/dw_1=-2(r(-w_1+w_0)-1)r'(-w_1+w_0)+2(r(w_1+w_0)-1)r'(w_1+w_0)+4(r(2w_1+w_0)-4)r'(2w_1+w_0)$

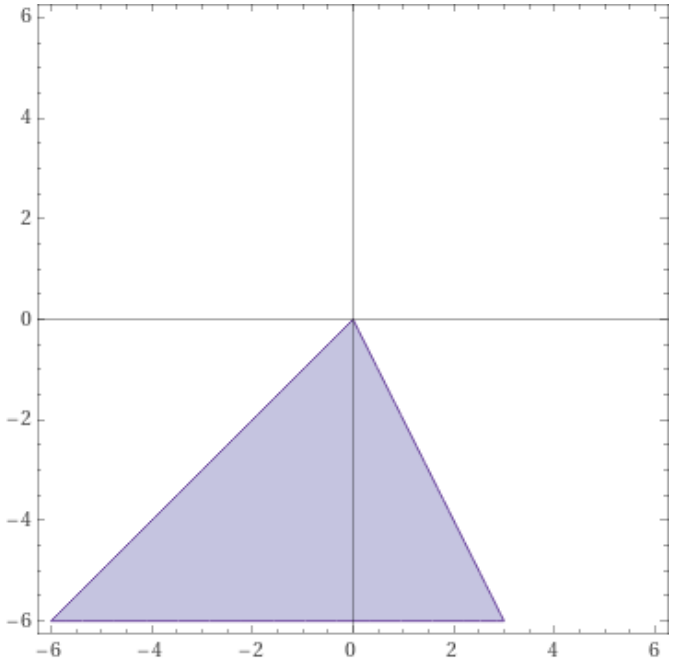
$dL/dw_0=2(r(-w_1+w_0)-1)r'(-w_1+w_0)+2r(w_0)r'(w_0)+2(r(w_1+w_0)-1)r'(w_1+w_0)+2(r(2w_1+w_0)-4)r'(2w_1+w_0)$

Это означает, что следующая точка ГС будет:
$a_2=(0,0)-h(0,0)=(0,0)=a_1$, то есть процесс ГС остановится.

Таких «плохих» значений w_1, w_0 немало. См. следующий слайд...

Пример

Чтобы найти все «плохие» начальные значения w_1, w_0, решим систему неравенств:



$$\begin{cases} -w_1 + w_0 \le 0, \\ w_0 \le 0, \\ w_1 + w_0 \le 0, \\ 2w_1 + w_0 \le 0 \end{cases}$$

– вот ее область решений.

Начальных значений для w_1, w_0 на которых ГС сразу оказывается на плато очень много.

А как избегать таких точек в качестве первого приближения?

Хорошее начальное приближение

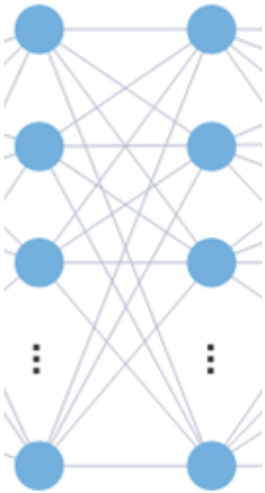
Начальные значения весов должны быть далеко от плато!
Иначе на первых итерациях ГС мы столкнемся с затуханиями градиента и процесс самопроизвольно остановится.

Во-вторых, метод инициализации весов должен хорошо работать для глубоких НС.

Нужно найти правило следующего рода: «генерировать начальные значения весов сети с помощью случайных чисел из интервала *[a,b]*» (числа *a, b* нужно будет найти).

Будем наблюдать за НС...

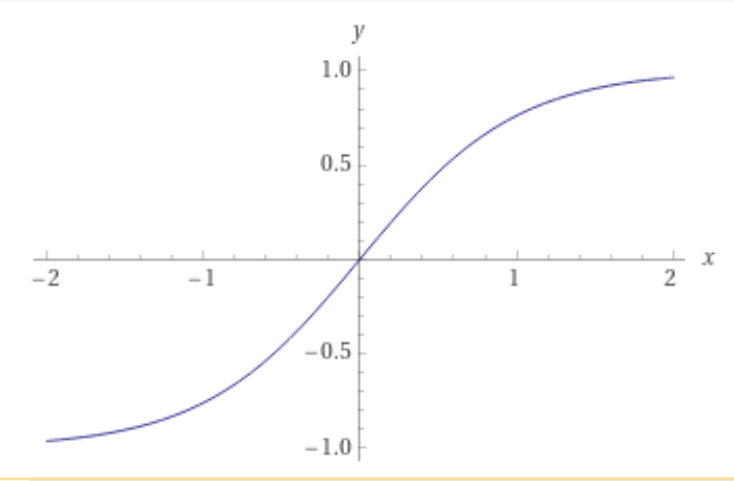
Есть пара полносвязных слоев. Считаем, что в них одинаковое число нейронов *n*. Следовательно, в каждый нейрон правого слоя входит *n* связей (и у каждой свой вес).



Будем наблюдать за НС...

Будем считать, что ФА *f* у всех нейронов одинаковая и удовлетворяет условиям: $f(x) \approx x,\ f'(x) \approx 1$

Например, таким условиям удовлетворяет tanh(x):



Также будем считать, что **веса-смещения всех нейронов равны 0** (или близки к нему).

Какими же должны быть веса-связи между нейронами?

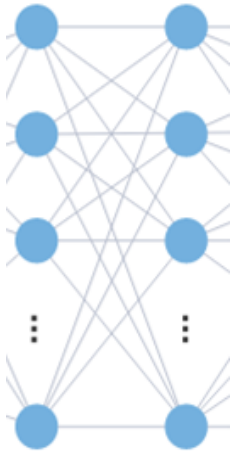
Будем наблюдать за НС...

Допустим, что каждый из нейронов левого слоя **до** применения ФА имеет выходное значение в диапазоне *[-a,a]*.
Тогда из этих нейронов выходят значения из *[f(-a),f(a)]=[-f(a),f(a)]*.

Допустим, что веса связей между слоями принадлежат отрезку *[-b,b]* (число *b* мы как раз хотим найти). Тогда каждый из нейронов правого слоя получает на вход число из диапазона *[-nbf(a),nbf(a)]*.

Поскольку смещения равны 0, то в итоге получаем, что при переходе между слоями сигнал из интервала ***[-a,a]*** переходит в сигнал из интервала ***[-nbf(a),nbf(a)]***.

Какой отрезок больше и что это значит?



Будем наблюдать за НС...

Из свойств ФА получаем, что нужно сравнивать отрезки: *[-a,a] [-nba,nba].*

Если *nba<a,* то это **плохо**! Диапазон сигналов нейронов уменьшается, и при достаточно глубокой сети последний слой будет выдавать константу 0.

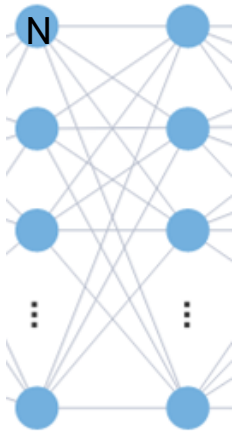
Получили пока одно условие: $nba \geq a \Rightarrow b \geq 1/n$
– **запомним его.**

Будем наблюдать за НС...

Теперь посмотрим, какие значения ЧП при back-prop возникают в нашей сети. Пусть все y(x) – значение, выдаваемое некоторым нейроном N y из левого слоя; $y_1(x),\ldots,y_n(x)$ – выходы из нейронов правого слоя.
Значения $y_1(x),\ldots,y_n(x)$ зависят от y(x):

$y_i(x)=f(\ldots+y(x)w_i+\ldots)$

где w_i – вес связи между нейроном N и i-м нейроном правого слоя.



И поэтому:

$$\partial y_i/\partial y = f'(\text{input of the neuron } y_i)w_i$$

Продемонстрируем это на примере.

Будем наблюдать за НС...

$$\partial y_i / \partial y = f'(\text{input of the neuron } y_i) w_i$$

Из выбора ФА *f* получаем: $\partial y_i / \partial y \approx w_i$

Тогда ЧП функции потерь *L* равны: $$\frac{\partial L}{\partial y} = \frac{\partial L}{\partial y_1}\frac{\partial y_1}{\partial y} + \ldots + \frac{\partial L}{\partial y_n}\frac{\partial y_n}{\partial y}$$

Их можно оценить как:

$$\frac{\partial L}{\partial y} \approx w_1 \frac{\partial L}{\partial y_1} + \ldots + w_n \frac{\partial L}{\partial y_n}$$

Веса w_i принадлежат интервалу *[-b,b]*. Пусть ЧП с правого слоя $\partial L / \partial y_i$ принадлежат некоторому интервалу *[-c,c]*.

Тогда

$$\partial L / \partial y \in [-nbc, nbc]$$

Будем наблюдать за НС...

Итак, ЧП с правого слоя принадлежат интервалу *[-c,c]*, а ЧП с левого слоя принадлежат интервалу *[-nbc,nbc]*.

Если *nbc<c*, то это **плохо**! Это означает, что когда алгоритм back-prop дойдёт до первого слоя сети, значения ЧП будут почти равны нулю (затухание градиента). Таким образом, мы снова получаем такое же условие

$$nbc \geq c \Rightarrow b \geq 1/n$$

Вывод: начальные значения весов-связи между нейронами полносвязного слоя нужно задавать случайными числами из диапазона *[-b,b]*, где $b \geq 1/n$

n - количество нейронов в каждом из двух полносвязных слоев.

В частности, (и для простоты) **можно выбрать диапазон** *[-1/n,1/n]*.

А если в слоях разное число нейронов?

В этом случае начальные значения весов-связей между нейронами полносвязного слоя нужно задавать случайными числами из диапазона

$$[-1/n,1/n],$$

где

$$n=(n_1+n_2)/2$$

– среднее арифметическое числа нейронов в соседних слоях.

Инициализация Ксавье

Выше была объяснена (весьма неформально) суть **инициализации Ксавье** весов НС:

веса-связи инициализируются случайными числами из интервала [-*1/n,1/n*], веса-смещения инициализируются нулями.

На самом деле веса генерируются не из интервала *[-1/n,1/n]*,
а с помощью равномерного распределения **с дисперсией** *1/n*.

Подробности — в книге →

Инициализация Ксавье

Практика показывает, что инициализация Ксавье улучшает обучение НС и повышает точность предсказания.

Также в этой книге можно узнать
про **инициализацию Хе**,
которая хорошо работает с ФА Relu.

Онлайн-курс

ВВЕДЕНИЕ В ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ

6. Инициализация весов и нормализация данных

Нормализация данных

Что такое нормализация?

Нормализация – это приведение данных к одному масштабу.

Например, часть данных у вас выражена в кг с диапазоном значений [0,100], а часть данных – в градусах Цельсия с возможными значениями [-30,40]. После нормализации все данные становятся безразмерными, и принадлежащими одному общему диапазону.

Оказывается, данные ТВ нужно нормализировать перед тренировкой НС.

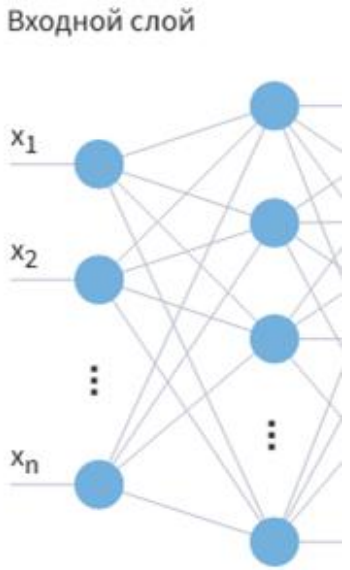
А почему?

Идея нормализации

Необходимость нормализации следует из наших рассуждений об алгоритме инициализации весов.

Там мы предполагали, что нейроны некоторого слоя НС **до применения** ФА выдают значения из диапазона *[-a,a]*. Это предположение должно быть верным для любого слоя НС, в том числе и **для входного слоя**.

Но позвольте! Признаки объектов x_i могут иметь разный масштаб и природу (например, x_1={*пол человека*}, x_2={*рост*}).
Как их загнать в один диапазон?



Идея нормализации...

… состоит в том, чтобы **все признаки объекта привести к одному масштабу**. Нормализация осуществляется с помощью «умных» формул, ниже будут указаны два самых популярных способа нормализации данных.

По сути нормализация – это преобразование **нецелевых столбцов** в таблице с данными

Объект	Признак P_1	…	Признак P_n	Целевой признак Y
A_1	p_{11}	…	p_{1n}	y_1
A_2	p_{21}	…	p_{2n}	y_2
…	…	…	…	…
A_m	p_{m1}	…	p_{mn}	y_m

1-й способ нормализации

Для каждого столбца X делаем следующее:

Пусть $X=\{x_1,\ldots,x_m\}$ – все данные из столбца X, $a=min\{x_1,\ldots,x_m\}$, $b=max\{x_1,\ldots,x_m\}$ - максимальное и минимальное значение в столбце.

Тогда нормализованный столбец состоит из значений: $$x'_i := \frac{x_i - a}{b - a}$$

После нормализации значения в столбце *X* принадлежат интервалу...

1-й способ нормализации

Для каждого столбца X делаем следующее:

Пусть $X=\{x_1,\ldots,x_m\}$ – все данные из столбца X, $a=min\{x_1,\ldots,x_m\}$, $b=max\{x_1,\ldots,x_m\}$ - максимальное и минимальное значение в столбце.

Тогда нормализованный столбец состоит из значений: $$x_i' := \frac{x_i - a}{b - a}$$

После нормализации значения в столбце *X* принадлежат интервалу *[0,1]*

2-й способ нормализации

Этот способ требует знания некоторых понятий из математической статистики.

Пусть $X=\{x_1,\ldots,x_m\}$ – все данные из столбца X. **Тогда величина**

$$\bar{x} := \frac{x_1 + \ldots + x_m}{m}$$

называется средним значением признака X.

Задачка: если из всех значений признака X вычесть его среднее значение, а потом вычислить среднее значение нового признака, то оно будет равно...

2-й способ нормализации

Этот способ требует знания некоторых понятий из математической статистики.

Пусть $X=\{x_1,\ldots,x_m\}$ – все данные из столбца X. **Тогда величина**

$$\bar{x} := \frac{x_1 + \ldots + x_m}{m}$$

называется средним значением признака X.

Задачка: если из всех значений признака X вычесть его среднее значение, а потом вычислить среднее значение нового признака, то оно будет равно **0**.

2-й способ нормализации

Пусть $X=\{x_1,\ldots,x_m\}$ – все данные из столбца X. **Величина**

$$s_X = \sqrt{\frac{1}{m-1}((x_1-\bar{x})^2+\ldots+(x_m-\bar{x})^2)}$$

называется отклонением признака *X*.

Тогда нормализованный столбец состоит из значений:

$$x'_i := \frac{x_i-\bar{x}}{s_X}$$

2-й способ нормализации

Можно доказать, что нормализованный столбец имеет среднее значение 0 и отклонение 1.

Что это даёт?
С вероятностью 95-99% значения нормализованного столбца лежат в диапазоне *[-3,3]* (**18+: правило трёх сигм**).

Пример нормализации

Для нашей старой таблички

Объекты	x	Y
A_1	-1	1
A_2	0	0
A_3	1	1
A_4	2	4

признак X после нормализации равен

Пример нормализации

1й способ:

Объекты	x	Y
A_1	(-1-(-1))/(2-(-1))=0	1
A_2	(0-(-1))/(2-(-1))=1/3	0
A_3	(1-(-1))/(2-(-1))=2/3	1
A_4	(2-(-1))/(2-(-1))=1	4

Пример нормализации

2й способ:

среднее значение=0.5;

отклонение $s = \sqrt{1/3((-1-0.5)^2 + (0-0.5)^2 + (1-0.5)^2 + (2-0.5)^2)} = 1.3$

Объекты	x	Y
A_1	(-1-0.5)/1.3=-1.15	1
A_2	(0-0.5)/1.3=-0.38	0
A_3	(1-0.5)/1.3=0.38	1
A_4	(2-0.5)/1.3=1.15	4

Выводы по нормализации

Данные тренировочной выборки нужно нормализовать перед началом тренировки НС.

Когда сеть будет натренирована, то при получении нового объекта для предсказания **его тоже нужно нормализовать, используя числовые характеристики ТВ**.

Пример нормализации

После нормализации таблицы

Объекты	x	Y
A_1	-1	1
A_2	0	0
A_3	1	1
A_4	2	4

мы тренируем НС. Когда НС построена, то она должна предсказывать значение целевого признака для новых объектов.
Но эти объекты нужно нормализовать. Например, берем новый объект:

A	-2	?

Пример нормализации

1й способ:

Объекты	x	Y
A_1	(-1-(-1))/(2-(-1))=0	1
A_2	(0-(-1))/(2-(-1))=1/3	0
A_3	(1-(-1))/(2-(-1))=2/3	1
A_4	(2-(-1))/(2-(-1))=1	4

даёт такое преобразование объекта:

A	(-2-(-1))/(2-(-1))=-1/3	?

Пример нормализации

2й способ:

Объекты	x	Y
A_1	(-1-0.5)/1.3=-1.15	1
A_2	(0-0.5)/1.3=-0.38	0
A_3	(1-0.5)/1.3=0.38	1
A_4	(2-0.5)/1.3=1.15	4

даёт такую нормализацию объекта:

A	(-2-0.5)/1.3=-1.92	?

Онлайн-курс

ВВЕДЕНИЕ В ИСКУССТВЕННЫЕ НЕЙРОННЫЕ СЕТИ

6. Инициализация весов и нормализация данных

Выводы

Выводы:

- Мы показали, насколько важной является правильная инициализация весов.
- Указали диапазон инициализации весов.
- Мы обсудили важность нормализации данных и указали два способа достижения оной.