Тятя! Тятя! Нейросети заменили продавца!

Ульянкин Филипп

Аннотация

В этой виньетке собрана коллекция ручных задачек про нейросетки на пару томных вечеров. Вместе с Машей можно попробовать по маленьким шажкам с ручкой и бумажкой раскрыть у себя в теле несколько чакр и немного глубже понять модели глубокого обучения[1].

Вместо введения

Однажды Маша услышала про какой-то Машин лёрнинг. Она сразу же смекнула, что именно она та самая Маша, кому этот лёрнинг должен принадлежать. Ещё она смекнула, что если хочет владеть лёрнингом по праву, ни одна живая душа не должна сомневаться в том, что она шарит. Поэтому она постоянно изучает что-то новое.

Её друг Миша захотел стать адептом Машиного лёрнинга, и спросил её о том, как можно за вечер зашарить алгоритм обратного распространения ошибки. Тогда Маша открыла свою первую книгу по глубокому обучению и прочитала в ней:

> Благодаря символическому дифференцированию вам никогда не придется заниматься реализацией агоритма обратного распространения вручную. Поэтому не будем тратить время на его формулировку[2].

Маше такая логика показалась странной. Поэтому она взяла книгу с более глубокой математикой. Там она прочитала, что:

Николенко

Тогда Маша взяла Библию глубокого обучения[3] и поняла, что по ней за один вечер точно не разберёшься. Слишком серьёзно всё написано. Для вечерних разборок нужно что-то более инфантильное.

У Маши оставался один выход: поскрести по лёрнингу и собрать инфантильную коллекцию ручных задачек, прорешивая которую новые адепты Машиного лёрнинга могли бы открывать у себя во чакру за чакрой. Так и появилась эта виньетка.

[1] Ахахах глубже глубокого, ахахах

[2] Франсуа Шолле, Глубокое обучение на Python, стр. 77

[3] Goodfellow I., Bengio Y., Courville A. Deep learning. – MIT press, 2016.

Содержание

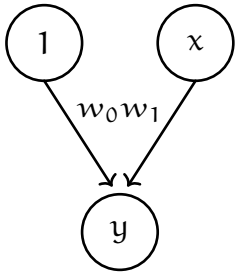
1 Всего лишь функция

Ты всего лишь машина, только имитация жизни. Робот сочинит симфонию? Робот превратит кусок холста в шедевр искусства?

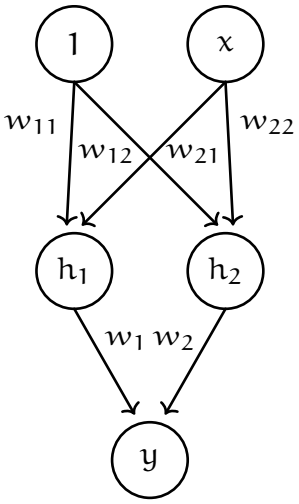
Из фильма «Я, робот» (2004)

Упражнение 1 (от регрессии к нейросетке)

Однажды вечером, по пути с работы[4] Маша зашла в свою любимую кофейню на Тверской. Там, на стене, она обнаружила очень интересную картину:



Хозяин кофейни, Добродум, объяснил Маше, что это Покрас-Лампас так нарисовал линейную регрессию,[5] и её легко можно переписать в виде формулы: $y_i = w_0 + w_1 \cdot x_i$. Пока Добродум готовил кофе, Маша накидала у себя на бумажке новую картинку:



а. Как такая функция будет выглядеть в виде формулы?

б. Правда ли, что y будет нелинейно зависеть от x?

в. Если нет, как это исправить и сделать зависимость нелинейной?

[4] она работает рисёрчером.

[5] эксклюзивный заказ был

Решение:

Когда мы переписывали картинку в виде уравнения регрессии, мы брали вход из кругляшей, умножали его на веса, написанные около стрелок и искали сумму.

Сделаем ровно то же самое для Машиной картинки. Буквы h внутри кругляшей скрытого слоя будут считаться как:

$$h_1 = w_{11} \cdot 1 + w_{21} \cdot x$$
$$h_2 = w_{12} \cdot 1 + w_{22} \cdot x$$

Итоговый y будет складываться из ашек:

$$y = w_1 \cdot h_1 + w_2 \cdot h_2.$$

Раскрываем h-ки и получаем для y итоговое уравнение:

$$\begin{aligned} y = w_1 \cdot h_1 + w_2 \cdot h_2 = \\ = w_1 \cdot (w_{11} + w_{21} \cdot x) + w_2 \cdot (w_{12} + w_{22} \cdot x) = \\ = \underbrace{(w_1 w_{11} + w_2 w_{12})}_{\gamma_1} + \underbrace{(w_1 w_{21} + w_2 w_{22})}_{\gamma_2} \cdot x \end{aligned}$$

Когда мы раскрыли скобки, мы получили ровно ту же самую линейную регрессию. Правда мы зачем-то довольно сложно параметризовали γ_1 и γ_2 через шесть параметров.

Чтобы сделать зависимость нелинейной, нужно немного преобразить каждую из h_i, взяв от них какую-нибудь нелинейную функцию. Например, сигмоиду:

$$f(h) = \frac{1}{1 + e^{-h}}.$$

Тогда формула преобразиться:

$$y = w_1 \cdot f(w_{11} + w_{21} \cdot x) + w_2 \cdot f(w_{12} + w_{22} \cdot x).$$

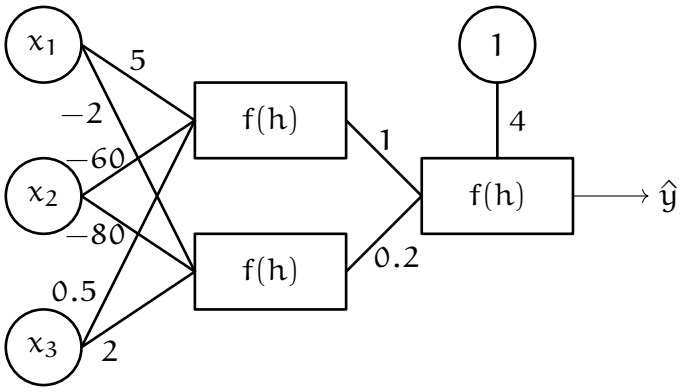
~~Смерти~~ Линейности больше нет. **Только что на ваших глазах произошло чудо. Регрессия превратилась в нейросеть.** Можно использовать вместо сигмоиды любую другую функцию. Например,

$$\mathrm{ReLU}(h) = \max(0, h).$$

Она называется релу.[6] Она простая и обычно для нелинейности её хватает. Но об этом поговорим позже.

Упражнение 2 (из картинки в формулу)

Добродум хочет понять насколько сильно будет заполнена кофейня в следующие выходные. Для этого он обучил нейросетку. На вход она принимает три фактора: температуру за окном, x_1, факт наличия на Тверской митинга, x_2 и пол бариста на смене, x_3. В качестве функции активации Добродум использует ReLU.



а. В эти выходные за барной[7] стойкой стоит Агнесса. Митинга не предвидится, температура будет в районе 20 градусов. Сколько человек придёт в кофейню к Добродуму?

б. На самом деле каждая нейросетка — это просто-напросто какая-то нелинейная сложная функция. Запишите нейросеть Добродума в виде функции.

Решение:

Будем постепенно идти по сетке и делать вычисления. Подаём все значения в первый нейрон, получаем:

$$h_1 = max(0, 5 \cdot 20 + (-60) \cdot 0 + 0.5 \cdot 1) = max(0, 100.5) = 100.5$$

Ровно то же самое делаем со вторым нейроном:

$$h_2 = max(0, -2 \cdot 20 + (-80) \cdot 0 + 2 \cdot 1) = max(0, -38) = 0$$

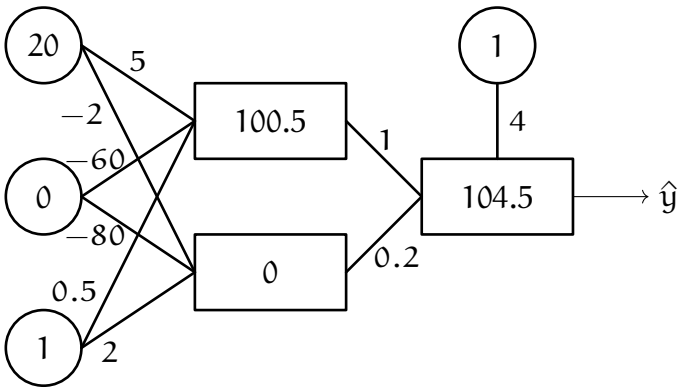
Дальше результат скрытых нейронов идёт во второй слой:

$$\hat{y} = max(0, 1 \cdot 100.5 + 0.2 \cdot 0 + 4 \cdot 1) = 104.5$$

Это и есть итоговый прогноз.

[6] внезапное название

[7] барной... конечно, кофейня у него...



Теперь по мотивам наших вычислений запишем нейронку как функцию. Начинать будем с конца:

$$\hat{y} = f(1 \cdot h_1 + 0.2 \cdot h_2 + 4 \cdot 1)$$

Подставляем вместо h_1 и h_2 вычисления, которые происходят на первом слое нейронки:

$$\begin{aligned}\hat{y} &= f(1 \cdot f(5 \cdot x_1 - 60 \cdot x_2 + 0.5 \cdot x_3) + 0.2 \cdot f(-2 \cdot x_1 - 80 \cdot x_2 + 2 \cdot x_3) + 4 \cdot 1) = \\ &= \max(0, \max(0, 5 \cdot x_1 - 60 \cdot x_2 + 0.5 \cdot x_3) + 0.2 \cdot \max(0, -2 \cdot x_1 - 80 \cdot x_2 + 2 \cdot x_3) + 4).\end{aligned}$$

Обучение нейронной сетки, на самом деле, эквивалентно обучению такой сложной нелинейной функции.

Упражнение 3 (из формулы в картинку)

Маша написала на бумажке функцию:

$$y = \max(0, 4 \cdot \max(0, 3 \cdot x_1 + 4 \cdot x_2 + 1) + 2 \cdot \max(0, 3 \cdot x_1 + 2 \cdot x_2 + 7) + 6)$$

Теперь она хочет, чтобы кто-нибудь из её адептов нарисовал её в виде нейросетки. Нарисуй.

Решение:

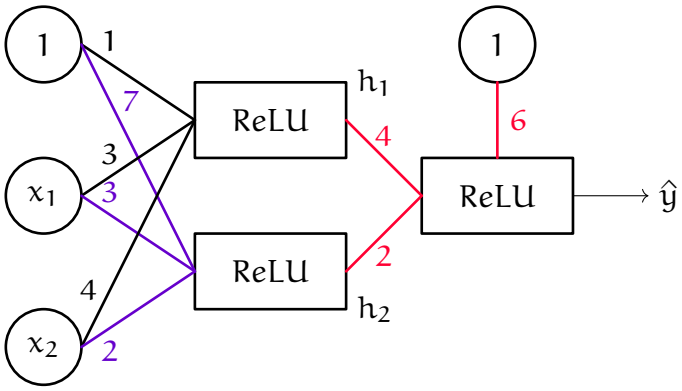
Начнём рисовать картинку с конца. На выход выплёвывается либо 0, либо комбинация из двух входов:

$$\hat{y} = \mathrm{ReLU}(4 \cdot h_1 + 2 \cdot h_2 + 6)$$

Каждый из входов — это снова либо 0, либо комбинация из двух входов.

$$y = \max(0, 4 \cdot \underbrace{\max(0, 3 \cdot x_1 + 4 \cdot x_2 + 1)}_{h_1} + 2 \cdot \underbrace{\max(0, 3 \cdot x_1 + 2 \cdot x_2 + 7)}_{h_2} + 6)$$

Получается, что на первом слое находится два нейрона, которые передают свои выходы в третий:

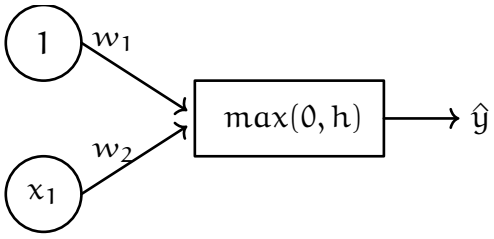


Упражнение 4 (армия регрессий)

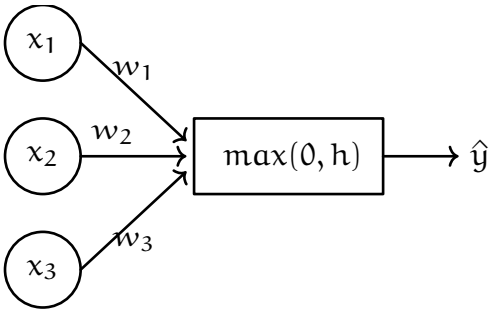
Парни очень любят Машу,[8] а Маша с недавних пор любит собирать персептроны и думать по вечерам об их весах и функциях активации. Сегодня она решила разобрать свои залежи из персептронов и как следует упорядочить их.

а. В ящике стола Маша нашла персептрон с картинки 1 Маша хочет подобрать веса так, чтобы он реализовывал логическое отрицание, то есть превращал $x_1 = 0$ в $y = 1$, а $x_1 = 1$ в $y = 0$ (так работает логичская функция: отрицание).

Картинка 1



Картинка 2



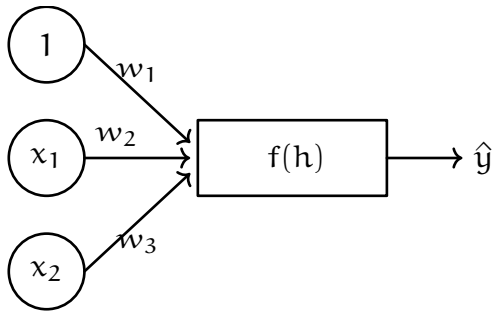
б. В тумбочке, среди носков, Маша нашла персептрон, с картинки 2, Маша хочет подобрать такие веса w_i, чтобы персептрон превращал x из таблички в соответствующие y:

x_1	x_2	x_3	y
1	1	2	0.5
1	−1	1	0

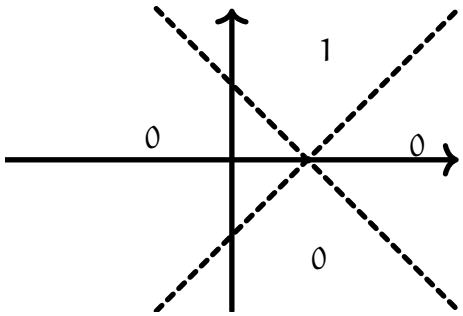
[8] когда у тебя есть лёрнинг, они так и лезут

в. Оказывается, что в ванной всё это время валялась куча персептронов с картинки 3 с неизвестной функцией активации (надо самому выбирать).

Картинка 3



Картинка 4



Маша провела на плоскости две прямые: $x_1 + x_2 = 1$ и $x_1 - x_2 = 1$. Она хочет собрать из персептронов нейросетку, которая будет классифицировать объекты с плоскости так, как показано на картинке 4.

Решение:

а. Начнём с первого пункта. Чтобы было легче запишем нейрон в виде уравнения:

$$\hat{y} = max(0, w_1 + w_2 \cdot x_1).$$

Нам нужно, чтобы

$$max(0, w_1 + w_2 \cdot 1) = 0$$
$$max(0, w_1 + w_2 \cdot 0) = 1$$

Из второго уравнения сразу получаем, что $w_1 = 1$, а w_2 на второе уравнение никак не влияет. Для того, чтобы в первом уравнении получить ноль, нужно взять $w_2 \leqslant -1$. Нейрон готов.

б. Снова выписываем несколько уравнений:

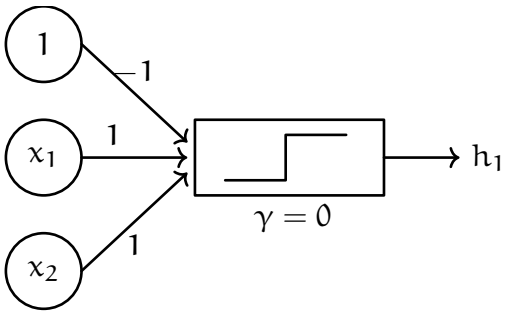
$$max(0, w_1 + w_2 + 2 \cdot w_3) = 0.5$$
$$max(0, w_1 - w_2 + w_3) = 0$$

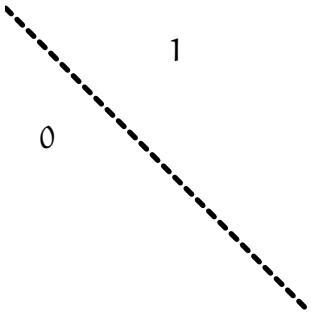
Тут решений может быть довольно много. Первое, что приходит в голову — это занулить w_1 и w_3 в первом уравнении, а w_2 поставить 0.5. Тогда во втором уравнении мы сразу же будем оказываться в отрицательной области и ReLU заботливо будет отдавать нам 0.

в. Давайте для разнообразия возьмём в качестве $f(h)$ пороговую функцию потерь

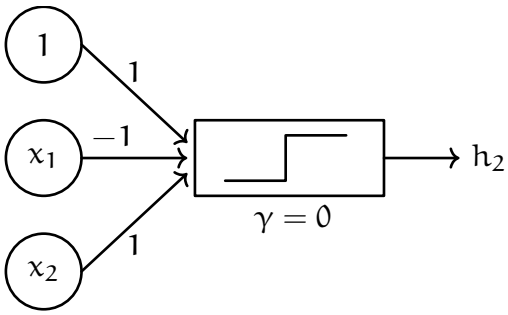
$$f(h) = \begin{cases} 1, h > 0 \\ 0, h \leqslant 0 \end{cases}.$$

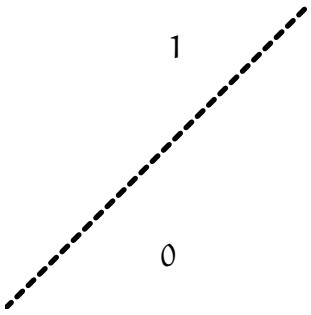
Один нейрон — это одна линия, проведённая на плоскости. Эта линия отделяет один класс от другого. Например, линию $x_1 + x_2 - 1 = 0$ мог бы описать нейрон





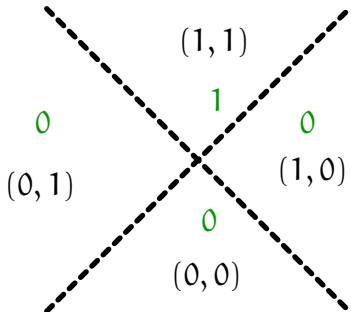
Порог γ для кусочной функции в каком-то смысле дублирует константу. Будем всегда брать его нулевым. Видим, что если мы получили комбинацию x_1, x_2 и 1, большую, чем ноль, мы оказались справа от прямой. Если хочется поменять метку 0 и 1 сторонами, можно умножить все коэффициенты на -1. **Наш персептрон понимает по какую сторону от прямой мы оказались,** то есть задаёт одну линейную разделяющую поверхность. По аналогии для второй прямой мы можем получить:

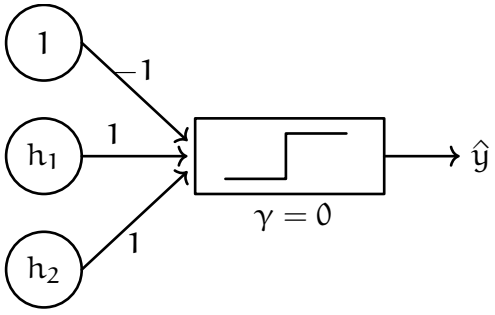




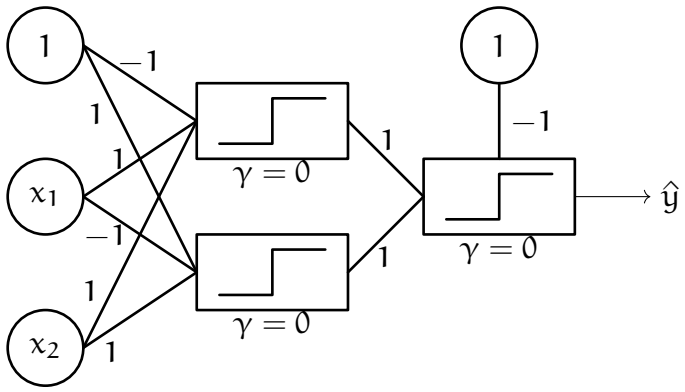
Итак, первый персептрон выбрал нам позицию относительно первой прямой, второй относительно второй. Остаётся только соединить эти результаты в один. Нейрон для скрепки должен реализовать для нас логическую функцию, которую задаёт табличка ниже. Там же нарисованы примеры весов, которые могли бы объединить выхлоп первого слоя в итоговый прогноз.

h_1	h_2	$\hat{y}$
1	1	1
1	0	0
0	1	0
0	0	0





Теперь мы можем нарисовать итоговую нейронную сеть, решающую задачу Маши. Она состоит из двух слоёв. Меньше не выйдет, так как каждый персептрон строит только одну разделяющую линию.



Кстати говоря, если бы мы ввели для нашей нейросетки дополнительный признак $x_1 \cdot x_2$, у нас бы получилось обойтись только одним персептроном. В нашей ситуации **нейросетка сама сварила на первом слое признак $x_1 \cdot x_2$, которого ей не хватало.**

Упражнение 5 (логические функции)

Маша вчера поссорилась с Пашей. Он сказал, что у неё нет логики. Чтобы доказать Паше обратное, Маша нашла теорему, которая говорит о том, что с помощью нейросетки можно аппроксимировать почти любую функцию, и теперь собирается заняться апроксимацией логических функций. Для начала она взяла самые простые, заданные следующими таблицами истинности:

x_1	x_2	$x_1 \cap x_2$
1	1	1
1	0	0
0	1	0
0	0	0

x_1	x_2	$x_1 \cup x_2$
1	1	1
1	0	1
0	1	1
0	0	0

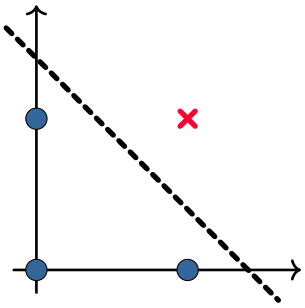
x_1	x_2	x_1 XoR x_2
1	1	0
1	0	1
0	1	1
0	0	0

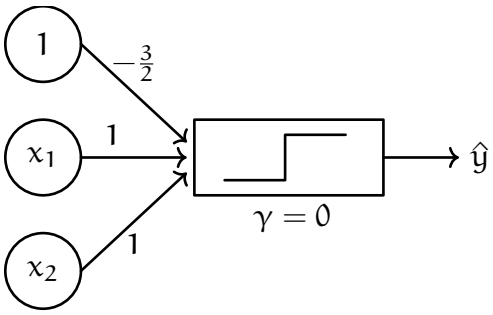
Первые два столбика идут на вход, третий получается на выходе. Первая операция — логическое "и вторая — "или". Операция из третьей таблицы называется "исключающим или"(XoR). Если внимательно приглядеться, то можно заметить, что XoR — это то же самое что и $[x_1 \neq x_2]$[9].

Решение:

На самом деле в предыдущем упражнении мы уже построили нейрон для пересечения, когда нам нужно было оказаться два раза по правильную сторону прямой. Посмотрим на тот же нейрон под другим углом:

[9]Тут квадратные скобки обозначают индикатор. Он выдаёт 1, если внутри него стоит правда и 0, если ложь.

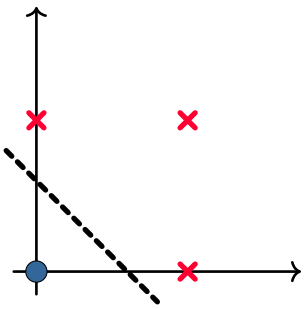


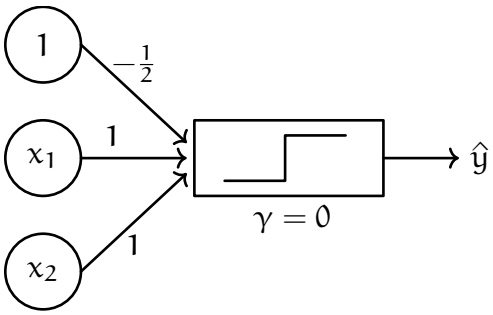


Если нарисовать все наши четыре точки на плоскости, становится ясно, что мы хотим отделить точку $(1, 1)$ от всех остальных. Сделать это можно практически любой линией. Например, в нейроне выше задана линия $x_2 = 1.5 - x_1$. Подойдёт и любая другая, отделяющая крест от точек. Пропустим ради приличия точки через наш нейрон:

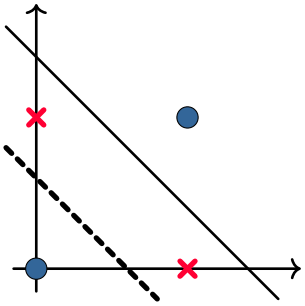
$$[-1.5 + 1 + 1 > 0] = [0.5 > 0] = 1$$
$$[-1.5 + 0 + 0 > 0] = [-1.5 > 0] = 0$$
$$[-1.5 + 0 + 1 > 0] = [-0.5 > 0] = 0$$
$$[-1.5 + 1 + 0 > 0] = [-0.5 > 0] = 0$$

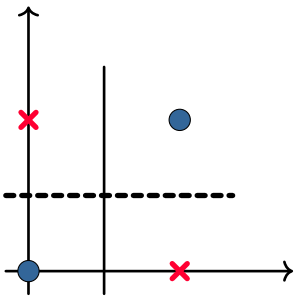
С объединением та же ситуация, только на этот раз линия должна пройти чуть ниже. Подойдёт $x_2 = 0.5 - x_1$.





С третьей операцией, исключающим или, начинаются проблемы. Чтобы разделить точки, нужно строить две линии. Сделать это можно многими способами. Но линий всегда будет две. То есть мы попадаем в ситуацию из прошлой задачки. Надо посмотреть первым слоем нейросетки, где мы оказались относительно каждой из линий, а вторым слоем соединить результаты.





Если немного пофантазировать, можно даже записать эту нейросеть через объединение и пересечение:

$$\hat{y} = [1 \cdot (x_1 \cup x_2) - 1 \cdot (x_1 \cap x_2) - 0.5 > 0]$$

Нейрон $(x_1 \cup x_2)$ выясняет по какую сторону от сплошной линии мы оказались, нейрон $x_1 \cap x_2$ делает то же самое для пунктирной линии. А дальше мы просто объединяем результат.

Упражнение 6 (ещё немного про XoR)

Маша заметила, что на XoR ушло очень много персептронов. Она поняла, что первые два персептрона пытаются сварить для третьего нелинейные признаки, которых нейросетке не хватает. Она решила самостоятельно добавить персептрону вход $x_3 = x_1 \cdot x_2$ и реализовать XoR одним персептроном. Можно ли это сделать?

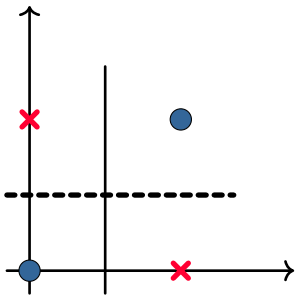
Решение:

Маша обратила внимание на очень важную штуку. Нам не хватает признаков, чтобы реализовать XoR за один нейрон. Поэтому первый слой нейросетки сам их для нас придумывает. Чем глубже нейросетку мы построим, тем более сложные и абстрактные признаки она будет выделять из данных и подавать дальше.

Если добавить ко входу $x_3 = x_1 \cdot x_2$, мы сделаем за нейросетку часть её работы и сможем обойтись одним нейроном. Например, вот таким:

$$\hat{y} = [x_1 + x_2 - 2 \cdot x_1 \cdot x_2 - 0.5 > 0]$$

Такая линия как раз будет задавать две скрещивающиеся прямые.



Это легко увидеть, если немного поколдовать над уравнением:

$$\begin{aligned}
&x_1 + x_2 - 2x_1x_2 - 0.5 = 0 \\
&2x_1 + 2x_2 - 4x_1x_2 - 1 = 0 \\
&2x_1(1 - 2x_2) + 2x_2 - 1 = 0 \\
&(1 - 2x_2) \cdot (2x_1 - 1) = 0
\end{aligned}$$

Получаем два решения. Прямую $x_2 = 0.5$ и прямую $x_1 = 0.5$.

Упражнение 7 (универсальный классификатор)

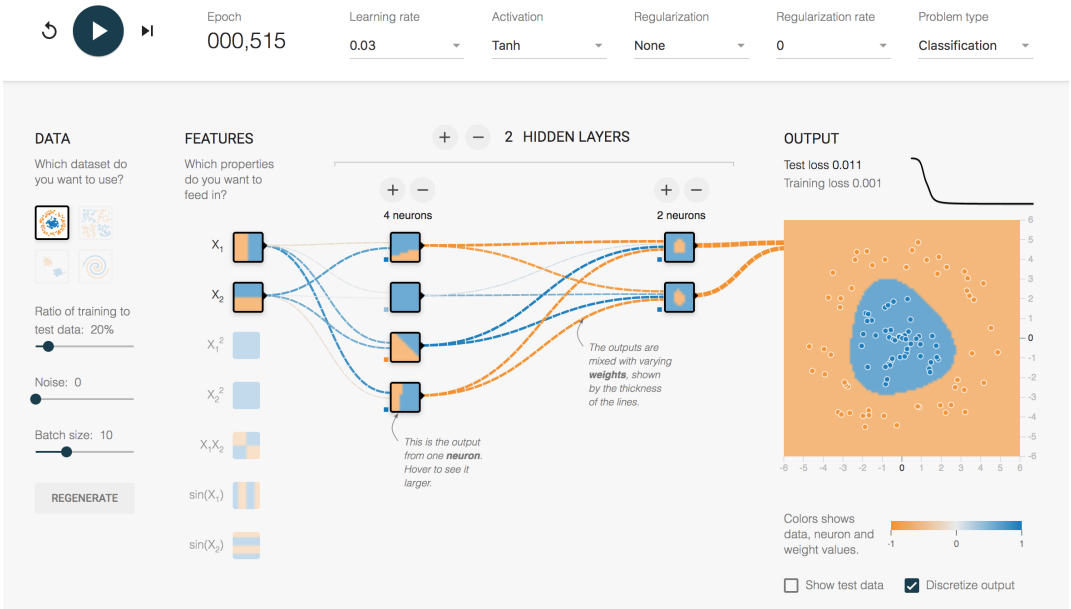
Маша задумалась о том, можно ли с помощью нейронной сетки с одним скрытым слоем и ступенчатой функцией активации решить абсолютно любую задачу классификации на два класса со сколь угодно большой точностью. Ей кажется, что да. Как это можно сделать?

Это бред, перепридумать!

Упражнение 8 (избыток)

На сайте `http://playground.tensorflow.org` Маша стала играться с простенькими нейросетками и обучила для решения задачи классификации трёхслойного монстра.

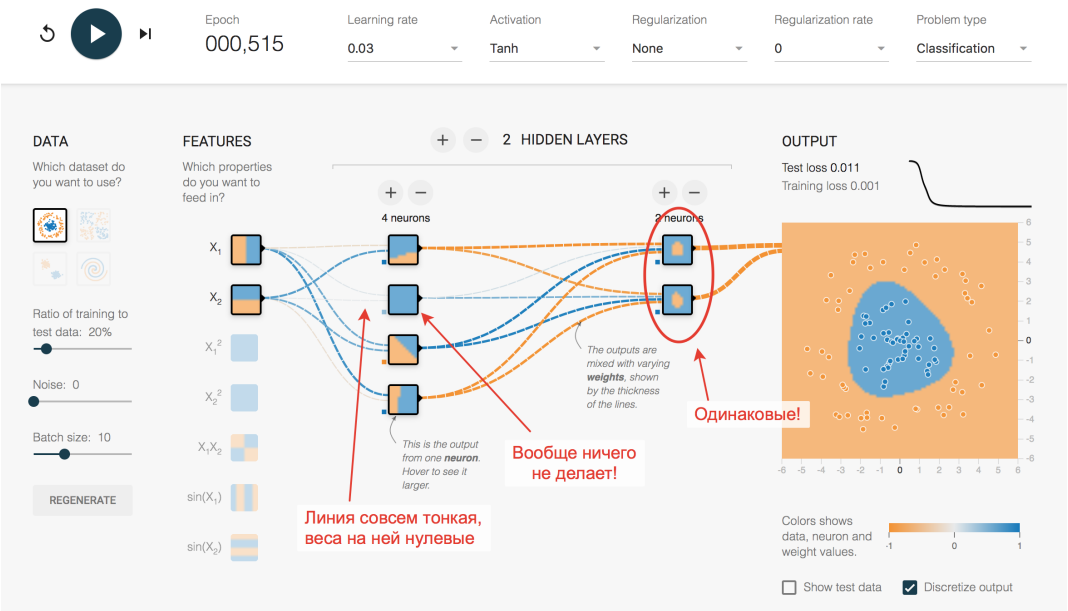
Голубым цветом обозначен первый класс, рыжим второй. Внутри каждого нейрона визуализирована та разделяющая поверхность, которую он выстраивает. Так, первый слой ищет разделяющую линию. Второй слой пытается из этих линий выстроить более сложные фигуры и так далее. Чем ярче связь между нейронами, тем больше весовой коэффициент, относящейся к ней. Синие связи — положительные, рыжие — отрицательные. Чем тусклее связь, тем он ближе к нулю.



Маша заметила, что с её архитектурой что-то не так. Какие у неё проблемы?

Решение:

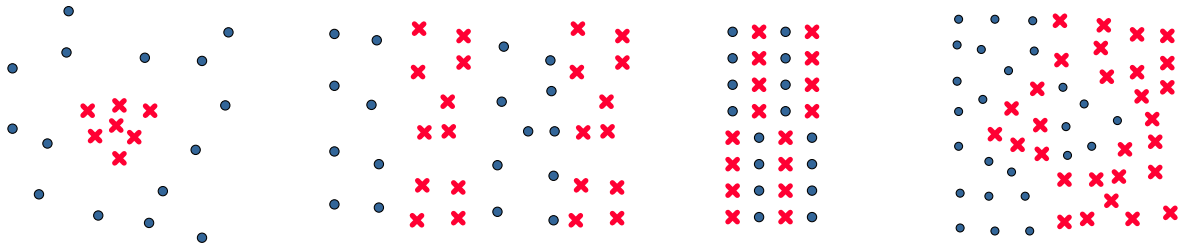
Нейросетка Маши оказалась избыточной. Во-первых, можно увидеть, что на первом слое есть нейрон, который вообще ничего не делает. Связи, которые идут к нему от входов настолько тусклые (коэффициенты при них равны нулю), что их даже не видно на картинке. От этого нейрона смело можно избавиться и сделать архитектуру проще. Во-вторых, можно заметить, что на последнем слое у нас есть два одинаковых нейрона. Один из них смело можно выбрасывать.



Для решения такой простой задачи классификации подойдёт более простая модель. Сколько минимально нужно нейронов, чтобы её решить вам и Маше предстоит выяснить в следующей задаче.

Упражнение 9 (минималочка)

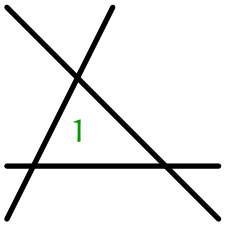
Шестилетняя сестрёнка ворвалась в квартиру Маши и разрисовала ей все обои:



Маша по жизни оптимистка. Поэтому она увидела не дополнительные траты на ремонт, а четыре задачи классификации. И теперь в её голове вопрос: сколько минимально нейронов нужно, чтобы эти задачи решить?

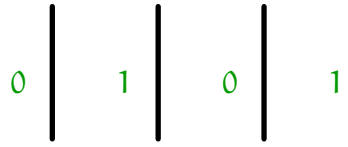
Решение:

a. Нам с помощью нейросетки надо выделить треугольник. Всё, что внутри будет относится к первому классу.



Получается на первом слое надо три нейрона. Каждый из них настроим так, что если мы попадаем внутрь треугольника, он выдаёт 1. Тогда на втором слое будет достаточно одного нейрона, который удостоверится, что все три результата с первого слоя оказались равны 1. Посмотрите теперь на предыдущую задачу, сходите на сайт с демкой и постройте оптимальную нейросетку.

б. Понятно, что первый слой должен построить нам три линии. Это три нейрона.



Второй слой должен принять решение: в какой из полос мы оказались.

в. Перед нами две XoR задачи. На первом слое будем строить четыре линии.

г. Первый слой строит пять линий.

Упражнение 10 (универсальный регрессор)

Маша доказала Паше, что у неё всё в полном порядке с логикой. Теперь она собирается доказать ему, что с помощью однослойной нейронной сетки можно приблизить любую непрерывную функцию от одного аргумента $f(x)$ со сколь угодно большой точностью[10].

Hint: Вспомните, что любую непрерывную функцию можно приблизить с помощью кусочно-линейной функции (ступеньки). Осознайте как с помощью пары нейронов можно описать такую ступеньку. Соедините все ступеньки в сумму с помощью выходного нейрона.

Решение:

Мы хотим приблизить функцию $f(x)$ с какой-то точностью. Будем делать это с помощью кусочно-линейных ступенек. Чем выше точность, тем больше будем рисовать ступенек:

Картинка с двумя приближениями для функции

Попробуем смоделировать одну ступеньку.

Картинка ступеньки

Если x, для которого мы ищем $f(x)$ попадает в неё, мы будем приближать $f(x)$ этой ступенькой. Ступенька состоит из двух линий. Выходит, что она будет описываться двумя нейронами. Если мы внутри ступеньки, значит $a \leqslant x \geqslant a + h$. Пара нейронов должна сравнить x с a и $a + h$ и на основе этого принять решение.

Картинка двух нейронов

[10] http://neuralnetworksanddeeplearning.com/chap4.html

Можно записать попадание x в ступеньку в виде нейрона также как мы делали это в задачке с таблицами истинности:

$$1 - [x \leqslant a] - [x \geqslant a + h]$$

Если оба условия — неправда, получаем 1. Мы в ступеньке. Если хотя бы одно из них выполнено — мы вылетаем за ступеньку. Оба сразу выполниться они не могут.

Будем так действовать для каждой ступеньки. Мы попадём только в одну из них. Значит внутренний слой выплюнет на нас 1 только из одной ветки. Остаётся только решить

ко ко ко

Упражнение 11 (число параметров)

Та, кому принадлежит машин лёрнинг собирается обучить нейронную сеть для решения задачи регрессии, На вход в ней идёт 12 переменных, в сетке есть 3 скрытых слоя. В пером слое 300 нейронов, во втором 200, в третьем 100.

a) Сколько параметров предстоит оценить Маше? Сколько наблюдений вы бы на её месте использовали?

b) Что Маша должна сделать с внешним слоем, если она собирается решать задачу классификации на два класса и получать на выходе вероятность принадлежности к первому классу?

c) Что делать Маше, если она хочет решать задачу классификации на K классов?

2 50 оттенков градиентного спуска

Производная это просто
Скорость роста, это скорость роста.
Возьми предел $\frac{\Delta y}{\Delta x}$ и получишь.
Чем выше она — тем круче.

Научно-технический рэп

Повторять до сходимости — это как жарить до готовности

Неизвестный студент Вышки

Упражнение 1 (погружаемся)

Маша Нестерова, хозяйка машин лёрнинга[11], собрала два наблюдения: $x_1 = 1, x_2 = 2, y_1 =$

[11] Лёрнинг ей папа подарил

$2, y_2 = 3$ и собирается обучить линейную регрессию $y = \beta \cdot x$. Маша очень хрупкая девушка, и ей не помешает помощь.

а. Получите теоретическую оценку методом наименьших квадратов.

б. Сделайте два шага градиентного спуска. В качестве стартовой точки используйте $\beta_0 = 0$. В качестве скорости обучения возьмите $\eta = 0.1$.

в. Сделайте два шага стохастического градиентного спуска. Пусть в SGD сначала попадает первое наблюдение, затем второе.

г. Если вы добрались до этого пункта, вы поняли градиентный спуск. Маша довольна. Начинаем заниматься тупой технической бессмыслицей. Сделайте два шага Momentum SGD. Возьмите $\alpha = 0.9, \eta = 0.1$

д. Сделайте два шага Momentum SGD с коррекцией Нестерова.

е. Сделайте два шага RMSprop. Возьмите $\alpha = 0.9, \eta = 0.1$

ж. Шоб ещё такого сделать? Придумал! Давайте сделаем два шага Adam. Возьмём $\beta_1 = \beta_2 = 0.9, \eta = 0.1$

Упражнение 2 (логистическая регрессия)

Маша решила, что нет смысла останавливаться на обычной регрессии, когда она знает, что есть ещё и логистическая:

$$z = \beta \cdot x \qquad p = P(y = 1) = \frac{1}{1 + e^{-z}}$$
$$\text{logloss} = -[y \cdot \ln p + (1 - y) \cdot \ln(1 - p)]$$

Запишите формулу, по которой можно пересчитывать веса в ходе градиентного спуска для логистической регрессии.

Оказалось, что $x = -5$, а $y = 1$. Сделайте один шаг градиентного спуска, если $\beta_0 = 1$, а скорость обучения $\gamma = 0.01$.

Решение:

Сначала нам надо найти $\text{logloss}'_\beta$. В принципе в этом и заключается вся сложность задачки. Давайте подставим вместо $\hat{p}$ в logloss сигмоиду.

$$\text{logloss} = -1 \left(y \cdot \ln \left(\frac{1}{1 + e^{-z}} \right) + (1 - y) \cdot \ln \left(1 - \frac{1}{1 + e^{-z}} \right) \right)$$

Теперь подставим вместо z уравнение регрессии:

$$\text{logloss} = -1 \left(y \cdot \ln \left(\frac{1}{1 + e^{-\beta \cdot x}} \right) + (1 - y) \cdot \ln \left(1 - \frac{1}{1 + e^{-\beta \cdot x}} \right) \right)$$

Это и есть наша функция потерь. От неё нам нужно найти производную. Давайте подготовимся.

Делай раз, найдём производную $\mathrm{logloss}$ по $\hat{p}$:

$$\mathrm{logloss}'_{\hat{p}} = -1\left(y \cdot \frac{1}{\hat{p}} - (1-y) \cdot \frac{1}{(1-p)}\right)$$

Делай два, найдём производную $\frac{1}{1+e^{-\beta x}}$ по β:

$$\left(\frac{1}{1+e^{-\beta x}}\right)'_{\beta} = -\frac{1}{(1+e^{-\beta x})^2} \cdot e^{-\beta x} \cdot (-x) = \frac{1}{1+e^{-\beta x}} \cdot \frac{e^{-\beta x}}{1+e^{-\beta x}} \cdot x =$$
$$= \frac{1}{1+e^{-\beta x}} \cdot \left(1 - \frac{1}{1+e^{-\beta x}}\right) \cdot x$$

По-другому это можно записать как $\hat{p} \cdot (1-\hat{p}) \cdot x$.

Делай три, находим полную производную:

$$\mathrm{logloss}'_{\beta} = -1\left(y \cdot \frac{1}{\hat{p}} \cdot \hat{p} \cdot (1-\hat{p})) \cdot x - (1-y) \cdot \frac{1}{(1-\hat{p})} \cdot \hat{p} \cdot (1-\hat{p})) \cdot x\right) =$$
$$= -y \cdot (1-\hat{p}) \cdot x + (1-y) \cdot \hat{p} \cdot x = (-y + y\hat{p} + \hat{p} - y\hat{p}) \cdot x = (\hat{p} - y) \cdot x$$

Найдём значение производной в точке $\beta_0 = 1$ для нашего наблюдения $x = -5, y = 1$:

$$\left(\frac{1}{1+e^{-1\cdot(-5)}} - 1\right) \cdot (-5) \approx 4.96$$

Делаем шаг градиентного спуска:

$$\beta_1 = 1 - 0.01 \cdot 4.96 \approx 0.95$$

3 Алгоритм обратного распространения ошибки (Backpropagation)

Что происходит, когда мы суём пальцы в розетку? Нас бьёт током! Мы делаем ошибку, и она распространяется по нашему телу назад.

Твоя мама

Упражнение 1 (граф вычислений)

Маша вспомнила картину из кофейни Добродума и решила нарисовать у себя дома свою такую же. Она хочет изобразить для функции

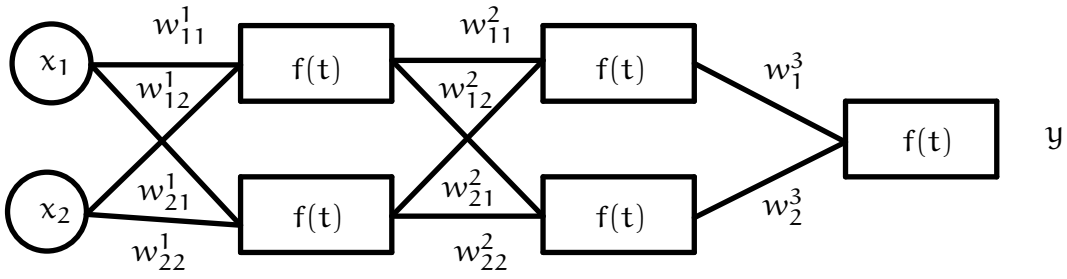
$$f(x, y) = x^2 + xy + (x + y)^2$$

граф вычислений. В кругляшах она будет записывать результаты вычислений. Каждое ребро будет обозначать элементарную операцию: плюс или умножить.

Когда картина будет нарисована, Маша хочет найти производные всех выходов из кругляшей по всем входам. Опираясь на получившийся граф Маша хочет выписать частные производные функции f по x и по y[12].

Упражнение 2 (придумываем backpropagation)

Маша умеет собирать нейросети. Например, у неё есть такая нейросеть:



Здесь w^k_{ij} — веса для k слоя, $f(t)$ — какая-то функция активации. Маша хочет научиться делать для такой нейронной сетки градиентный спуск:

a. Запишите Машину нейросеть как сложную функцию. Сначала в обычной записи, а затем в матричном виде.

[12]По мотивам книги Николенко "Глубокое обучение"(стр. 79)

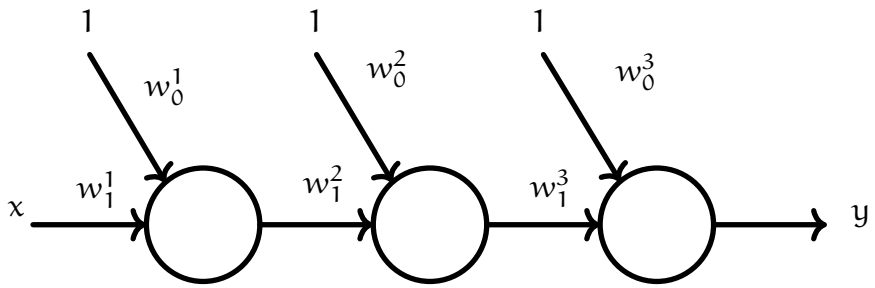
б. Пусть $L(W_1, W_2, W_3) = \frac{1}{2} \cdot (y - \hat{y})^2$ — функция потерь, где W_k — веса k—го слоя. Найдите производные функции L по всем весам W_k.

в. В производных постоянно повторяются одни и те же части. Постоянно искать их не очень оптимально. Выделите эти часть в прямоугольнички цветными ручками.

г. Выпишите все производные в том виде, в котором их было бы удобно использовать для алгоритма обратного распространения ошибки, а затем, сформулируйте сам алгоритм.

Упражнение 3 (backpropagation руками)

Маша как-то раз решала задачу классифкации. С тех пор у неё в кармане завалялась нейросеть:



В качестве функции активации Маша использовала сигмоиду: $f(t) = \frac{e^t}{1+e^t}$. Как это обычно бывает, Маша обнаружила её в своих штанах после стирки и очень обрадовалась. Теперь она хочет сделать два шага стохастического градиентного спуска, используя алгоритм обратного распространения ошибки.

У неё есть два наблюдения: $x_1 = 1, x_2 = 5$, $y_1 = 1$, $y_2 = 0$. Скорость обучения $\gamma = 1$. В качестве инициализации взяты нулевые веса. Сначала берётся второе наблюдение, затем первое. Помогите Маше.

Упражнение 4 (незаметный backpropagation)

Маша собрала нейросеть:

$$y = \max\left(0; X \cdot \begin{pmatrix} 1 & -1 \\ 0.5 & 0 \end{pmatrix}\right) \cdot \begin{pmatrix} 0.5 \\ 1 \end{pmatrix}$$

Теперь Маша внимательно смотрит на неё.

а. Первый слой нашей нейросетки — линейный. По какой формуле делается forward pass? Предположим, что на вход пришло наблюдение $x = (1, 2)$. Сделайте через этот слой forward pass и найдите выход из слоя.

б. Найдите для первого слоя производную выхода по входу. При обратном движении по нейросетке, в первый слой пришёл накопленный градиент $(-1, 0)$. Каким будет новое накопленное значение градиента, которое выплюнет из себя линейный слой? По какой формуле делается backward pass?

в. Второй слой нейросетки —- функция активации, ReLU. По какой формуле делается forward pass? На вход в него поступило значение $(2, -1)$. Сделайте через него forward pass.

г. Найдите для второго слоя производную выхода по входу. При обратном движении по нейросетке во второй слой пришёл накопленный градиент $(-1, -2)$. Каким будет новое накопленное значение градиента, которое выплюнет из себя ReLU? По какой формуле делается backward pass?

д. Третий слой нейросетки — линейный. По какой формуле делается forward pass? Пусть на вход поступило значение $(2, 0)$. Сделайте через него forward pass.

е. Найдите для третьего слоя производную выхода по входу. При обратном движении по нейросетке, в третий слой пришёл накопленный градиент -2. Каким будет новое накопленное значение градиента, которое выплюнет из себя линейный слой? По какой формуле делается backward pass?

ж. Мы решаем задачу Регрессии. В качестве функции ошибки мы используем MSE. Пусть для рассматриваемого наблюдения реальное значение $y = 0$. Найдите значение MSE. Чему равна производная MSE по входу (прогнозу)? Каким будет накопленное значение градиента, которое MSE выплюнет из себя в предыдущий слой нейросетки, если изначально значение градиента инициализировано единицей?

з. Пусть скорость обучения $\gamma = 1$. Сделайте для весов нейросети шаг градиентного спуска.

Посидела Маша, посидела, и поняла, что неправильно она всё делает. В реальности перед ней не задача регрессии, а задача классификации.

а. Маша навинтила поверх второго линейного слоя сигмоиду. Как будет для неё выглядеть forward pass? Сделайте его. Найдите для сигмоиды производную выхода по входу.

б. В качестве функции потерь Маша использует logloss. Как для этой функции потерь выглядит forward pass? Сделайте его. Найдите для logloss производную выхода по входу.

в. Как будет выглядеть backward pass через logloss и сигмоиду? Прделайте его. Как изменится процедура градиентного спуска для остальной части сети?

Упражнение 5 (Ещё один backpropagation)

Может усложнить задачу?

Пусть у нас есть нейронка:

$$y = f(X \cdot W_2) \cdot W_1$$

Как для функции потерь $L(W_1, W_2) = (y - \hat{y})^2$ будет выглядеть алгоритм обратного распространения ошибки, если $f(t) = ReLU(t) = max(0; t)$? Найдите все выходы, все промежуточные производные. Опишите правило, по которому производная будет накапливаться, а также сам шаг градиентного спуска.

Какую-нибудь задачу про Нестерова в бэкпропе по мотивам вопроса с пары про то как правильно обновлять веса

4 Активация и потери

Мудрая цитата про активацию

Автор цитаты

Упражнение 1 (про сигмоиду)

Функция $f(t) = \frac{e^t}{1+e^t}$ называется сигмоидой. Вообще говоря сигмоидой называют любую "sобразную функцию.

а. Что происходит при $t \to +\infty$? А при $t \to -\infty$?

б. Как связаны между собой $f(t)$ и $f(-t)$?

в. Как связаны между собой $f'(t)$ и $f'(-t)$?

г. Как связаны между собой $f(t)$ и $f'(t)$?

д. Найдите $f(0)$, $f'(0)$ и $\ln f(0)$.

е. Найдите обратную функцию $f^{-1}(t)$

ж. Как связаны между собой $\frac{d \ln f(t)}{dt}$ и $f(-t)$?

з. Постройте графики функций $f(t)$ и $f'(t)$.

и. Разложите $h(\beta_1, \beta_2) = \ln f(y_i(\beta_1 + \beta_2 x_i))$ в ряд Тейлора до второго порядка в окрестности точки $\beta_1 = 0$, $\beta_2 = 0$.

к. Выпишите формулы для forward pass и backward pass через слой с сигмоидой.

л. Правда ли, что сигмоида способствует затуханию градиента и параличу нейронной сети? Какое максимальное значение принимает её производная?

Упражнение 2 (про logloss)

У Маши три наблюдения, первое наблюдение — кит, остальные — муравьи. Киты кодируются $y_i = 1$, муравьи — $y_i = 0$. В качестве регрессоров Маша берёт номера наблюдений $x_i = i$. После этого Маша оценивает логистическую регрессию с константой.

а. Выпишите эмпирическую функцию риска, которую минимизирует Маша;

б. При каких оценках коэффициентов логистической регрессии эта функция достигает своего минимума?

Упражнение 3 (про $\mathtt{softmax}$)

Маша чуть внимательнее присмотрелась к своему третьему наблюдению и поняла, что это не кит, а бобёр. Теперь ей нужно решать задачу классификации на три класса. Она решил использовать для этого нейросеть с softmax-слоем на выходе. Предположим, что сетка обучилась и на двух новых наблюдениях, перед самым $\mathtt{softmax}$-слоем она выплюнула $1, 2, 5$ и $2, 5, 1$.

а. Чему равны вероятности получить кита, муравья и бобра для обеих ситуаций?

б. Пусть первым был кит, а вторым бобёр. Чему будет равна $\mathtt{logloss}$-ошибка?

в. Пусть у Маши есть два класса. Она хочет выучить нейросеть. Она может учить нейронку с одним выходом и сигмоидой в качестве функции активации либо нейронку с двумя выходами и $\mathtt{softmax}$ в качестве функции активации. Как выходы этих двух нейронок взаимосвязаны между собой?

Тут сделать пункт про то почему софтмакс называется софтмаксом

Упражнение 4 (про тангенс)

Функция $f(t) = \tanh(t) = \frac{2}{1+e^{-2t}} - 1$ называется гиперболическим тангенсом.

а. Что происходит при $t \to +\infty$? А при $t \to -\infty$?

б. Как связаны между собой $f(t)$ и $f'(t)$?

в. Выпишите формулы для forward pass и backward pass через слой с тангенсом.

г. Правда ли, что тангенс способствует затуханию градиента и параличу нейронной сети? Какое максимальное значение принимает производная тангенса?

д. пункт про то, почему часто функцию юзают в RNN

Упражнение 5 (про ReLU)

Функция $f(t) = \mathrm{ReLU}(t) = \max(t, 0)$ называется ReLU.

а.

Задача про ReLU и сигмоиду (Николенко)

Задача про паралич сигмоиды и ReLU

Упражнение 6 (про разные выходы)

Та, в чьих руках находится лёрнинг, решила немного поэкспериментировать с выходами из своей сетки.

а. Для начала Маша решила, что хочет решать задачу классификации на два класса и получать на выходе вероятность принадлежности к первому. Что ей надо сделать с последним слоем сетки?

б. Теперь Маша хочет решать задачу классификации на K классов. Что ей делать с последним слоем?

в. Новые вводные! Маша хочет спрогнозировать рейтинг фильма на "Кинопоиске". Он измеряется по шкале от 0 до 10 и принимает любое непрерывное значение. Как Маша может приспособить для этого свою нейронку?

г. У Маши есть куча новостей. Каждая новость может быть спортивной, политической или экономической. Иногда новость может относится сразу к нескольким категориям. Как Маше собрать нейросетку для решения этой задачи? Как будет выглядеть при этом функция ошибки?

д. Маша пошла в кафе. А там куча народу. Сейчас она сидит за столиком, попивает ванильный топлёный кортадо и думает о нём, о лёрнинге. Сейчас мысль такая: как можно спрогнозировать число людей в кафе так, чтобы на выходе сетка всегда прогнозировала целое число. Надо ли как-то при этом менять функцию потерь?

Hint: вспомните про пуассоновскую регрессию.

Упражнение 7 (температура генерации)

Иногда в функцию $\mathtt{softmax}$ добавляют дополнительный параметр T, который называют температурой. Тогда она приобретает вид

$$f(z) = \frac{e^{\frac{z_i}{T}}}{\sum_{k=1}^{K} e^{\frac{z_k}{T}}}$$

Обычно это делается, когда с помощью нейросетки нужно сгенерировать какой-нибудь новый объект. Пусть у нас есть три класса. Наша нейросеть выдала на последнем слое числа $1, 2, 5$.

а. Какое итоговое распределение вероятностей мы получим, если $T = 10$?

б. А если $T = 1$?

в. А если $T = 0.1$?

г. Какое распределение получится при $T \to 0$?

д. А при $T \to \infty$?

е. Предположим, что объектов на порядок больше. Например, это реплики, которые Алиса может сказать вам в ответ на какую-то фразу. Понятное дело, что вашей фразе будет релевантно какое-то подмножество ответов. Какое значение температуры сэмплирования T смогут сделать реплики Алисы непредсказуемыми? А какие сделают их однотипными?

5 Регуляризация

Цитата про переобучение

Автор цитаты

Упражнение 1 (Маша и покемоны)

Маша измерила вес трёх покемонов, $y_1 = 6$, $y_2 = 6$, $y_3 = 10$. Она хочет спрогнозировать вес следующего покемона. Модель для веса покемонов у Маши очень простая, $y_i = \beta + \varepsilon_i$, поэтому прогнозирует Маша по формуле $\hat{y}_i = \hat{\beta}$.

Для оценки параметра β Маша использует следующую целевую функцию:

$$\sum (y_i - \hat{\beta})^2 + \lambda \cdot \hat{\beta}^2$$

а) Найдите оптимальное $\hat{\beta}$ при $\lambda = 0$.

б) Найдите оптимальное $\hat{\beta}$ при произвольном λ. Правда ли, что чем больше λ, тем меньше β?

в) Подберите оптимальное λ с помощью кросс-валидации leave one out («выкинь одного»). При такой валидации на первом шаге мы оцениваем модель на всей выборке без первого наблюдения, а на первом тестируем её. На втором шаге мы оцениваем модель на всей выборке без второго наблюдения, а на втором тестируем её. И так далее n раз. Каждое наблюдение является отдельным фолдом.

г) Найдите оптимальное $\hat{\beta}$ при λ_{CV}.

Упражнение 2 (а вот и моя остановочка)

Сделать задачу по связи ранней остановки и регуляризатора. Как в книжке про диплернинг

Упражнение 3 (дропаут)

Маша собирается обучить нейронную сеть для решения задачи регрессии, На вход в неё идёт 12 переменных, в сетке есть 3 скрытых слоя. В пером слое 300 нейронов, во втором 200, в третьем 100.

a) Сколько параметров предстоит оценить Маше? Сколько наблюдений вы бы на её месте использовали?

b) Пусть в каждом слое была отключена половина нейронов. Сколько коэффициентов необходимо оценить?

c) Предположим, что Маша решила после первого слоя добавить в свою сетку Dropout с вероятностью p. Какова вероятность того, что отключится весь слой?

d) Маша добавила Dropout с вероятностью p. после каждого слоя. Какова вероятность того, что один из слоёв отключится и сетка не сможет учиться?

e) Пусть случайная величина N — это число включённых нейронов. Найдите её математическое ожидание и дисперсию. Если Маша хочет проредить сетку на четверть, какое значение p она должна поставить?

f) Пусть случайная величина P — это число параметров в нейросети, которое необходимо оценить. Найдите её математическое ожидание и дисперсию. Почему найденное вами математическое ожидание выглядит очень логично? Что оно вам напоминает? Обратите внимание на то, что смерть одного из параметров легко может привести к смерти другого.

Бэкпроп через дропаут

Бэкпроп через батчнорм, смысл батчнорма

Упражнение 4 (инициализация весов)

а. Маша использует для активации симметричную функцию. Для инициализации весов она хочет использовать распределение

$$w_i \sim U\left[-\frac{1}{\sqrt{n_{in}}}; \frac{1}{\sqrt{n_{in}}}\right].$$

Покажите, что это будет приводить к затуханию дисперсии при переходе от одного слоя к другому.

б. Какими нужно взять параметры равномерного распределения, чтобы дисперсия не затухала?

в. Маша хочет инициализировать веса из нормального распределения. Какими нужно взять параметры, чтобы дисперсия не затухала?

г. Несимметричный случай

Упражнение 5 (ReLU и инициализация весов)

Внутри нейрона в качестве функции активации используется ReLU. На вход идёт 10 признаков. В качестве инициализации для весов используется нормальное распределение, $N(0, 1)$. С какой вероятностью нейрон будет выдавать на выход нулевое наблюдение, если

Предположения на входы? Какое распределение и с какими параметрами надо использовать, чтобы этого не произошло? Сюда же про инициализацию Хе.

задача про инициализацию от Воронцова

родить задачу из статьи dropout vs batchnorm

6 Всего лишь кубики LEGO

Какая-нибудь цитата про LEGO

автор цитаты

6.1 Свёртка

Упражнение 1 (Картинка)

На вход в нейронную сетку идёт изображение размера 28×28. Маша вытягивает эту картинку в вектор. Он состоит из значений каждого пикселя.

Дальше в сетке идёт полносвязный слой из 1000 нейронов. После полносвязный слой, который осуществляет классификацию изображения на 10 классов. Сколько параметров нужно оценить? **Решение:**

У нас $28^2 = 784$ входа. Весов между входным и полносвязным слоями будет

$$(784 + 1) \cdot 1000 = 785000.$$

Единица отвечает за константу для каждого из 1000 нейронов. Между полносвязным и итоговым слоем

$$(1000 + 1) \cdot 10 = 10010.$$

Задача сделать свёртку своими руками

Задача придумать ядро для классификации крестиков и ноликов

Придумать ядро которое подсчитывает число слешей на картинке

Список ядер с вопросами что они делают с картинкой

Упражнение 2 (Свёрточный и полносвязный слои)

Записать свёрточный слой в виде полносвязного. Рисунок + матрица.

Упражнения про падинги и про то какой рецептиф филд

6.2 Рекурентные сетки

простые задачи про RNN и LSTM

простые задачи про RNN и LSTM

какие-нибудь упражнения про w2v

упражнение про разные модные виды ячеек типа резнетов и тп

7 Итоговый тест в стиле Носко

a. Вопрос про батчнормализацию первым слоем вместо нормализации в предобработке.