Тятя! Тятя! Нейросети заменили продавца!

Ульянкин Ппилиф

Аннотация

В этой виньетке собрана коллекция ручных задачек про нейросетки. 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

> Я попала в сети, которые ты метил, я самая счастливая на всей планете.
>
> *Юлианна Караулова*

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 коллекцию учебников по глубокому обучению. В каких-то из них было написано, что ей никогда не придётся реализовывать алгоритм обратного распространения ошибки, а значит и смысла тратить время на его формулировку нет[2]. В каких-то она находила слишком сложную математику, с которой за один вечер точно не разберёшься.[3] В каких-то алгоритм был описан понятно, но оставалось много недосказанностей[4].

Маша решила, что для вечерних разборок нужно что-то более инфантильное. Тогда она решила поскрести по лёрнингу и собрать коллекцию ручных задачек, прорешивая которую, новые адепты Машиного лёрнинга могли бы открывать у себя диплернинговые чакры. Так и появилась эта виньетка.

[1] Ахахах глубже глубокого, ахахах

[2] Франсуа Шолле, Глубокое обучение на Python

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

[4] Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. Погружение в мир нейронных сетей - Санкт-Петербург, 2018.

Содержание

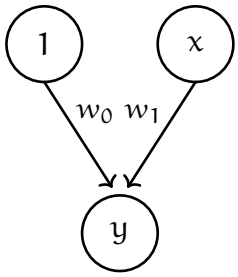
1 Всего лишь функция

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

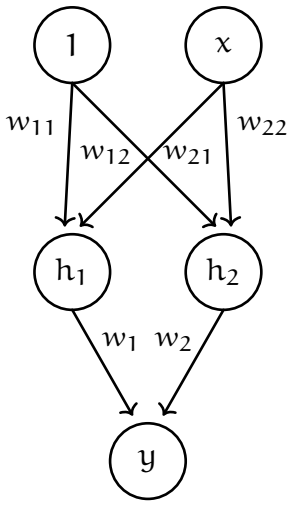
Из фильма «Я, робот» (2004)

Упражнение 1 (от регрессии к нейросетке)

Однажды вечером, по пути с работы[5] 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, 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:

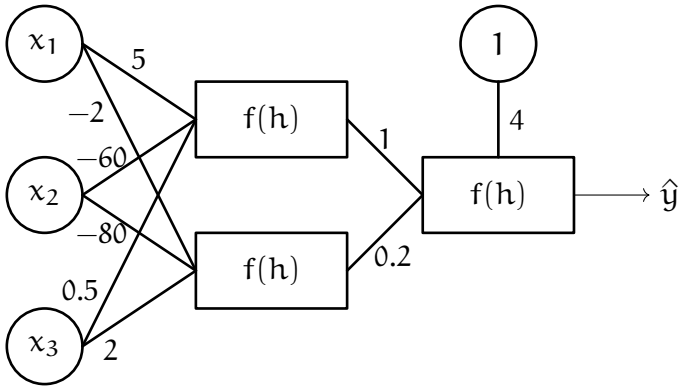


Как такая функция будет выглядеть в виде формулы? Правда ли, что y будет нелинейно зависеть от x? Если нет, как это исправить и сделать зависимость нелинейной?

[5] она работает рисёрчером.

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[6] стойкой стоит Агнесса. Митинга не предвидится, температура будет в районе 20 градусов. Спрогнозируйте, 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.

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. Нарисуйте.

Упражнение 4 (армия регрессий)

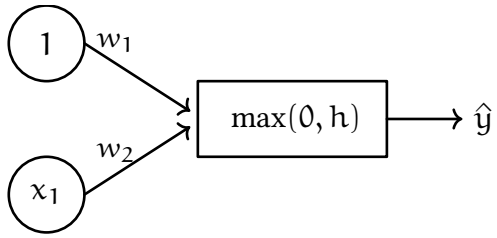
Парни очень любят Машу,[7] а Маша с недавних пор любит собирать персептроны и думать по вечерам об их весах и функциях активации. Сегодня она решила разобрать свои залежи из персептронов и как следует упорядочить их.

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 $y = 0$.

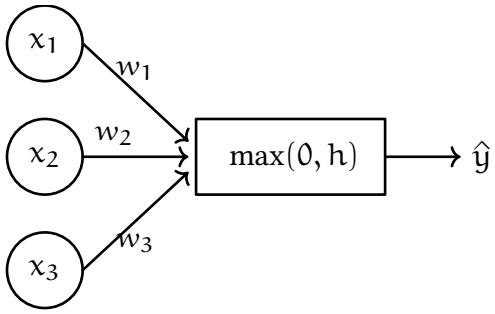
[6] барной... конечно, кофейня у него...

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

Картинка 1



Картинка 2

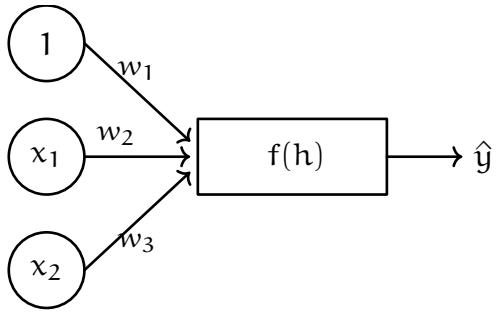


б. В тумбочке, среди носков, Маша нашла персептрон, с картинки 2, Маша хочет подобрать такие веса w_i, чтобы персептрон превращал x из таблички в соответствующие y:

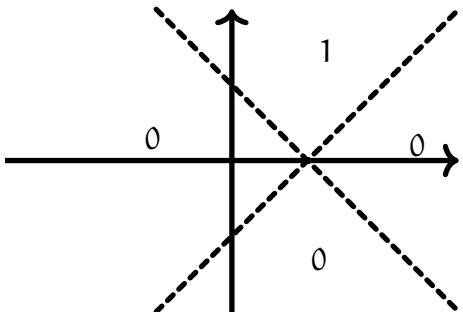
x_1	x_2	x_3	y
1	1	2	0.5
1	−1	1	0

в. Оказывается, что в ванной всё это время валялась куча персептронов с картинки 3 с неизвестной функцией активации.

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. В качестве функции возьмите единичную ступеньку (Функцию Хевисайда).

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 аппроксимацией 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 $[x_1 \neq x_2]$[8].

Упражнение 6 (ещё немного про XoR)

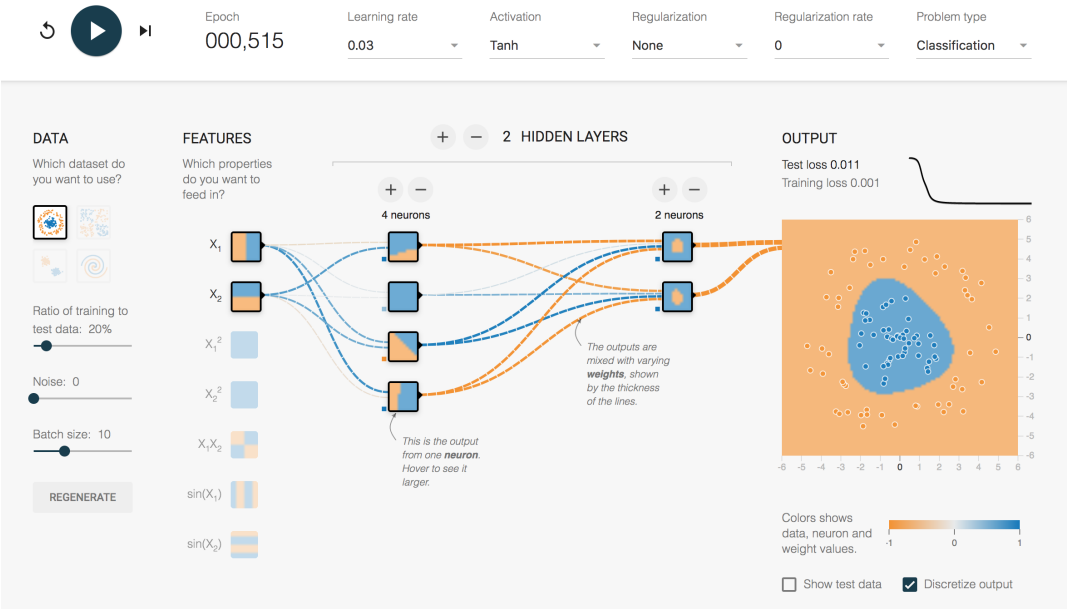
Маша заметила, что на XoR ушло очень много персептронов. Она поняла, что первые два персептрона пытаются сварить для третьего нелинейные признаки, которых нейросетке не хватает. Она решила самостоятельно добавить персептрону вход $x_3 = x_1 \cdot x_2$ и реализовать XoR одним персептроном. Можно ли это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

[8]Тут квадратные скобки обозначают индикатор. Он выдаёт 1, если внутри него стоит правда и 0, если ложь. Такая запись называется скобкой Айверсона. Попробуйте записать через неё единичную ступеньку Хевисайда.

Упражнение 7 (избыток)

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

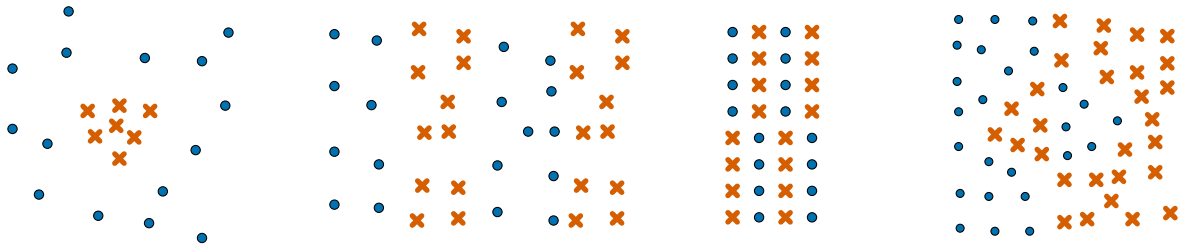
Голубым цветом обозначен первый класс, рыжим второй. Внутри каждого нейрона визуализирована та разделяющая поверхность, которую он выстраивает. Так, первый слой ищет разделяющую линию. Второй слой пытается из этих линий выстроить более сложные фигуры и так далее. Чем ярче связь между нейронами, тем больше 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?

Упражнение 8 (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?

Упражнение 9 (универсальный регрессор)

Маша доказала Паше, что у неё всё в полном порядке с логикой. Теперь она собирается доказать ему, что с помощью двухслойной нейронной сетки можно приблизить любую непрерывную функцию от одного аргумента $f(x)$ со сколь угодно большой точностью[9].

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.

Упражнение 10 (число параметров)

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

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

2 Что выплёвывает нейросеть

Плюют в душу обычно те, кому не удалось в неё влезть.

Пацанский паблик категории Б

Упражнение 1 (про сигмоиду)

Любую s-образную функцию называют сигмоидой. Наиболее сильно прославилась под таким названием функция $f(t) = \frac{e^t}{1+e^t}$. Слава о ней добралась до Маши и теперь она хочет немного поисследовать её свойства.

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. Говорят, что сигмоида — это гладкий аналог единичной ступеньки. Попробуйте построить на компьютере графики $f(t)$, $f(10 \cdot t)$, $f(100 \cdot t)$, $f(1000 \cdot t)$. Как они себя ведут?

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. В качестве функции потерь используются логистические потери.

а. Выпишите для данной задачи функцию потерь, которую минимизирует Маша.

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

Упражнение 3 (про 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.

Маша уже обучила нейронную сетку и хочет построить прогнозы для двух наблюдений. Слой, который находится перед softmax выдал для этих двух наблюдений следующий результат: $1, -2, 0$ и $0.5, -1, 0$.

а. Чему равны вероятности получить кита, муравья и бобра для этих двух наблюдений?

б. Пусть первым был кит, а вторым бобёр. Чему будет равна 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 softmax в качестве функции активации. Как выходы этих двух нейронок взаимосвязаны между собой?

г. Объясните, почему softmax считают сглаженным вариантом максимума.

Упражнение 4 (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. У Маши есть картинки с уточками и чайками. Маша хочет научить нейросеть искать на картинке птицу, обводить её в прямоугольник (bounding box), а затем классифицировать то, что попало в прямоугольник. Как должен выглядеть выход из такой нейросети? Как должна выглядеть функция потерь?

е. Маша задумалась, 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.

3 Пятьдесят оттенков градиентного спуска

Повторять до сходимости — это как жарить до готовности

Неизвестный студент Вышки

Упражнение 1 (50 оттенков спуска)

Маша Нестерова, хозяйка машин лёрнинга[10], собрала два наблюдения: $x_1 = 1, x_2 = 2, y_1 = 2, y_2 = 3$ и собирается обучить линейную регрессию $y = w \cdot x$. 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 $w_0 = 0$. 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. Сделайте 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:

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

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 $w_0 = 1$, а скорость обучения $\gamma = 0.01$.

[10] Лёрнинг ей папа подарил

4 Алгоритм обратного распространения ошибки

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

Твоя мама

Упражнение 1 (граф вычислений)

Как найти производную a по b в графе вычислений? Находим не посещённый путь из a в b, перемножаем все производные на рёбрах получившегося пути. Добавляем это произведение в сумму. Так делаем для всех путей.

Маша хочет попробовать этот алгоритм на функции

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

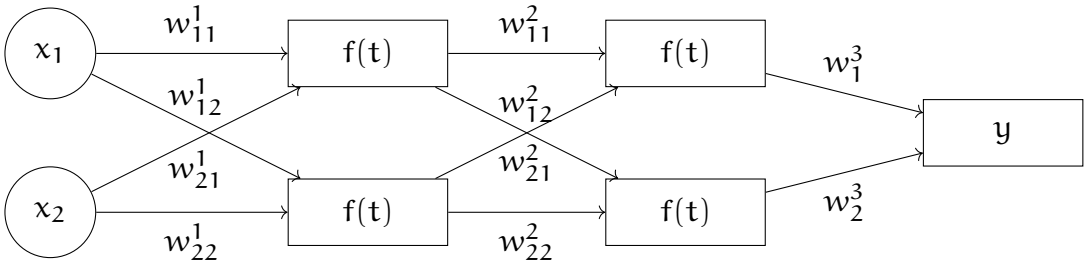
Помогите ей нарисовать граф вычислений и найти $\frac{\partial f}{\partial x}$ и $\frac{\partial f}{\partial y}$. В каждой вершине графа записывайте результат вычисления одной элементарной операции: сложений или умножения.

граф вычислений. В вершинах графа она будет записывать результаты вычислений. Каждое ребро будет обозначать элементарную операцию: плюс или умножить[11].

В тексте ниже огромное число ошибок, я их со временем исправлю!

Упражнение 2 (придумываем backpropagation)

У Маши есть нейросеть с картинки ниже, где w_{ij}^k — веса для k слоя, $f(t)$ — какая-то функция активации. Маша хочет научиться делать для такой нейронной сетки градиентный спуск.



а. Запишите Машину нейросеть, как сложную функцию. Сначала в виде нескольких уравнений, а затем в матричном виде.

б. Предположим, что Маша решает задачу регрессии. Она прогоняет через нейросетку одно наблюдение. На выходе она вычисляет знчение функции потерь $L(W_1, W_2, W_3) = \frac{1}{2} \cdot$

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

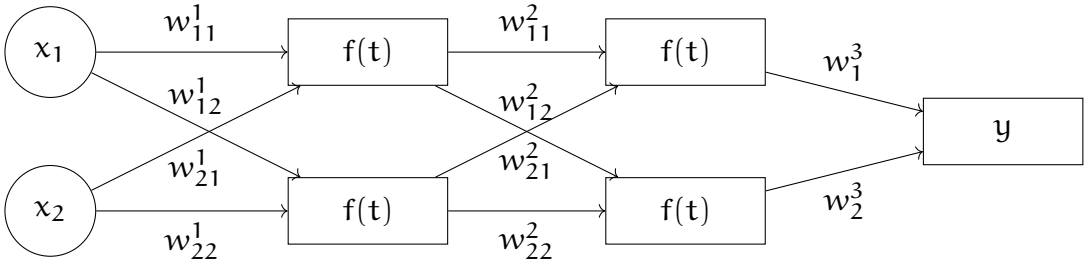
$(y - \hat{y})^2$ — функция потерь, где W_k — веса k—го слоя. Найдите производные функции L по всем весам W_k.

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. Нарисуйте под него удобную схемку.

Упражнение 3 (Backpropagation без константы)

У Маши есть нейросеть с картинки ниже. Она использует функцию потерь $L(W_1, W_2, W_3) = \frac{1}{2} \cdot (y - \hat{y})^2$. В качестве функции активации Маша выбрала сигмоиду $\sigma(t) = \frac{e^t}{1+e^t}$.



Выпишите для Машиной нейросетки алгоритм обратного распространения ошибки в общем виде. Пусть Маша инициализировала веса нейронной сети нулями. У неё есть два наблюдения

№	x_1	x_2	y
1	1	1	1
2	5	2	0

Сделайте руками два шага алгоритма обратного распространения ошибки. Пусть скорость обучения $\eta = 1$. Стохастический градиентный спуск решил, что сначала для шага будет использоваться второе наблюдение, а затем первое.

5 Всего лишь кубики LEGO

5.1 Функции активации

Желание - Ржавый - Семнадцать - Рассвет - Печь - Девять - Добросердечный - Возвращение на Родину - Один - Грузовой вагон.

Код активации Зимнего Солдата

Упражнение 1 (про сигмоиду)

Любую "sобразную функцию называют сигмоидой. Наиболее сильно прославилась под таким названием функция $f(t) = \frac{e^t}{1+e^t}$. Слава о ней добралась до Маши и теперь она хочет немного поисследовать её свойства.

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

б. Какое максимальное значение принимает производная сигмоиды? Объясните как это способствует затуханию градиента и параличу нейронной сети?

Упражнение 2 (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 3 (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

Parametric Rectifier (PReLU) Выписать уравнения для бэкпропа по параметру α

Упражнение 4 (температура генерации)

Иногда в функцию 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?

5.2 Регуляризация

Цитата про переобучение

Автор цитаты

Упражнение 5 (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 6 (а вот и моя остановочка)

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

Упражнение 7 (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.

Добавить вопросиков про дропконнект

Бэкпроп через дропаут

5.3 Нормализация по батчам

Чашка хорошего чая восстановит мою нормальность.

Артур из «Автостопом по галактике»

Бэкпроп через батчнорм, смысл батчнорма

родить задачу из статьи dropout vs batchnorm

5.4 Инициализация

цитата об этом

автор

Упражнение 8 (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 9 (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 Xe.

задача про инициализацию от Воронцова

5.5 Стрельба по ногам

Упражнение 10 (Проблемы с архитектурой)

Миша принёс Маше несколько разных архитектур. Они выглядят довольно странно. Помогите Маше разобраться, что именно Миша сделал неправильно.

а. Решается задача регрессии, предсказываются цены на недвижимость.

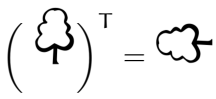
б. Решается задача классификация картинок на 10 классов. Исходный размер картинок 28×28.

в. Решается задача классификация картинок на 10 классов. Исходный размер картинок 100×100.

6 Свёрточные сетки

7 Рекуррентные сетки

8 Матричное дифФффФффФфференцирование



«Джек и бобовый стебель» (1890)

Эта часть виньетки необязательна для изучения. В ней мы подробно поговорим про матричные производные. С их помощью удобно заниматься оптимизацией, в том числе бэкпропом[12].

Упражнение 1

Найдите следующие производные:

а. $f(x) = x^2$, где x скаляр

б. $f(x) = a^T x$, где a и x векторы размера $1 \times n$

в. $f(x) = x^T A x$, где x вектор размера $1 \times n$, A матрица размера $n \times n$

г. $f(x) = \ln(x^T A x)$, где x вектор размера $1 \times n$, A матрица размера $n \times n$

д. $f(x) = a^T X A X a$, где x вектор размера $1 \times n$, A матрица размера $n \times n$

е. $f(x) = x x^T x$, где x вектор размера $1 \times n$

Упражнение 2

Давайте пополним таблицу дифференциалов несколькими новыми функциями, специфичными для матриц. Найдём матричные дифференциалы функций:

а. $f(X) = X^{-1}$, где матрица X размера $n \times n$

б. $f(X) = \det X$, где матрица X размера $n \times n$

в. $f(X) = \mathrm{tr}(X)$, где матрица X размера $n \times n$

г. Ещё больше матричных производных можно найти в книге The Matrix Cookbook[13]

Упражнение 3

Найдите следующие производные:

а. $f(X) = \mathrm{tr}(AXB)$, где матрица A размера $p \times m$, матрица B размера $n \times p$, матрица X размера $m \times n$.

б. $f(X) = \mathrm{tr}(AX^T X)$, где матрица A размера $n \times n$, матрица X размера $m \times n$.

в. $f(X) = \ln \det X$

г. $f(X) = \mathrm{tr}(AX^T X B X^{-T})$

д. $f(X) = \det(X^T A X)$

[12] Часть задач взята из прототипа задачника по ML Бориса Демешева, часть из конспектов по ML Жени Соколова

[13] https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

е. $f(x) = x^T Ab$, где матрица A размера $n \times n$, вектора x и b размера $n \times 1$.

ж. $f(A) = x^T Ab$.

Упражнение 4

Рассмотрим задачу линейной регресии

$$L(w) = (y - Xw)^T (y - Xw) \to \min_w.$$

а. Найдите $L(w)$, выведите формулу для оптимального w.

б. Как выглядит шаг градиентного спуска в матричном виде?

в. Найдите $d^2L(w)$. Убедитесь, что мы действительно в точке минимума.

Упражнение 5

В случае Ridge-регрессии минимизируется функция со штрафом:

$$L(w) = (y - xw)^T (y - xw) + \lambda w^2,$$

где λ — положительный параметр, штрафующий функцию за слишком большие значения w.

а. Найдите $dL(w)$, выведите формулу для оптимального w.

б. Как выглядит шаг градиентного спуска в матричном виде?

в. Найдите $d^2L(w)$. Убедитесь, что мы действительно в точке минимума.

Упражнение 6

Пусть x_i — вектор-столбец $k \times 1$, y_i — скаляр, равный $+1$ или -1, w — вектор-столбец размера $k \times 1$. Рассмотрим логистическую функцию потерь с l_2 регуляризацией

$$L(w) = \sum_{i=1}^{n} \ln(1 + \exp(-y_i x_i^T w)) + \lambda w^T w$$

а. Найдите dL;

б. Найдите вектор-столбец ∇L.

в. Как для этой функции потерь выглядит шаг градиентного спуска в матричном виде?

Упражнение 7

Упражняемся в матричном методе максимального правдоподобия. Допустим, что выборка размера n пришла к нам из многомерного нормального распределения с неизвестными вектором средних μ и ковариационной матрицей Σ. В этом задании нужно найти оценки максимального

правдоподобия для $\hat{\mu}$ и $\hat{\Sigma}$. Обратите внимание, что выборкой здесь будет не $x_1, \dots, x_n$, а

$$\begin{pmatrix} x_{11}, \dots, x_{n1} \\ \dots \\ x_{n1}, \dots, x_{nm} \end{pmatrix}$$

Упражнение 8

Найдите симметричную матрицу X наиболее близкую к матрице A по норме Фробениуса, $\sum_{i,j}(x_{ij} - a_{ij})^2$. Тут мы просто из каждого элемента вычитаем каждый и смотрим на сумму квадратов таких разностей. То есть решите задачку условной матричной минимизации

$$\begin{cases} \|X - A\|^2 \to \min_A \\ X^T = X \end{cases}$$

Hint: Надо будет выписать Лагранджиан. А ещё пригодится тот факт, что $\sum_{i,j}(x_{ij} - a_{ij})^2 = \|X - A\|^2 = \mathrm{tr}((X - A)^T(X - A))$.