Тятя! Тятя! Наши сети притащили мертвеца!

эконом РАНХиГС
осень 2019

Задачи к посиделке 3

Обратное распространение ошибки

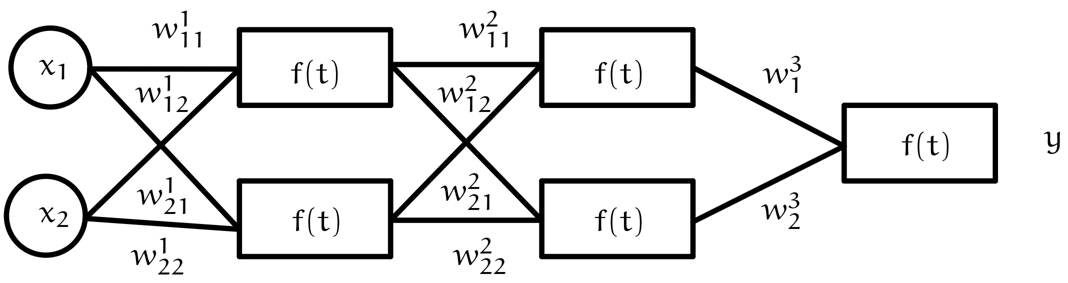
Что происходит, когда мы суём пальцы в розетку? Нас бьёт током! Мы делаем ошибку, и она распространяется по нашему телу назад.

Задача 1

Изобразите для функции $f(x, y) = x^2 + xy + (x + y)^2$ граф вычислений. Найдите производные всех выходов по всем входам. Опираясь на граф выпишите частные производные функции f.[1]

Задача 2

Дана нейросетка:



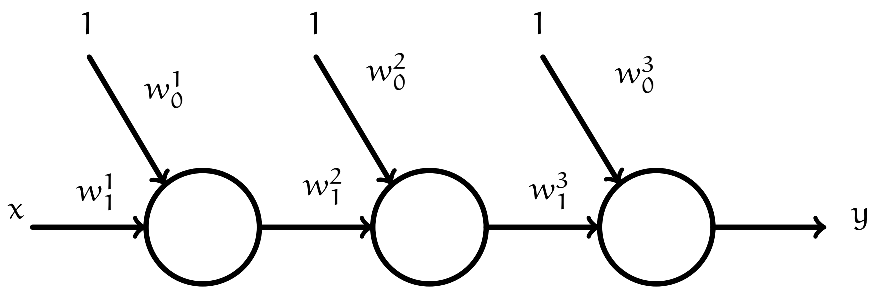
1. Перепишите её как сложную функцию.
2. Запишите эту функцию в матричном виде.
3. Предположим, что $L(W_1, W_2, W_3) = \frac{1}{2} \cdot (y - \hat{y})^2$ — функция потерь, где W_i — веса i—го слоя. Найдите производную функции L по всем весам W_i.

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

4. Выглядит не очень оптимально, правда? Выпишите все производные в том виде, в котором их было бы удобно использовать для алгоритма обратного распространения ошибки, а затем, сформулируйте сам алгоритм.

Задача 3

Как-то раз Вовочка решал задачу классификации. С тех пор у него в кармане завалялась нейросеть:



В качестве функции активации используется сигмоид: $f(t) = \frac{e^t}{1+e^t}$. 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. Как это обычно бывает, Вовочка обнаружил её в своих штанах после стирки и очень обрадовался. Теперь он собирается сделать два шага стохастического градиентного спуска, используя алгоритм обратного распространения ошибки. Помогите ему.

Задача 4

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 $f(t) = \mathrm{ReLU}(t) = max(0; t)$? Найдите все выходы, все промежуточные производные. Опишите правило, по которому производная будет накапливаться, а также сам шаг градиентного спуска.

Задача 5

Маша (ОПЯТЬ ОНА?!) 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.

1. 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.
2. 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?
3. 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.
4. 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?
5. Третий слой нейросетки — линейный. По какой формуле делается forward pass? Пусть на вход поступило значение $(2, 0)$. Сделайте через него forward pass.
6. 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?
7. 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 инициализированно единицей?
8. 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.

1. Маша навинтила поверх второго линейного слоя сигмоиду. Как будет для неё выглядеть forward pass? Сделайте его. Найдите для сигмоиды производную выхода по входу.
2. В качестве функции потерь Маша использует logloss. Как для этой функции потерь выглядит forward pass? Сделайте его. Найдите для logloss производную выхода по входу.
3. Как будет выглядеть backward pass через logloss и сигмоиду? Прделайте его. Как изменится процедура градиентного спуска для остальной части сети?

Матричное дифФфФфФфириенцирование[2]

Задача 6

В этой задачке нужно просто найти НЕмного разных производных:

1. $f(x) = a^T x$, где a и x векторы размера $1 \times n$
2. $f(x) = x^T A x$, где x вектор размера $1 \times n$, A матрица размера $n \times n$
3. $f(x) = \ln(x^T A x)$, где x вектор размера $1 \times n$, A матрица размера $n \times n$

[2]Часть задач взята из прототипа задачника по ML Демешева, часть из Конспектов Соколова

4. $f(x) = a^T X A X a$, где x вектор размера $1 \times n$, A матрица размера $n \times n$
5. $f(x) = x x^T x$, где x вектор размера $1 \times n$
6. $f(X) = X^{-1}$, где матрица X размера $n \times n$
7. $f(X) = \det X$, где матрица X размера $n \times n$

Задача 7

В этой задачке нужно просто найти много разных производных:

1. $f(X) = \mathrm{tr}(AXB)$, где матрица A размера $p \times m$, матрица B размера $n \times p$, матрица X размера $m \times n$.
2. $f(X) = \mathrm{tr}(AX^T X)$, где матрица A размера $n \times n$, матрица X размера $m \times n$.
3. $f(X) = \ln \det X$
4. $f(X) = \ln AX^{-1}B$
5. $f(X) = \mathrm{tr}(AX^T XBX^{-T})$
6. $f(X) = \ln \det(X^T AX)$
7. $f(x) = x^T Ab$, где матрица A размера $n \times n$, вектора x и b размера $n \times 1$.
8. $f(A) = x^T Ab$.

Задача 8

Рассмотрим задачу линейной регрессии

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

1. Найдите $dQ(w)$, выведите формулу для оптимального w.
2. Как выглядит шаг градиентного спуска в матричном виде?
3. Найдите $d^2Q(w)$. Убедитесь, что мы действительно в точке минимума.

Задача 9

В случае Ridge-регрессии минимизируется функция

$$Q(w) = (y - Xw)^T (y - Xw) + \lambda w^T w,$$

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

1. Найдите $dQ(w)$, выведите формулу для оптимального w.
2. Как выглядит шаг градиентного спуска в матричном виде?
3. Найдите $d^2Q(w)$. Убедитесь, что мы действительно в точке минимума.

В случае Lasso-регрессии мы имеем дело с функцией

$$Q(w) = (y - Xw)^T (y - Xw) + \lambda |w|,$$

1. Найдите $dQ(w)$, выведите формулу для оптимального w.

2. Как выглядит шаг градиентного спуска в матричном виде?

Задача 10

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

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

1. Найдите dQ;
2. Найдите вектор-столбец ∇Q.

Задача 11

Упражняемся в матричном методе максимального правдоподобия! Допустим, что векторы $X_1, \dots, X_m$ выбраны 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 $X_1, \dots, X_m$, а

$$\begin{pmatrix} x_{11}, \dots, x_{m1} \\ \dots \\ x_{1n}, \dots, x_{mn} \end{pmatrix}$$

Задача 12

Найдите симметричную матрицу X наиболее близкую а 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 $\sum_{i,j}(x_{ij} - a_{ij})^2 = ||X - A||^2 = tr((X - A)^T(X - A))$. То, что это так мы доказали на семинаре :) Вспоминайте!