Математика для Data Science. Математический анализ. Решения задач

Содержание

Замечание. Вот этим цветом отмечены ссылки на страницы внутри этого файла.

5.2 Линейная регрессия и градиентный спуск

Задача 1

Пример с арбузами

Наша обучающая выборка состоит из 4 арбузов. Каждый арбуз имеет 3 признака: диаметр, вес и количество полос. Все три признака являются действительными числами (сантиметры, граммы, штуки). Мы хотим измерять вкусность арбузов. Но вкусность мерять сложно. Поэтому вместо вкусности мы будем мерять количество грамм сахара, содержащегося в одном килограмме арбуза. Это и будет нашей целевой функцией. Будем называть её "сахарностью".

Вот досье наших арбузов:

1. Арбуз Изабелла $(16, 3500, 20)$, сахарность 10
2. Арбуз Авдотья $(14, 3800, 23)$, сахарность 12
3. Арбуз Дракo $(17, 3100, 18)$, сахарность 13
4. Арбуз Тухачевский $(20, 2500, 30)$, сахарность 170

Мы откуда-то взяли модель с параметрами $(\theta_0, \theta_1, \theta_2, \theta_3) = (20, 1, \frac{1}{1000}, -2)$. Она не обучена и ответы будет давать плохие, но для разбора обозначений подойдёт.

Напишите в поле ответа, чему равны следующие величины. И обсудите ваш ответ с преподавателем.

1. $x_3^{(1)}$
2. $x_1^{(3)}$
3. $x_0^{(2)}$
4. m
5. n
6. $y^{(2)}$
7. $\hat{y}^{(2)}$
8. $L(y^{(2)}, \hat{y}^{(2)})$
9. $h_\theta(x^{(2)})$
10. $(h_\theta(x^{(2)}) - y^{(2)})^2$

Подсказка. Вспомните обозначения с двух предыдущих шагов. В решении мы будем писать, на какой пункт в обозначениях мы ссылаемся.

Решение.

1. $x_3^{(1)}$ — это значение третьего признака первого объекта нашей выборки (см. пункт 2 обозначений). То есть, это количество полос у арбуза Изабелла. Итого, $x_3^{(1)} = 20$.
2. $x_1^{(3)}$ — это значение первого признака у третьего объекта нашей выборки (см. п. 2). То есть, это диаметр арбуза Драко. Значит, $x_1^{(3)} = 17$.
3. $x_0^{(2)}$ — это нулевой (фиктивный) признак второго объекта (см. п. 9), который всегда равен 1. Тогда и $x_0^{(2)} = 1$.
4. m — это количество объектов в выборке (см. п. 4). В нашем случае арбузов 4, то есть $m = 4$.
5. n — это количество признаков у объектов нашей выборки (см. п. 2). У нас $n = 3$.
6. $y^{(2)}$ — это значение целевой функции на втором объекте. А сахарность второго арбуза (Авдотья) равна $y^{(2)} = 12$.
7. $\hat{y}^{(2)}$ — это предсказание нашей модели для второго объекта (см. п. 9). При этом

$$\hat{y}^{(2)} = \sum_{j=0}^{3} \theta_j x_j^{(2)} = \theta_0 x_0^{(2)} + \theta_1 x_1^{(2)} + \theta_2 x_2^{(2)} + \theta_3 x_3^{(2)}.$$

   Подставим сюда $(\theta_0, \theta_1, \theta_2, \theta_3) = (20, 1, \frac{1}{1000}, -2)$, тогда, продолжая цепочку равенств, получаем, что $\hat{y}^{(2)} = 20 \cdot x_0^{(2)} + 1 \cdot x_1^{(2)} + \frac{1}{1000} \cdot x_2^{(2)} - 2 \cdot x_3^{(2)}$. Теперь подставим параметры арбуза Авдотья: $x^{(2)} = (x_1^{(2)}, x_2^{(2)}, x_3^{(2)}) = (14, 3800, 23)$ и $x_0^{(2)} = 1$, получим $\hat{y}^{(2)} = 20 \cdot 1 + 1 \cdot 14 + \frac{1}{1000} \cdot 3800 - 2 \cdot 23 = -8.2$.
8. $L(y^{(2)}, \hat{y}^{(2)})$ — это значение квадратичной функции потерь на втором объекте (см. п. 6). $L(y^{(2)}, \hat{y}^{(2)}) := (\hat{y}^{(2)} - y^{(2)})^2$. Подставим в это выражение значения $y^{(2)}$ и $\hat{y}^{(2)}$, полученные в двух предыдущих пунктах этой задачи. Получаем: $L(y^{(2)}, \hat{y}^{(2)}) = (-8.2 - 12)^2 = 408.04$.
9. $h_\theta(x^{(2)}) = \hat{y}^{(2)}$ (см. п. 9). А это значение мы уже считали в 7 пункте этой задачи: $\hat{y^{(2)}} = -8.2$
10. $(h_\theta(x^{(2)}) - y^{(2)})^2 = (\hat{y}^{(2)} - y^{(2)})^2$. Это мы тоже уже посчитали в 8 пункте этой задачи: $(\hat{y}^{(2)} - y^{(2)})^2 = 408.04$.

Задача 2

В шагах про обозначения в пункте 7 был такой параграф:

> Значением функции потерь для всей выборки называют среднее значение функции потерь по всей выборке, то есть $\frac{1}{m}\sum\limits_{i=1}^{m} L(y^{(i)}, \hat{y}^{(i)}) = \frac{1}{m}\sum\limits_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2$. По некоторым причинам коэффициент $\frac{1}{m}$ перед суммой не важен (обсудим это позже), поэтому обычно за функцию потерь берут $\frac{1}{2}\sum\limits_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2$.

Почему нам не важен коэффициент $\frac{1}{m}$ перед суммой, и мы можем заменить его на любое другое ненулевое число?

Подсказка. Вспомните, какова наша финальная цель обучения и для чего мы вводили функцию потерь.

Решение. Как мы помним, наша цель — найти точку минимума функции потерь. Если мы найдём $\hat{y}^{(1)}, \ldots, \hat{y}^{(m)}$, при которых достигается локальный минимум функции $\frac{1}{m}\sum\limits_{i=1}^{m} L(y^{(i)}, \hat{y}^{(i)}) = \frac{1}{m}\sum\limits_{i=1}^{m}(\hat{y}^{(i)} - y^{(i)})^2$, то при этих же значениях $\hat{y}^{(1)}, \ldots, \hat{y}^{(m)}$ будет достигаться локальный минимум функции $c \cdot \sum\limits_{i=1}^{m} L(y^{(i)}, \hat{y}^{(i)})$, где $c \in \mathbb{R}$ — ненулевое число. В частности, можно взять $c = \frac{1}{2}$.

Задача 3

Давайте для простоты рассмотрим случай $m = 1$. То есть когда наша обучающая выборка состоит из одного объекта.

Тогда $J(\theta) := \frac{1}{2}\sum\limits_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 = \frac{1}{2}\left(h_\theta(x^{(1)}) - y^{(1)}\right)^2$

Воспользуемся формулой для h_θ с десятого шага этого урока: $\frac{1}{2}\left(h_\theta(x^{(1)}) - y^{(1)}\right)^2 = \frac{1}{2}\left(\sum\limits_{j=0}^{n}\theta_j x_j^{(1)} - y^{(1)}\right)^2$.

То есть $J(\theta) = \frac{1}{2}\left(\sum\limits_{j=0}^{n}\theta_j x_j^{(1)} - y^{(1)}\right)^2$.

Найдите $\nabla J(\theta) := \left(\frac{\partial J}{\partial \theta_0}(\theta), \frac{\partial J}{\partial \theta_1}(\theta), \ldots, \frac{\partial J}{\partial \theta_n}(\theta)\right)$ в этом случае (т.е. найдите все частные производные J в точке θ).

Возможно, вам понадобятся правила для вычисления производных.

Кстати, теперь должно быть понятно, почему мы взяли коэффициент $\frac{1}{2}$ вместо $\frac{1}{m}$.

Подсказка. Посчитайте, чему равна частная производная $\frac{\partial J}{\partial \theta_0}(\theta)$. Для этого представьте, что все остальные параметры $\theta_1, \theta_2, \ldots, \theta_n$ — постоянные величины. Тогда на функцию $J(\theta)$ можно смотреть как на функцию от одной переменной θ_0, от которой можно посчитать обычную производную. Эта производная и будет равна искомой $\frac{\partial J}{\partial \theta_0}(\theta)$. А $\frac{\partial J}{\partial \theta_i}(\theta)$ для $i = 1, 2, \ldots n$ будут считаться аналогично.

Решение. Как и написано в подсказке, посчитаем $\frac{\partial J}{\partial \theta_0}(\theta)$.

По условию $J(\theta) = \frac{1}{2}\left(\sum\limits_{j=0}^{n}\theta_j x_j^{(1)} - y^{(1)}\right)^2$. Зафиксируем значения $\theta_1, \theta_2, \ldots, \theta_n$, тогда $J(\theta) = J(\theta_0) = \frac{1}{2}\left(\theta_0 x_0^{(1)} + \sum\limits_{j=1}^{n}\theta_j x_j^{(1)} - y^{(1)}\right)^2$. При этом на выражение $\sum\limits_{j=1}^{n}\theta_j x_j^{(1)} - y^{(1)}$ мы смотрим как на некоторую константу. Обозначим её за $c := \sum\limits_{j=1}^{n}\theta_j x_j^{(1)} - y^{(1)}$. Итак, мы хотим посчитать производную от функции одной переменной $J(\theta_0) = \frac{1}{2}(\theta_0 x_0^{(1)} + c)^2$.

Пользуясь правилом взятия производной сложной функции, получаем: $J'(\theta_0) = \frac{1}{2} \cdot 2 \cdot (\theta_0 x_0^{(1)} + c) \cdot x_0^{(1)} = (\theta_0 x_0^{(1)} + c) \cdot x_0^{(1)}$.

Вспоминая, что мы обозначили за c, получаем, что $\frac{\partial J}{\partial \theta_0}(\theta) = \left(\theta_0 x_0^{(1)} + \sum\limits_{j=1}^{n} \theta_j x_j^{(1)} - y^{(1)}\right) \cdot x_0^{(1)} =$

$= \left(\sum\limits_{j=0}^{n} \theta_j x_j^{(1)} - y^{(1)}\right) \cdot x_0^{(1)}$.

Проведя аналогичные рассуждения, получим, что для $i \in \{0, 1, \ldots, n\}$ выполнено

$$\frac{\partial J}{\partial \theta_i}(\theta) = \left(\sum_{j=0}^{n} \theta_j x_j^{(1)} - y^{(1)}\right) \cdot x_i^{(1)}.$$

Объясним то же самое словами. Итак, чтобы в задаче линейной регрессии для одного объекта посчитать частную производную от функции потерь по i-ому параметру θ_i, мы должны:

1. посчитать разность между предсказанием нашей линейной регрессии и реальным ответом,
2. умножить полученное число на значение i-ого признака нашего объекта.

Замечание. Когда мы считали $\frac{\partial J}{\partial \theta_0}(\theta)$, мы могли вспомнить, что $x_0^{(1)} = 1$. Соответственно, ответ немного упростится: $\frac{\partial J}{\partial \theta_0}(\theta) = \sum\limits_{j=0}^{n} \theta_j x_j^{(1)} - y^{(1)}$.

Теперь запишем ответ к задаче: $\nabla J(\theta) = \left(\sum\limits_{j=0}^{n} \theta_j x_j^{(1)} - y^{(1)}\right) \cdot \left(x_0^{(1)}, x_1^{(1)}, \ldots, x_n^{(1)}\right)$. Здесь мы вынесли множитель $\left(\sum\limits_{j=0}^{n} \theta_j x_j^{(1)} - y^{(1)}\right)$, на который умножается каждая из координат градиента, за скобку. Ответ можно записать и ещё короче: $\nabla J(\theta) = \left(\sum\limits_{j=0}^{n} \theta_j x_j^{(1)} - y^{(1)}\right) \cdot x^{(1)}$.

Или, эквивалентно, $\nabla J(\theta) = \left(\hat{y}^{(1)} - y^{(1)}\right) \cdot x^{(1)}$

Задача 4

Найдите $\nabla J(\theta) := \left(\frac{\partial J}{\partial \theta_0}(\theta), \frac{\partial J}{\partial \theta_1}(\theta), \ldots, \frac{\partial J}{\partial \theta_n}(\theta)\right)$ в общем случае. То есть когда мы не требуем, чтобы m было равно 1.

Подсказка. $J(\theta)$ предствляется в виде суммы выражений, градиенты которых мы научились считать в прошлой задаче.

Решение. Вспомним определение: $J(\theta) := \frac{1}{2} \sum\limits_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$.

Обозначим значение функции потерь для i-ого объекта выборки за $J^{(i)}(\theta) := \frac{1}{2} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$.

Тогда функцию потерь на всей выборке можно представить в виде $J(\theta) = \sum\limits_{i=1}^{m} J^{(i)}(\theta)$.

Из того, что частная производная от суммы функций равна сумме частных производных этих функций, будет следовать, что градиент от суммы равен сумме градиентов: $\nabla J(\theta) = \sum\limits_{i=1}^{m} \nabla J^{(i)}(\theta)$.

А градиент функции потерь для одного объекта мы научились считать в предыдущей задаче: $\nabla J^{(i)}(\theta) = \left(\sum\limits_{j=0}^{n} \theta_j x_j^{(i)} - y^{(i)}\right) \cdot x^{(i)}$.

Итого ответ в нашей задаче $\nabla J(\theta) = \sum\limits_{i=1}^{m} \left(\sum\limits_{j=0}^{n} \theta_j x_j^{(i)} - y^{(i)}\right) \cdot x^{(i)}$.

Или, эквивалентно, $\nabla J(\theta) = \sum\limits_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)}\right) \cdot x^{(i)}$.

5.3 Стохастический градиентный спуск и English

Задача 1

Почему спуск называется "стохастическим"? Слово "стохастический" это синоним слова "случайный". Применяя SGD, мы перемешиваем объекты из обучающей выборки случайным образом. Поэтому в SGD есть элемент случайности. В частности, если вы сделаете SGD два раза (с одинаковыми стартовыми точками и learning rate), вы можете получить два разных результата.

Если мы сделаем обычный градиентный спуск два раза (с одинаковыми стартовыми точками и learning rate), мы можем получить два разных результата?

Под "разными результатами" мы имеем в виду следующее. Для какого-то i на i-ом шаге второго запуска градиентный спуск окажется не в той же точке, что на i-ом шаге первого запуска. Другими словами, на каком-то шаге путь первого градиентного спуска будет отличаться от пути второго градиентного спуска.

Подсказка. Вспомните алгоритм градиентного спуска.

Решение. Итак, смотрим указанный в подсказке шаг:

Пусть дана функция $f : \mathbb{R}^n \to \mathbb{R}$, начальная точка x_0 и learning rate λ. Тогда алгоритм градиентного спуска для поиска минимума функции f выглядит следующим образом.

1. Обозначим за i номер шага градиентного спуска. Сейчас $i = 1$.
2. Вычислим градиент $\nabla f(x_i)$.
3. Если $\nabla f(x_i) = 0$, то алгоритм останавливается. Иначе делаем шаг градиентного спуска: переходим в точку $x_{i+1} = x_i - \lambda \nabla f(x_i)$.
4. Заменяем i на $i + 1$ и повторяем шаги 2, 3, 4.

По формуле из пункта 3 видим, что все точки, в которые попадает градиентный спуск для функции f, однозначно задаются номером шага, стартовой точкой x_0 и learning rate.

Действительно, подставим в формулу из пункта 3 $x_i = x_{i-1} - \lambda \nabla f(x_{i-1})$:

$x_{i+1} = x_i - \lambda \nabla f(x_i) = x_{i-1} - \lambda \nabla f(x_{i-1}) - \lambda \nabla f(x_{i-1} - \lambda \nabla f(x_{i-1}))$. Мы сможем и дальше продолжать эту цепочку равенств, выражая точки с большим индексом через точки с меньшим индексом, пока не придём к точке x_0.

Задача 2

В GD мы делаем шаг, вычисляя градиент функции потерь для всех m объектов обучающей выборки. В SGD мы делаем шаг, вычисляя градиент функции потерь только для 1 объекта обучающей выборки. Как делать шаг, вычисляя градиенты функции потерь для 16 объектов обучающей выборки?

Другими словами, придайте смысл выражению "делать шаг, вычисляя градиенты функции потерь для 16 объектов обучающей выборки".

Подсказка. Вспомните формулу, которую мы получили в задаче про градиент от функции потерь.

Решение.

Вспомним, что один шаг градиентного спуска задаётся следующим преобразованием координат вектора $(\theta_0, \theta_1, \ldots, \theta_n)$:

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial J}{\partial \theta_j}(\theta) = \theta_j - \alpha \sum_{i=1}^{m} \left(\hat{y}^{(i)} - y^{(i)} \right) \cdot x_j^{(i)} \text{ для всех } j \in \{0, 1, \ldots, n\}$$

А один шаг *стохастического градиентного спуска (или SGD)* же выглядит следующим образом.

$$\theta_j \leftarrow \theta_j - \alpha \left(\hat{y}^{(i)} - y^{(i)} \right) \cdot x_j^{(i)} \text{ для всех } j \in \{0, 1, \ldots, n\}$$

А делать шаг, вычисляя градиенты функции потерь для 16 объектов обучающей выборки мы будем по формуле

$$\theta_j \leftarrow \theta_j - \alpha \sum_{i=1}^{16} \left(\hat{y}^{(i)} - y^{(i)} \right) \cdot x_j^{(i)} \text{ для всех } j \in \{0, 1, \ldots, n\}$$

На следующем шаге мы будем суммировать уже от 17 до 32, затем от 33 до 48 и т.д.

Как и в стохастическом градиентном спуске, в начале каждой эпохи выборка перемешивается.

В этой задаче мы для простоты считали, что мы решаем задачу линейной регрессии. А в общем случае мы будем суммировать градиенты функций потерь для каждого из 16 объектов:

$$\theta_j \leftarrow \theta_j - \alpha \sum_{i=1}^{16} \nabla J^{(i)}(\theta) \text{ для всех } j \in \{0, 1, \ldots, n\}$$

Здесь $J^{(i)}(\theta)$ — значение функции потерь для i-ого объекта выборки.

5.4 Градиентный спуск с моментом

Задача 1

Напомним формулу для V_t:
$V_t = (1-\beta)(\beta^{t-1}S_1 + \beta^{t-2}S_2 + \cdots + \beta^2 S_{t-2} + \beta^1 S_{t-1} + S_t)$,
при этом $0 < \beta < 1$.
Придайте смысл фразе "V_t придаёт большее значение более поздним шагам".

Подсказка. Посмотрите на коэффициенты при S_1 и при S_t. Какой из них больше?

Решение. $V_t = (1-\beta)\beta^{t-1}S_1 + (1-\beta)\beta^{t-2}S_2 + \cdots + (1-\beta)\beta^2 S_{t-2} + (1-\beta)\beta^1 S_{t-1} + (1-\beta)\beta^0 S_t$

Пусть k и q — такие натуральные числа, что $1 \le k < q \le t$. Рассмотрим, на какой коэффициент в формуле выше будет умножаться S_k, а на какой S_q. Так мы узнаем, какое "значение" придаёт V_t более поздним шагам (S_q) относительно более ранних (S_k).

Итак, S_k будет умножаться на $(1-\beta)\beta^{t-k}$, а S_q — на $(1-\beta)\beta^{t-q}$. Поскольку $k < q$, то $t-k > t-q$. А так как $0 < \beta < 1$, то $\beta^{t-k} < \beta^{t-q}$. Итак, коэффициент, на который умножается S_q больше коэффициента, на который умножается S_k. Что и означает, что V_t придаёт большее значение более поздним шагам.

Задача 2

На четвёртом шаге этого урока мы сделали так:

- $V_0 = 0$, потому что V_0 агрегирует информацию с 0 шагов
- $V_1 = \beta V_0 + (1-\beta)S_1 = (1-\beta)S_1$

То есть при $\beta = 0.9$ мы получаем $V_1 = (1-0.9)S_1 = 0.1S_1$.

Почему было бы неразумно начать с 1, взяв просто $V_1 = S_1$?

Коль скоро V_1 агрегирует информацию всего с одного шага, выбор $V_1 = S_1$ кажется естественным.

Подсказка. Посмотрите, как будет выглядеть формула для V_t в случае $V_1 = S_1$.

Решение.

Вспомним формулу $V_t = (1-\beta)(\beta^{t-1}S_1 + \beta^{t-2}S_2 + \cdots + \beta^2 S_{t-2} + \beta^1 S_{t-1} + S_t)$.

Рассмотрим, как бы она выглядела при $\widetilde{V}_1 = S_1$. Будем писать $\widetilde{V}_i$, чтобы подчеркнуть, что мы используем другую формулу для первого элемента.

$\widetilde{V}_2 = \beta\widetilde{V}_1 + (1-\beta)S_2 = \beta S_1 + (1-\beta)S_2$

$\widetilde{V}_3 = \beta\widetilde{V}_2 + (1-\beta)S_3 = \beta^2 S_1 + \beta(1-\beta)S_2 + (1-\beta)S_3$

$\widetilde{V}_4 = \beta\widetilde{V}_3 + (1-\beta)S_4 = \beta^3 S_1 + \beta^2(1-\beta)S_2 + \beta(1-\beta)S_3 + (1-\beta)S_4$

Или, в общем случае $\widetilde{V}_t = \beta^{t-1}S_1 + (1-\beta)(\beta^{t-2}S_2 + \cdots + \beta^2 S_{t-2} + \beta^1 S_{t-1} + S_t)$.

Видим, что V_t и $\widetilde{V}_t$ отличаются только множителем при S_1: в формуле для V_t он равен $(1-\beta)\beta^{t-1}$, а в формуле для $\widetilde{V}_t$ — β^{t-1}. При этом, так как $\beta < 1$, то $(1-\beta)\beta^{t-1} < \beta^{t-1}$. А если взять β близким к единице, то

эти два коэффициента будут отличаться на порядок. То есть, мы получили, что $\widetilde{V}_t$ придаёт большее значение S_1, чем V_t. А мы не хотим, чтобы значение градиента в первой точке сильно влияло на V_t, мы хотим, чтобы большее значение придавалось более поздним шагам. Поэтому мы будем брать именно $V_0 = 0$, а не $V_1 = S_1$.

Задача 3

Сейчас мы докажем формулу, которая понадобится нам на следующем шаге.
Докажите, что $(1 + \beta + \beta^2 + \cdots + \beta^{t-2} + \beta^{t-1})(1 - \beta) = 1 - \beta^t$ для любого $t \in \mathbb{N}$ и любого $\beta \in \mathbb{R}$.
В частности, для любого $\beta \neq 1$ выполнено $1 + \beta + \beta^2 + \cdots + \beta^{t-2} + \beta^{t-1} = \frac{1-\beta^t}{1-\beta}$.
Обсудите следующий шаг с преподавателем.

Подсказка. Раскройте скобки в выражении слева.

Решение. Раскроем скобки: $(1 + \beta + \beta^2 + \cdots + \beta^{t-2} + \beta^{t-1})(1 - \beta) =$
$= 1 + \beta + \beta^2 + \cdots + \beta^{t-2} + \beta^{t-1} - \beta - \beta^2 - \beta^3 - \cdots - \beta^{t-1} - \beta^t$. Все β^i, где $i \in \{1, 2, \ldots, t-1\}$ сократятся. Останется как раз $1 - \beta^t$.

5.5 RMSprop

Задача 1

Чтобы лучше понять конструкцию RMSprop давайте на конкретном примере посмотрим, как он работает.
Пример. Пусть в первой точке r_1 градиент был равен S_1. Мы сделали первый шаг градиентного спуска с RMSprop и попали в точку r_2. В ней градиент равен S_2. На какой вектор мы будем сдвигаться на втором шаге градиентного спуска?
Пусть градиенты функции потерь на первом и втором шаге такие:

- $S_1 = (s_{1,1}, s_{1,2}, s_{1,3}) = (5, 2, 0)$,
- $S_2 = (s_{2,1}, s_{2,2}, s_{2,3}) = (-7, 3, 0)$.

А коэффициенты такие: $\beta_2 = 0.9$, $\varepsilon = 10^{-8}$ и $\alpha = 1$.
Решение. Мы хотим найти вектор, на который мы сместимся на втором шаге. Для этого нужно найти вектор $U_2 = (u_{2,1},\ u_{2,2}, u_{2,3})$ – вектор, состоящий из экспоненциально взвешенных средних каждой из трёх координат за первые два шага.
По формуле с предыдущего шага $u_{2,j} = 0.09 s_{1,j}^2 + 0.1 s_{2,j}^2$, где $j \in \{1, 2, 3\}$.

1. Рассмотрим первую координату $j = 1$. Получаем, что $u_{2,1} = 0.09 s_{1,1}^2 + 0.1 s_{2,1}^2 = 0.09 \cdot 5^2 + 0.1 \cdot (-7)^2 = 7.15$. Чтобы посчитать первую координату вектора, на который мы сместимся на втором шаге, мы хотим найти, чему равно $(-1) \cdot s_{2,1}$, поделённое на $\sqrt{\frac{u_{2,1}}{1-0.9^2}} + 10^{-8}$. Минус перед $s_{2,1}$ возник из-за умножения на $-\alpha = -1$. Найдём сначала величину: $\sqrt{\frac{u_{2,1}}{1-0.9^2}} = \sqrt{\frac{7.15}{1-0.9^2}} \approx 6.13$. Знак $\approx$ здесь и далее означает, что мы вычисляем с точностью до второго знака после запятой. Итак, первая координата вектора, на который мы сделаем шаг, равна $\frac{-\ s_{2,1}}{6.13+10^{-8}} = \frac{7}{6.13+10^{-8}} \approx 1.14$.

2. Для второй координаты $j = 2$ мы получим, что $u_{2,2} = 0.09 s_{1,2}^2 + 0.1 s_{2,2}^2 = 0.09 \cdot 2^2 + 0.1 \cdot 3^2 = 1.26$. Средняя длина шага при этом будет равна $\sqrt{\frac{u_{2,2}}{1-0.9^2}} = \sqrt{\frac{1.26}{1-0.9^2}} \approx 2.58$. А вторая координата вектора, на который мы сделаем шаг, равна $\frac{-s_{2,2}}{2.58+10^{-8}} = \frac{-3}{2.58+10^{-8}} \approx -1.16$.

3. Для третьей координаты $j = 3$ мы получим, что $u_{2,3} = 0.09 s_{1,3}^2 + 0.1 s_{2,3}^2 = 0.09 \cdot 0^2 + 0.1 \cdot 0^2 = 0$, $\sqrt{\frac{u_{2,3}}{1-0.9^2}} = \sqrt{\frac{0}{1-0.9^2}} = 0$ и третья координата вектора, на который мы сделаем шаг, равна $\frac{-s_{2,3}}{0+10^{-8}} = \frac{0}{0+10^{-8}} = 0$. Кстати, здесь мы как раз видим, для чего нам нужен был ε: без него возникло бы деление на ноль.

Итого, мы получили, что $U_2 = (7.15, 1.26, 0)$, а вектор, на который мы сместимся на втором шаге, равен $(1.14, -1.16, 0)$. Обратите внимание, насколько абсолютные значения каждой из координат близки к единице (они оказываются близки к learing rate $\alpha = 1$).

Задача. Продолжаем рассматривать ситуацию из примера выше. Мы сделали второй шаг градиентного шага на вектор $(1.14, -1.16, 0)$ и оказались в точке r_3. Градиент в ней оказался такой:

$S_3 = (s_{3,1}, s_{3,2}, s_{3,3}) = (6, 4, -1)$.

Найдите вектор, на который мы сместимся на третьем шаге (в процессе найдя вектор U_3).

Подсказка. Посмотрите на формулу для $u_{3,j}$ с предыдущего шага.

Решение. На предыдущем шаге мы посчитали, что $u_{3,j} = 0.1 \cdot (0.81 s_{1,j}^2 + 0.9 s_{2,j}^2 + s_{3,j}^2)$.

1. Для $j = 1$ получаем $u_{3,1} = 0.081 s_{1,1}^2 + 0.09 s_{2,1}^2 + 0.1 s_{3,1}^2 = 0.081 \cdot 5^2 + 0.09(-7)^2 + 0.1 \cdot 6^2 = 10.035$.

   Итак, первая координата вектора, на который мы сделаем шаг, равна

$$\frac{-s_{3,1}}{\sqrt{\frac{u_{3,1}}{1-0.9^3}} + 10^{-8}} = \frac{-6}{\sqrt{\frac{10.035}{1-0.9^3}} + 10^{-8}} \approx -0.99$$

2. Для $j = 2$ получаем $u_{3,2} = 0.081 s_{1,2}^2 + 0.09 s_{2,2}^2 + 0.1 s_{3,2}^2 = 0.081 \cdot 2^2 + 0.09 \cdot 3^2 + 0.1 \cdot 4^2 = 2.734$.

   Итак, вторая координата вектора, на который мы сделаем шаг, равна

$$\frac{-s_{3,2}}{\sqrt{\frac{u_{3,2}}{1-0.9^3}} + 10^{-8}} = \frac{-4}{\sqrt{\frac{2.734}{1-0.9^3}} + 10^{-8}} \approx -1.26$$

3. Для $j = 3$ получаем $u_{3,3} = 0.081 s_{1,3}^2 + 0.09 s_{2,3}^2 + 0.1 s_{3,3}^2 = 0.081 \cdot 0^2 + 0.09 \cdot 0^2 + 0.1 \cdot (-1)^2 = 0.1$.

   Итак, третья координата вектора, на который мы сделаем шаг, равна $\frac{-s_{3,3}}{\sqrt{\frac{u_{3,3}}{1-0.9^3}} + 10^{-8}} = \frac{1}{\sqrt{\frac{0.1}{1-0.9^3}} + 10^{-8}} \approx 1.65$

Наконец, ответ в задаче такой: $U_3 = (10.035, 2.734, 0.1)$ и вектор, на который мы сместимся на третьем шаге, равен $(-0.99, -1.26, 1.65)$.

Задача 2

Не менее важная проблема, чем описанная в первом шаге этого урока.

Иногда наш градиентный спуск может попасть на плато – область, где градиенты очень маленькие. Поэтому перемещение по плато идёт медленно, и его желательно ускорить. Рассмотрим такой игрушечный пример.

Пусть learning rate нашего RMSprop равен $\alpha = 0.2$. Параметр $\beta_2 = 0.9$.

На шаге t мы имеем $U_t = (100, 100)$. В этот момент мы попали на плато – во всех последующих точках у нас будет градиент $S = \left(-\frac{1}{20}, \frac{1}{200}\right)$.

Посмотрите, как будет меняться вектор, на который мы смещаемся, в течение следующих шагов. Для этого будет полезно выразить вектора $U_{t+1}, U_{t+2}, \dots$ через S, β_2 и U_t.

Задача. Покажите, что через большое количество шагов вектор, на который мы смещаемся, будет близок к $(\alpha, -\alpha)$. В тех же условиях обычный градиентный спуск давал бы смещение на вектор $\left(\frac{\alpha}{20}, -\frac{\alpha}{200}\right)$.

Тем самым RMSprop как бы подталкивает величину изменений по каждой координате к числу α. Это решает и проблему слишком больших изменений, и проблему слишком маленьких изменений.

Подсказка. Через S, β_2, k и U_t можно выразить вектор U_{t+k}, где $k \in \mathbb{N}$. А затем посмотреть, к чему стремится это выражение при $k \to \infty$.

Решение.

Пусть $k \in \mathbb{N}$. Поскольку все $S_{t+k} = S_t = \left(-\frac{1}{20}, \frac{1}{200}\right) = S$, то за S^2 соответственно обозначим $\left(\frac{1}{20^2}, \frac{1}{200^2}\right)$.

Далее, $U_{t+1} = U_t \cdot \beta_2 + S^2 \cdot (1 - \beta_2)$

$U_{t+2} = (U_t \cdot \beta_2 + S^2 \cdot (1 - \beta_2)) \cdot \beta_2 + S^2(1 - \beta_2) = U_t \cdot \beta_2^2 + S^2(1 - \beta_2)(\beta_2 + 1)$

$U_{t+3} = (U_t \cdot \beta_2^2 + S^2(1 - \beta_2)(\beta_2 + 1)) \cdot \beta_2 + S^2(1 - \beta_2) = U_t \cdot \beta_2^3 + S^2(1 - \beta_2)(\beta_2^2 + \beta_2 + 1)$

Видно, что $U_{t+k} = U_t \cdot \beta_2^k + S^2(1 - \beta_2)(\beta_2^{k-1} + \beta_2^{k-2} + \dots + \beta_2 + 1)$

При этом, поскольку по задаче 3 из урока 5.3 $\beta_2^{k-1} + \beta_2^{k-2} + \dots + \beta_2 + 1 = \frac{1-\beta_2^k}{1-\beta_2}$, окончательно получаем $U_{t+k} = U_t \cdot \beta_2^k + S^2(1 - \beta_2^k)$.

И, подставляя $\beta_2 = 0.9$, получаем $U_{t+k} = 0.9^k U_t + S^2(1 - 0.9^k)$. Как мы помним по первой неделе нашего курса, $\lim\limits_{k\to\infty} 0.9^k = 0$. Значит, при больших k выражение 0.9^k близко к нулю, а $1 - 0.9^k$ — близко к единице, то есть U_{t+k} близко к S^2.

Далее, мы должны посчитать $\sqrt{U_{t+k}/(1 - 0.9^{t+k})}$. При больших k числитель близок к S^2, а знаменатель — к единице. Итого корень близок к $|S| = \left(\frac{1}{20}, \frac{1}{200}\right)$.

Тогда вектор, на который мы будем сдвигаться на шаге $t+k$ равен $-\alpha \cdot S_{t+k}/(\sqrt{U_{t+k}/(1 - 0.9^{t+k})}+\varepsilon)$ (здесь мы делим покоординатно), что близко к $-\alpha S/|S|$ при больших k. Поскольку $S/|S| = (-1, 1)$, то $-\alpha S/|S| = (\alpha, -\alpha)$.

5.6 Adam

Задача 1

Своими словами перескажите алгоритм Adam (с открытым текстом второй страницы статьи)
Для каждой строки алгоритма скажите, что мы делаем и зачем.

Подсказка. Перечитайте шаги этого урока

Решение.
Алгоритм Adam
Пусть даны:

- функция $f(\theta)$, минимум которой мы хотим найти (в нашем случае это функция потерь), где θ — параметры (в общем случае это многомерный вектор)
- θ_0 — стартовое значение параметров θ
- число α — learning rate, по дефолту можно взять $\alpha = 0.001$
- числа $\beta_1, \beta_2 \in [0, 1)$, по дефолту можно взять $\beta_1 = 0.9$ и $\beta_2 = 0.999$
- число $\epsilon > 0$, по дефолту можно взять $\epsilon = 10^{-8}$

Пусть θ_t — значение параметров θ на шаге номер t. Обозначим среднее градиентов за m_t, а среднее квадратов градиентов — за v_t. За g_t обозначим значение градиента функции f на шаге номер t. g_t^2 означает, что мы каждую координату вектора g_t возводим в квадрат. β_1^t означает возведение β_1 в степень t. То же самое для β_2^t.

В алгоритме все преобразования векторов происходят покоординатно. Например, если написана формула вида $\frac{a_t}{b_t}$ (как в пункте 5), то подразумевается, что первая координата вектора a_t делится на первую координату вектора b_t, вторая — делится на вторую и т.д.

Положим $m_0 = 0$ и $v_0 = 0$. Далее в скобочках будем писать комментарии к каждому из пунктов.

1. Обозначим номер шага за t. Сейчас $t = 1$.
2. Если θ_t последние несколько шагов менялось несильно, то алгоритм останавливается.

   Иначе вычисляем g_t — значение градиента функции f на шаге номер t: $g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$.

   (Ведь мы хотим найти точку минимума функции f, а из утверждения прошлой недели следует, что шаг мы должны делать в направлении, противоположном направлению градиента)
3. Вычисляем среднее градиентов: $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$

   (Вектор m_t нужен нам, чтобы помнить, какие были градиенты на прошлых шагах)

   Далее вычисляем экспоненциально взвешенное среднее градиентов: $\widehat{m}_t \leftarrow m_t / (1 - \beta_1^t)$

   (Как мы обсуждали ранее, коэффициент $\frac{1}{1-\beta_1^t}$ позволяет делать первые шаги больше, к тому же с ним формула действительно по смыслу ближе к "среднему")
4. Вычисляем среднее квадратов градиентов: $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$

   (С помощью v_t мы хотим уменьшить слишком большие координаты вектора $\hat{m}_t$ и увеличить слишком маленькие координаты вектора $\hat{m}_t$. Кроме того, v_t помогает решить проблему с плато - см. задачу 3 по RMSprop)

   И вычисляем экспоненциально взвешенное среднее квадратов градиентов: $\widehat{v}_t \leftarrow v_t / (1 - \beta_2^t)$

5. Делаем шаг: переходим в точку $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t / \left(\sqrt{\hat{v}_t} + \epsilon\right)$

   (Если бы формула выглядела просто как $\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \hat{m}_t$, то это был бы алгоритм градиентного спуска с моментом. По причине из 4 пункта мы добавляем ещё и деление на $\sqrt{\hat{v}_t} + \epsilon$, при этом ϵ нужен только для того, чтобы случайно не поделить на ноль)
6. Заменяем t на $t + 1$ и повторяем пункты 2 – 5.