Математика для Data Science. Математический анализ. Условия 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. $(h_\theta(x^{(2)}) - y^{(2)})^2$

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.

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 $\frac{1}{m}$.

Пример. Рассмотрим случай $n = 1$. То есть наш единственный объект описывается парой чисел $(x_0^{(1)}, x_1^{(1)})$. Пусть $(x_0^{(1)}, x_1^{(1)}) = (4, 7)$, и $y^{(1)} = 5$.

Тогда функция потерь это

$$J(\theta) = \frac{1}{2}\left((\theta_0\ x_0^{(1)} + \theta_1\ x_1^{(1)}) -\ y^{(1)}\right)^2 = \frac{1}{2}\left((4\theta_0\ + 7\theta_1) -\ 5\right)^2 = \frac{1}{2}\cdot\left((4\theta_0 + 7\theta_1) -\ 5\right)\cdot\left((4\theta_0 + 7\theta_1) -\ 5\right).$$

1. Найдём $\frac{\partial J}{\partial \theta_0}$. Функция J это произведение константы $\frac{1}{2}$ и двух одинаковых сомножителей $((4\theta_0+7\theta_1)-5)$. По правилу дифференцирования произведения получаем:

$$\frac{\partial J}{\partial \theta_0} = \frac{1}{2}\left(\frac{\partial((4\theta_0+7\theta_1)-5)}{\partial \theta_0}\cdot((4\theta_0+7\theta_1)-5)+((4\theta_0+7\theta_1)-5)\cdot\frac{\partial((4\theta_0+7\theta_1)-5)}{\partial \theta_0}\right) = =\frac{1}{2}\left(2\frac{\partial((4\theta_0+7\theta_1)-5}{\partial \theta_0}\right.$$

2. Найдём $\frac{\partial J}{\partial \theta_1}$. Аналогично вычислению для θ_0 получаем:

$$\frac{\partial J}{\partial \theta_0} = \frac{\partial((4\theta_0+7\theta_1)-5)}{\partial \theta_1}\cdot((4\theta_0+7\theta_1)-5) = 7((4\theta_0+7\theta_1)-5).$$

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.

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.

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".

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".
Пример. Пусть $\beta = 0.9, t = 5$. Тогда
S_1 входит в V_5 с коэффициентом $(1-\beta)\beta^{t-1} = 0.1 \cdot 0.9^4 = 0.06561$ S_3 входит в V_5 с коэффициентом $(1-\beta)\beta^{t-3} = 0.1 \cdot 0.9^2 = 0.081$ заметим, что $0.081 > 0.06561$ Третий шаг был позже, чем первый шаг. В V_t вектор S_3 входит с бОльшим коэффициентом, чем вектора S_1.

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.

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.

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).

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.

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.