Познавач на Височина

Решение с Линейна Регресия

Николай Пашов

30 май 2022 г.

Съдържание

1 Описание на решението

Същината на задачата е чрез линейна регресия да се намери функция $f : \mathbb{R}^3 \to \mathbb{R}$, която изчислява височината на човек, разполагайки с възрастта и тежестта. Знаем, че:

$$f(age, weight) = a_0 + a_1 * age + a_2 * weight$$

Дефинираме функция на грешката:

$$L(y, gt) = (y - gt)^2$$

, където $y = f(age, weight)$ за случайни $age, weight$, а gt (ground truth) е истинската стойност от базата данни, която съответства на дадените $age, weight$.

Задачата се свежда до оптимизационна задача на функцията L чрез параметрите a_0, a_1, a_2. Методът за оптизация е градиентно спускане.

Алгоритъмът се състои от множество итерации през данните за трениране, в които се случва оптимизация на параметрите посредством техните частни производни спрямо L. Добавя се множител lr (learning rate), с който да се контролира оптимизационната стъпка:

$$a_x \leftarrow \frac{\partial L}{\partial a_x} * lr$$

С прости сметки се вижда, че:

$$\frac{\partial L}{\partial a_x} = \frac{\partial L}{\partial f} * \frac{\partial f}{\partial a_x} = 2 * (y - gt) * a_x, x \neq 0$$

$$\frac{\partial L}{\partial a_x} = \frac{\partial L}{\partial f} * \frac{\partial f}{\partial a_x} = 2 * (y - gt) * 1, x = 0$$

Съответната сметка се прави по време на итериране през данните за трениране, откъдето се взимат стойностите на y и gt.

Крайната оценка на тренирането се извършва с друга функция на грешката:

$$Acc = \frac{1}{n} * \sum_{i=0}^{n} abs(prediction(test[i][age], test[i][weight]), test[i][height])$$

1.1 Псевдокод

$(train, test) \leftarrow input\_data.split(0.8, 0.2)$
$[a_0, a_1, a_2] \leftarrow [random(a_0, a_1, a_2)]$
$f(age, weight) = a_0 + age * a_1 + weight * a_2$
$iteration \leftarrow 0$
$lr \leftarrow 2e - 6$
while $it \neq length(train)$ **do**
    $prediction \leftarrow f(train[it][age], train[it][weight])$
    $a_0 \leftarrow a_0 - 2$*(prediction - data[it][height]) $* lr)$
    $a_1 \leftarrow a_1 - 2 * (prediction - data[it][height]) * a_1 * lr)$
    $a_2 \leftarrow a_2 - 2 * (prediction - data[it][height]) * a_2 * lr)$

While цикълът се изпълнява произволен брой пъти. Броят изпълнение е хиперпараметър, който подлежи на експериментиране

2 Инструкции за компилиране

Имплементацията е реализирана на Python 3.10. Очаква се да бъде изпълнима и на по-стари версии, заради backward compatibility, но не е тествана и не се гарантира от автора. Изисква се инсталиран Python и python командата в терминала да извиква интерпретатор на Python 3.10. За изпълнение на програмата:

cd <code_directory>
python <filename>

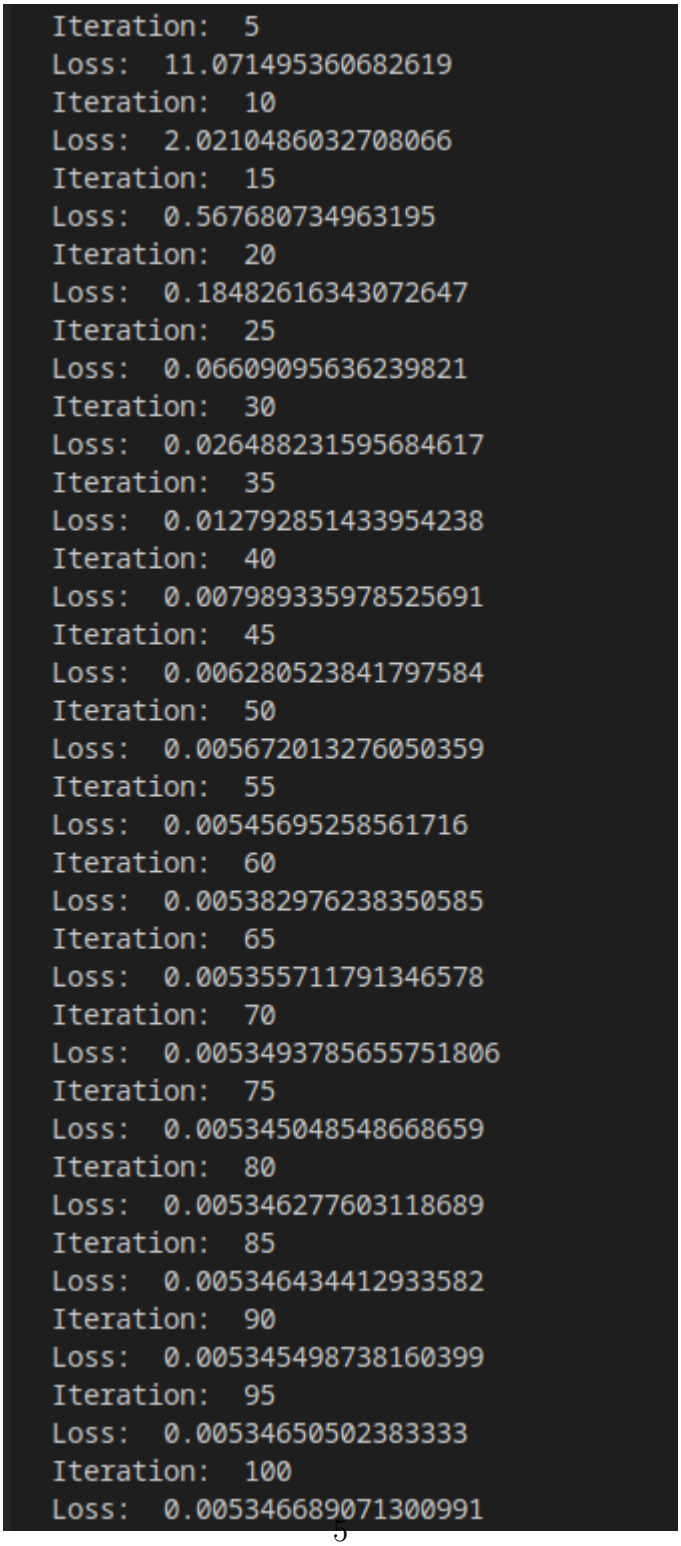
Изисква се .csv файлът с данните за трениране да е в същата директория и да се казва "age_weight.csv"

3 Примерни резултати

3.1 Експериментиране

В програмата е реализирано търсене на оптимална скорост на трениране. То работи по следния начин:

- Задава фиксирани скорости (lr) на трениране в масив.
- Разцепва train множеството на train и validation (Пропорции - 4:1).



Фигура 1: Трениране върху тестово множество със 100 итерации. $lr = 6e - 04$. Loss е функцията L която се пресмята на всеки 5 обучителни цикъла (обхождане с трениране на параметрите върху множеството за трениране). Стойността е осреднена върху елементите от множеството за трениране.

```
lr:  2e-05  Acc:  0.05489679759376924
lr:  5e-05  Acc:  0.0131378065704029
lr:  7e-05  Acc:  0.06421912394091132
lr:  0.0001  Acc:  0.06136979687598651
lr:  0.0004  Acc:  0.023950766288360213
lr:  0.0007  Acc:  0.024044680138715156
lr:  0.001  Acc:  0.02397123441187671
lr:  0.003  Acc:  0.02407774423056882
lr:  0.005  Acc:  0.27407520285230946
lr:  0.01  Acc:  0.6296062331128828
Best learning rate according to hyperparameter search:  5e-05
Actual best learning rate:  7e-05
```

Фигура 2: Експеримент с търсене на скорост на учене.

- Извършва трениране за всяко lr и пресмята Acc за съответното lr.
- Предикторите от всичките тренирания се тестват спрямо множеството за тестване и се пресмята кое lr е в действителност най-доброто.

4 Анализ

В настоящия проект беше реализиран оптимизационен алгоритъм за линейна функция чрез линейна регресия. Намаляващата функция на грешката L при трениране индицира, че обучение наистина се случва и алгоритъмът се подобрява. Крайните стойности на функцията на грешката Acc варираха между 0.1 и 0.003, което са добри стойности в сравнение с диапазона на числата, които се познават (височината на хората).

Функцията за търсене на най-добра скорост на трениране невинаги открива най-оптималната скорост. Възможна причина за това е малкото количество данни. Въпреки това, може да се заключи, че при големи стойности на lr ($lr > 3e - 3$) функцията на грешката е осезателно по-голяма. Очакваната причина за това е, че с толкова голяма стъпка локалните минимуми на функцията лесно се "прескачат".