Тятя! Тятя! Нейросети заменили продавца!

Пилиф Ульянкин

https://github.com/FUlyankin/neural_nets_prob

Листочек 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[1], 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 три шага градиентного спуска. В качестве стартовой точки используйте $w_0 = 0$. В качестве скорости обучения возьмите $\eta = 0.1$.

в. Сделайте четыре 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 $\beta_1 = \beta_2 = 0.9, \eta = 0.1$

Решение:

а. Найдём теоретическую оценку стандартным МНК. Минимизируем MSE:

$$\text{MSE} = \frac{1}{2} \cdot ((2 - w)^2 + (3 - 2w)^2) \to \min_w$$

[1] Лёрнинг ей папа подарил

Берём производную, решаем уравнение и получаем ответ:

$$-(2-w)-2(3-2w)=0 \Rightarrow \hat{w}=\frac{8}{5}=1.6$$

б. Чтобы сделать три шага градиентного спуска, нужно найти градиент. Наша функция потерь выглядит как

$$L(w)=(y-wx)^2.$$

Мы подбираем один параметр, значит градиентом в данном случае будет просто одно число — производная по этому параметру.

$$\nabla L(w_0,x,y)=\frac{\partial L}{\partial w}=-2x(y-wx).$$

Стартовая точка $w_0=0$. Мы хотим сделать шаг

$$w_1=w_0-\eta\cdot\nabla L(w_0)$$

Посчитаем градиент в точке w_0 по всей выборке:

$$\nabla L(w_0)=\frac{1}{n}\cdot\sum_{i=1}^{n}\nabla L(w_0,x_i,y_i)=\frac{1}{2}\cdot(-2(2-w_0)-2\cdot 2(3-2w_0))=-8$$

Делаем **первый шаг**:

$$w_1=0+0.1\cdot 8=0.8$$

По аналогии, **второй шаг**:

$$\nabla L(w_1)=\frac{1}{2}\cdot(-2(2-w_1)-2\cdot 2(3-2w_1))=-4$$
$$w_2=w_1-\eta\cdot\nabla L(w_1)=0.8+0.1\cdot 4=1.2$$

По аналогии, **третий шаг**:

$$\nabla L(w_2)=\frac{1}{2}\cdot(-2(2-w_2)-2\cdot 2(3-2w_2))=-2$$
$$w_3=w_2-\eta\cdot\nabla L(w_2)=1.2+0.1\cdot 2=1.4$$

в. Теперь то же самое, на градиентный спуск стохастический. Мы будем считать $\nabla(w_t)$ не как среднее по всей выборке, а как значение градиента в одной случайно выбранной точке.

Первый шаг:

$$\nabla L(w_0)=-2(2-w_0\cdot 1)=-4$$
$$\beta_1=w_0-\eta\cdot\nabla L(w_0)=0+0.1\cdot 4=0.4$$

Второй шаг:

$$\nabla L(w_1) = -2 \cdot 2 \cdot (3 - w_1 \cdot 2) = -8.8$$
$$w_2 = w_1 - \eta \cdot \nabla L(w_1) = 0.4 + 0.1 \cdot 8.8 = 1.28$$

Третий шаг:

$$\nabla L(w_2) = -2(2 - w_1 \cdot 1) = -1.44$$
$$w_3 = w_2 - \eta \cdot \nabla L(w_2) = 1.28 + 0.144 = 1.424$$

Четвёртый шаг:

$$\nabla L(w_3) = -2 \cdot 2(3 - w_3 \cdot 2)$$
$$w_4 = w_3 - \eta \cdot \nabla L(w_4) = 1.424 + 0.1 \cdot 0.608 = 1.4848$$

г. Автор не очень хочет расписывать решение дальнейших четырёх пунктов. Но вы обязательно это сделайте.

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 $\gamma = 0.01$.

Решение:

Сначала нам надо найти $\text{logloss}'_\beta$. В принципе в этом и заключается вся сложность задачки. Давайте подставим вместо $\hat{p}$ в logloss сигмоиду.

$$\text{logloss} = -1 \left(y \cdot \ln \left(\frac{1}{1 + e^{-z}} \right) + (1 - y) \cdot \ln \left(1 - \frac{1}{1 + e^{-z}} \right) \right)$$

Теперь подставим вместо z уравнение регрессии:

$$\text{logloss} = -1 \left(y \cdot \ln \left(\frac{1}{1 + e^{-w \cdot x}} \right) + (1 - y) \cdot \ln \left(1 - \frac{1}{1 + e^{-w \cdot x}} \right) \right)$$

Это и есть наша функция потерь. От неё нам нужно найти производную. Давайте подготовимся.

Делай раз, найдём производную logloss по $\hat{p}$:

$$\text{logloss}'_{\hat{p}} = -1\left(y \cdot \frac{1}{\hat{p}} - (1-y) \cdot \frac{1}{(1-p)}\right)$$

Делай два, найдём производную $\frac{1}{1+e^{-wx}}$ по w:

$$\left(\frac{1}{1+e^{-wx}}\right)'_w = -\frac{1}{(1+e^{-wx})^2} \cdot e^{-wx} \cdot (-x) = \frac{1}{1+e^{-wx}} \cdot \frac{e^{-wx}}{1+e^{-wx}} \cdot x =$$
$$= \frac{1}{1+e^{-wx}} \cdot \left(1 - \frac{1}{1+e^{-wx}}\right) \cdot x$$

По-другому это можно записать как $\hat{p} \cdot (1-\hat{p}) \cdot x$.

Делай три, находим полную производную:

$$\text{logloss}'_{\beta} = -1\left(y \cdot \frac{1}{\hat{p}} \cdot \hat{p} \cdot (1-\hat{p})) \cdot x - (1-y) \cdot \frac{1}{(1-\hat{p})} \cdot \hat{p} \cdot (1-\hat{p})) \cdot x\right) =$$
$$= -y \cdot (1-\hat{p}) \cdot x + (1-y) \cdot \hat{p} \cdot x = (-y + y\hat{p} + \hat{p} - y\hat{p}) \cdot x = (\hat{p} - y) \cdot x$$

Найдём значение производной в точке $w_0 = 1$ для нашего наблюдения $x = -5, y = 1$:

$$\left(\frac{1}{1+e^{-1\cdot(-5)}} - 1\right) \cdot (-5) \approx 4.96$$

Делаем шаг градиентного спуска:

$$w_1 = 1 - 0.01 \cdot 4.96 \approx 0.95$$

Упражнение 3 (вопросики)

Убедитесь, что вы можете дать ответы на следующие вопросы:

- Как вы думаете, почему считается, что SGD лучше работает для оптимизации функций, имеющих больше одного экстремума?
- Предположим, что у функции потерь есть несколько локальных минимумов. Как можно адаптировать градиентный спуск так, чтобы он находил глобальный минимум чаще?
- Что будет происходить со стохастическим градиентным спуском, если длина его шага не будет уменьшаться от итерации к итерации?

Надо, чтобы кто-нибудь написал решение.

Упражнение 4 (скорости обучения)

В стохастическом градиентном спуске веса изменяются по формуле

$$w_t = w_{t-1} - \eta_t \cdot \nabla L(w_{t-1}, x_i, y_i),$$

где наблюдение i выбрано случайно, скорость обучения зависит от номера итерации.

Условия Роббинса-Монро гарантируют сходимость алгоритма к оптимуму для выпуклых дифференцируемых функций. Они говорят, что ряд из скоростей $\sum_{t=0}^{\infty} \eta_t$ должен расходиться, а ряд $\sum_{t=0}^{\infty} \eta_t^2$ сходиться. То есть скорость спуска должна падать не слишком медленно, но и не слишком быстро. Какие из последовательностей, перечисленных ниже, можно использовать для описания изменения скорости алгоритма?

а. $\eta_t = \frac{1}{t}$

б. $\eta_t = \frac{0.1}{t^{0.3}}$

в. $\eta_t = \frac{1}{\sqrt{t}}$

г. $\eta_t = \frac{1}{t^2}$

д. $\eta_t = e^{-t}$

е. $\eta_t = \lambda \cdot \left(\frac{s_0}{s_0 + t}\right)^p$, где λ, p и s_0 — параметры

Надо, чтобы кто-нибудь написал решение.