دوره‌ی آموزشی «علم داده»
Data Science Course

جلسه سی‌ام (بخش چهارم)

آشنایی با سه بهینه‌ساز بسیار مهم در یادگیری ماشین

AdaGrad, RMSprop & Adam

مدرس: محمد فزونی
عضو هیئت علمی دانشگاه گنبدکاووس

Learning rate schedules

نیازی نیست که تمام جزئیات این روش‌ها رو بدونیم. فقط نوع عملکردشون کافیه

AdaGrad **RMSProp**

TensorFlow

AdaGrad

/adaptive gradient algorithm/

2011

It dynamically varies the learning rate at each update and for each weight **individually**

این روش سعی داره نرخ یادگیری رو برای هر وزن، در هر بروزرسانی تغییر بده

قبلاً ما چه می‌کردیم؟

$$w(t+1) = w(t) - \eta \frac{\partial L}{\partial w}(t)$$

The next weight (at time t+1)

The previous weight (at time t)

Follow the update rule

AdaGrad

/adaptive gradient algorithm/

$$w(t+1) = w(t) - \eta \frac{\partial L}{\partial w}(t)$$

$$w(t+1) - w(t) = -\eta \frac{\partial L}{\partial w}(t)$$

$$\Delta w = -\eta \frac{\partial L}{\partial w}(\mathrm{t})$$

AdaGrad

/adaptive gradient algorithm/

$$\Delta w = -\eta \frac{\partial L}{\partial w}(\mathrm{t})$$

$$\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t) + \epsilon}} \frac{\partial L}{\partial w_i}(t)$$

AdaGrad

/adaptive gradient algorithm/

$$\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t) + \epsilon}} \frac{\partial L}{\partial w_i}(t)$$

$$\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t) + \epsilon}} \frac{\partial L}{\partial w_i}(t)$$

Weight index. The update rule is individual for each weight.

Iteration (time) at which we are updating

AdaGrad

/adaptive gradient algorithm/

$$\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t) + \epsilon}} \frac{\partial L}{\partial w_i}(t)$$

Adaptation magic

AdaGrad

/adaptive gradient algorithm/

$$\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t)} + \epsilon} \frac{\partial L}{\partial w_i}(t)$$

$$G_i(t) = G_i(t-1) + \left(\frac{\partial L}{\partial w_i}(t)\right)^2$$

with beginning point $G_i(0) = 0$

$G_i(0) = 0$

$G_i(1) = 0 + \text{non-neg}$

$G_i(2) = [0 + \text{non-neg}] + \text{non-neg}$

.

.

.

G (t) is monotonously increasing function (each consequent G is bigger or equal to the previous one)

AdaGrad

/adaptive gradient algorithm/

$$\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t)} + \epsilon} \frac{\partial L}{\partial w_i}(t)$$

$$G_i(t) = G_i(t-1) + \left(\frac{\partial L}{\partial w_i}(t)\right)^2$$

with beginning point $G_i(0) = 0$

- Smart
- Adaptive learning rate schedule
- Based on the training itself
- **Per weight**

Different weights do not reach their optimal values simultaneously

AdaGrad

/adaptive gradient algorithm/

$$\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t)} + \epsilon}\frac{\partial L}{\partial w_i}(t)$$

$$G_i(t) = G_i(t-1) + \left(\frac{\partial L}{\partial w_i}(t)\right)^2$$

with beginning point $G_i(0) = 0$

RMSprop

/root mean square propagation/

$$\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t)} + \epsilon}\frac{\partial L}{\partial w_i}(t)$$

$$G_i(t) = \beta G_i(t-1) + (1-\beta)\left(\frac{\partial L}{\partial w_i}(t)\right)^2$$

with beginning point $G_i(0) = 0$

AdaGrad

/adaptive gradient algorithm/

$$\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t)} + \epsilon}\frac{\partial L}{\partial w_i}(t)$$

$$G_i(t) = G_i(t-1) + \left(\frac{\partial L}{\partial w_i}(t)\right)^2$$

with beginning point $G_i(0) = 0$

RMSprop

/root mean square propagation/

$$\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t)} + \epsilon}\frac{\partial L}{\partial w_i}(t)$$

$$G_i(t) = \beta G_i(t-1) + (1-\beta)\left(\frac{\partial L}{\partial w_i}(t)\right)^2$$

with beginning point $G_i(0) = 0$

β – yet another hyperparameter
usually, around ~0.9

No longer monotonous, so it can adapt **upwards and downwards**

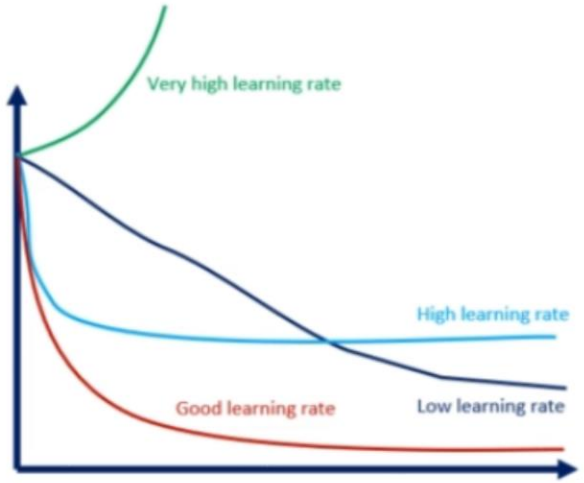
AdaGrad, RMSprop ?

$$\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t) + \epsilon}} \frac{\partial L}{\partial w_i}(t)$$

- Smart
- Adaptive learning rate schedule
- Based on the training itself
- **Per weight**

Can it get any better???

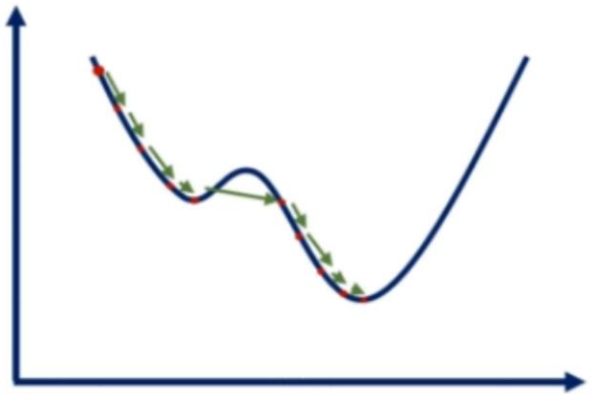
Learning rate schedules



$$\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t) + \epsilon}} \frac{\partial L}{\partial w_i}(t)$$

AdaGrad, RMSprop

Momentum



$$w \leftarrow w(t) - \eta \frac{\partial L}{\partial w}(t) - \alpha \eta \frac{\partial L}{\partial w}(t-1)$$

Current update

Update a moment ago

Adam

/adaptive moment estimation/

$$\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t)} + \epsilon}\frac{\partial L}{\partial w_i}(t)$$

AdaGrad, RMSprop

$$w \leftarrow w(t) - \eta\frac{\partial L}{\partial w}(t) - \alpha\eta\frac{\partial L}{\partial w}(t-1)$$

Current update

Update a moment ago

The most advanced optimizer (very fast and efficient)


The trends are ever-changing and you should stay informed an up to date

Adam

/adaptive moment estimation/

RMSprop

$$\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t) + \epsilon}} \frac{\partial L}{\partial w_i}(t)$$

Momentum

$$w \leftarrow w(t) - \eta \frac{\partial L}{\partial w}(t) - \alpha \eta \frac{\partial L}{\partial w}(t-1)$$

$$\Delta w_i(t) = -\frac{\eta}{\sqrt{G_i(t) + \epsilon}} M_i(t)$$

$$M_i(t) = \alpha M_i(t-1) + (1-\alpha) \frac{\partial L}{\partial w_i}(t)$$

As with all science, data science is a long chain of academic research building on top of each other

Enough is Enough

قبل از اینکه بریم سراغ چندتا پروژه‌ی انجام شده، در ویدیوی بعدی کمی راجع به پیش‌پردازش صحبت خواهیم کرد و بعد از اون

چیزهای باحال‌تری رو خواهم گفت