

1. Mini-batch Gradient Descent

Tags

Batch vs mini-batch gradient descent

Mini-batch gradient descent

Batch vs mini-batch gradient descent

Batch vs. mini-batch gradient descent

Vectorization allows you to efficiently compute on m examples.

$$\begin{aligned}
 X &= \begin{bmatrix} x^{(1)} & x^{(2)} & x^{(3)} & \dots & x^{(1000)} & | & x^{(1001)} & \dots & x^{(2000)} & | & \dots & | & \dots & x^{(m)} \end{bmatrix} \\
 &\quad (n_x, m) \qquad \underbrace{\hspace{10em}}_{X^{\{1\}} \quad (n_x, 1000)} \quad \underbrace{\hspace{10em}}_{X^{\{2\}} \quad (n_x, 1000)} \quad \dots \quad \underbrace{\hspace{10em}}_{X^{\{5,000\}} \quad (n_x, 1000)}
 \end{aligned}$$

$$\begin{aligned}
 Y &= \begin{bmatrix} y^{(1)} & y^{(2)} & y^{(3)} & \dots & y^{(1000)} & | & y^{(1001)} & \dots & y^{(2000)} & | & \dots & | & \dots & y^{(m)} \end{bmatrix} \\
 &\quad (1, m) \qquad \underbrace{\hspace{10em}}_{Y^{\{1\}} \quad (1, 1000)} \quad \underbrace{\hspace{10em}}_{Y^{\{2\}} \quad (1, 1000)} \quad \dots \quad \underbrace{\hspace{10em}}_{Y^{\{5,000\}} \quad (1, 1000)}
 \end{aligned}$$

What if $m = 5,000,000$?

5,000 mini-batches of 1,000 each

Mini-batch t : $X^{\{t\}}, Y^{\{t\}}$

$$\begin{array}{l}
 x^{(i)} \\
 z^{[l]} \\
 X^{\{t\}}, Y^{\{t\}}
 \end{array}$$

Andrew Ng

- $x\{t\}$ $y\{t\}$ 는 미니배치를 의미한다

- $x\{1\}$ 의 shape: $(n_x, 1000)$
- $x\{2\}$ 의 shape: $(n_x, 1000)$
- ...

Mini-batch gradient descent

Mini-batch gradient descent

repeat $\{$
for $t = 1, \dots, 5000 \}$

Forward prop on $X^{(t)}$.

$$Z^{(t)} = W^{(t)} X^{(t)} + b^{(t)}$$

$$A^{(t)} = g^{(t)}(Z^{(t)})$$

$$\vdots$$

$$A^{(t)} = g^{(t)}(Z^{(t)})$$

Vectorized implementation
(1000 examples)

for $X^{(t)}, Y^{(t)}$

Compute cost $J^{(t)} = \frac{1}{1000} \sum_{i=1}^n L(\hat{y}^{(i)}, y^{(i)}) + \frac{\lambda}{2 \cdot 1000} \sum \|W^{(t)}\|_F^2$

Backprop to compute gradients w.r.t $J^{(t)}$ (using $(X^{(t)}, Y^{(t)})$)

$$W := W^{(t)} - \alpha dW^{(t)}, \quad b := b^{(t)} - \alpha db^{(t)}$$

$\}$

"1 epoch"

pass through training set.

1 step of gradient descent
using $X^{(t)}, Y^{(t)}$.
(as if $m=1000$)

X, Y

Andrew Ng

L has been used previously to indicate the number of the layer in the network, in this particular example was also used to indicate the number of samples in the mini-batch.

- 1 epoch - "means pass through training set"
- mini-batch > batch: much faster!
-