Training with gradient Descent — Illustration using linear regression — Regular method

Background

Recall
$$\hat{y}^d = w_0 + w_1 x_1^d + \ldots + w_K x_K^d$$
$$= \sum_{j=0}^{K} w_j x_K^d \quad \text{where } x_K^d = 1 \text{ for all } d.$$

$$\frac{\partial E(\vec{w}, D)}{\partial w_i} = \frac{\partial}{\partial w_i} \frac{1}{2} \sum_d (y^d - \hat{y}^d)^2$$
$$= \frac{1}{2} \sum_d \frac{\partial}{\partial w_i} (y^d - \hat{y}^d)^2$$
$$= \frac{1}{2} \sum_d 2 . (y^d - \hat{y}^d) \cdot \frac{\partial}{\partial w_i} (y^d - \hat{y}^d)$$
$$= \sum_d (y^d - \hat{y}^d) \frac{\partial (y^d - (w_0 x_0^d + w_1 x_1^d + \ldots w_i x_i^d + \ldots w_K x_K^d))}{\partial w_i}$$
$$= \sum_d (y^d - \hat{y}^d)(-x_i^d).$$

$$\Delta w_i = -\alpha \frac{\partial E(\vec{w}, D)}{\partial w_i}$$
$$= \alpha . \sum (y^d - \hat{y}^d) x_i^d$$

Training Algorithm

- Initialize each w_i for $i \in \{1, \ldots, k\}$.
- Repeat
  - Initialize Δw_i to zero for all i
  - For each $d \in D$
    - Use current weights to compute y^d $\left(= \sum_{j=0}^{k} w_j x_j^d\right)$
    - For each i

1 $$\Delta w_i \leftarrow \Delta w_i + \alpha(y^d - \hat{y}^d) \cdot x_i^d$$

  - For each i

2 $$w_i \leftarrow w_i + \Delta w_i$$

Until convergence.

Training {Stochastic method.

- Initialize w_i for all i
- Repeat

  For each $d \in D$

  - Use current weights to compute y^d

  - For each i

① $$\Delta w_i \leftarrow \alpha (y^d - \hat{y}^d) x_i^d$$

② $$w_i \leftarrow w_i + \Delta w_i$$

Until Convergence.

Notice position of steps ① & ② in the two versions

One Epoch of training (Regular)

Epoch – run once through all of D

- Assume y^d is given by $y^d = 4x_1^d - 2x_2^d + 3$

  This is our target with $w_0 = 3$ $w_1 = 4$ $w_2 = -2$

- Assume D contains two points d_1 and d_2

  where $d_1 = (2, 1)$ or $(1, 2, 1)$ after setting x_0' as 1.

  & $d_2 = (1, 2)$ or $(1, 1, 2)$.

- Note y^{d_1} (the target value) is

  $y^{d_1} = 4 \cdot 2 - 2 \cdot 1 + 3 = 8 - 2 + 3 = 9.$

and

$y^{d_2} = 4 \cdot 1 - 2 \cdot 2 + 3 = 4 - 4 + 3 = 3$

Assume our current model is determined by $w_0 = 1$ $w_1 = 1$ $w_2 = 1$

- Then $\hat{y}^{d_1} = 1.2 + 1.1 + 1 = 2+1+1 = 4$

$\hat{y}^{d_2} = 1.1 + 1.2 + 1 = 1+2+1 = 4.$

So error on $d_1 = \left(y^{d_1} - \hat{y}^{d_1}\right) = 9-4 = 5$

error $d_2 = \left(y^{d_2} - \hat{y}^{d_2}\right) = 3-4 = -1$

- So for our current model given by $\vec{w} = \langle 1, 1, 1 \rangle$

$$E(\vec{w}, D) = \frac{1}{2} \sum_{j=1}^{2} \left(y^{d_j} - y^{d_j}\right)^2$$

$$= \frac{1}{2}\left((5)^2 + (-1)^2 \right)$$

$$= \frac{1}{2}(25+1)$$

$$= 13.$$

$\Delta w_1 = 0 = \Delta w_2 = \Delta w_0$ Let $\alpha = 0.1$

Computation at (1) for d_1

$$\Delta w_1 \leftarrow 0 + (0.1) \cdot 5 \cdot 2 = 1$$
$$\Delta w_2 \leftarrow 0 + (0.1) \cdot 5 \cdot 1 = 0.5$$
$$\Delta w_0 \leftarrow 0 + (0.1) \cdot 5 \cdot 1 = 0.5$$

Computation at (1) for d_2

$$\Delta w_1 = 1 + (0.1) \cdot (-1) \cdot (1) = 0.9$$
$$\Delta w_2 = 0.5 + (0.1) \cdot (-1) \cdot (2) = \cancel{0.5}\ 0.3$$
$$\Delta w_0 = 0.5 + (0.1)(-1) \cdot (1) = 0.4$$

At step 2

$$w_1 \leftarrow 1 + 0.9 = 1.9$$
$$w_2 \leftarrow 1 + 0.3 = 1.3$$
$$w_0 \leftarrow 1 + 0.4 = 1.4$$

So ~~new~~ weights of the model have changed to $w_0 = 1.4$ $w_1 = 1.9$ & $w_2 = 1.3$

Lets compute $E(\vec{w}, D)$ now.

~~error on~~ $\hat{y}^{d_1} = (1.4)(1) + (1.9)(2) + (1.3)(1)$

$= 1.4 + 3.8 + 1.3$

$= 6.5$.

$\hat{y}^{d_2} = (1.4)(1) + (1.9)(1) + (1.3)(2)$

$= 1.4 + 1.9 + 2.6$

$= 5.9$.

$$E(\vec{w}, D) = \frac{1}{2}\left((9 - 6.5)^2 + (3 - 5.9)^2 \right)$$

$$= \frac{1}{2}\left((2.5)^2 + (-2.9)^2 \right)$$

$$= \frac{1}{2}\left(6.25 + 8.41 \right)$$

$$= \frac{1}{2}(14.66)$$

$$= 7.33$$

Recall previously it was 13