Classification & Regression

2020年9月30日 8:33

Apply SGD to linear regression.

$$L(W, X, Y) = \frac{1}{2N} \sum_i (w^T x_i - y_i)^2$$

$$W_{t+1} = W_t - \frac{\eta}{N} \sum_i (W_t^T x_i - y_i) x_i$$

If not regression, $\Rightarrow$ classification

one way to do: $f(x) = \text{sign}(w^T x)$

$\Rightarrow$ simple perceptron $\begin{cases} y = \{-1, 1\} \\ \text{if } w^T x \cdot y < 0 \Rightarrow w = w + x \cdot y \end{cases}$

Convergence proof

Assume $\begin{cases} \exists\, w^*, \ \|w^*\| = 1 \\ \exists\, \gamma > 0. \ \ y_i \langle w^*, x_i \rangle \geq \gamma \quad \forall i. \\ \forall i, \ \ \|x_i\| \leq R \end{cases}$

$\Rightarrow$ the algorithm makes at most $\frac{R^2}{\gamma^2}$ mistakes.

Proof. start from $w_0 = 0$.

when make a mistake at t. $W_{t+1} = W_t + y_t x_t$

$$\Rightarrow \langle W_{t+1}, W^* \rangle = \langle W_t, W^* \rangle + \langle y_t x_t, W^* \rangle \geq \langle W_t, W^* \rangle + \gamma$$

$$\Rightarrow \|W_{t+1}\| = \|W_{t+1}\| \cdot \|W^*\| \geq \langle W_{t+1}, W^* \rangle \geq t\gamma$$

On the other hand,

$$\|W_{t+1}\|^2 = \|W_t + y_t x_t\|^2 = \|W_t\|^2 + \|y_t x_t\|^2 + 2\langle y_t x_t, W_t \rangle$$

(change only if $\langle y_t x_t, W_t \rangle < 0$)

$$\Rightarrow \|W_{t+1}\| \leq \|W_t\|^2 + R^2$$

$$\Rightarrow t^2 \gamma^2 \leq \|W_{t+1}\|^2 \leq t R^2$$

$$\Rightarrow t \leq \frac{R^2}{\gamma^2}$$

Logistic Regression

$f(x) \Rightarrow$ prob. in class A.

Logistic Regression

$f(x) \Rightarrow$ prob. in class A.

$\downarrow$

logistic: $f(x) = \frac{1}{1+e^{-w^T x}}$

Cross entropy.

- compute loss between two distribution
- Entropy: $H(X) = -\sum_i p_i \log p_i$
- Cross Entropy: $XE(y, p) = -\sum_i y_i \log p_i$ $\longrightarrow$ scale of gradient automatically fixed

$\geq H(y)$

$\downarrow$

KL divergence Remark: XE is asymmetric!

$= XE(y,p) - H(y)$

Linear Regression & Classification can learn everything once the features are correct

$\hookrightarrow$ Deep Learning learns features and the last step is always regression/ classification