ECE544
9/19/2013

TODAY : MSE

(1) LINEAR MSE

t_k^n = TARGET FOR k^{TH} CLASS, n^{TH} TOKEN
$\in \{-1, 1\}$, $1 \le k \le C$, $1 \le n \le N$

T $[N \times C]$ = MATRIX OF TARGETS

$\vec{t}^n$ $[C \times 1]$ = TARGETS FOR n^{TH} TOKEN
$\vec{t}_k$ $[N \times 1]$ = TARGETS FOR k^{TH} CLASS
(ALL N TOKENS)

$$T = [\vec{t}_1, \dots, \vec{t}_C] = \begin{bmatrix} \vec{t}^{1T} \\ \vdots \\ \vec{t}^{NT} \end{bmatrix}$$

ϕ_j^n = j^{TH} FEATURE OF n^{TH} TOKEN

$$\Phi\ [N \times M] = [\vec{\phi}_1, \dots, \vec{\phi}_M] = \begin{bmatrix} \vec{\phi}^{1T} \\ \vdots \\ \vec{\phi}^{NT} \end{bmatrix}$$

$\vec{\phi}^n$ $[M \times 1]$ = ALL M FEATURES, FOR n^{TH} TOKEN

$\vec{\phi}_j$ $[N \times 1]$ = j^{TH} FEATURE FOR ALL N TOKENS

w_j^k = WEIGHT ASSOCIATING k^{th} CLASS TO j^{th} FEATURE

$$W \; [C \times M] = [\vec{w}^1, \ldots, \vec{w}^M] = \begin{bmatrix} \vec{w}_1^T \\ \vdots \\ \vec{w}_C^T \end{bmatrix}$$

$\vec{w}^k$ $[M \times 1]$ = WEIGHT VECTOR FOR k^{th} CLASS

$$y_k(\vec{\Phi}^n) = \vec{w}_k^T \vec{\Phi}^n$$

IF YOU WANT A CONSTANT OFFSET, YOU NEED TO CHOOSE $\Phi_1^n = 1$ OR EQUIVALENT

$\vec{w}_j$ = WEIGHTS FOR ALL CLASSES ASSOCIATED WITH j^{th} FEATURE

LINEAR MSE

$$\mathcal{E} = \sum_{n=1}^{N} \sum_{k=1}^{C} \left| t_k^n - y_k(\vec{\Phi}^n) \right|^2$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{C} \left| t_k^n - \vec{\Phi}^{nT} \vec{w}^k \right|^2$$

$$= \sum_{n=1}^{N} \left\| \vec{t}^{nT} - \vec{\Phi}^{nT} W \right\|_2^2$$

$$= \| T - \Phi W \|_2^2 = \sum_{k=1}^{C} \| \vec{t}_k - \Phi \vec{w}^k \|^2$$

$\|A\|_F^2$ = SQUARED FROBENIUS NORM $= \sum_i \sum_j |a_{ij}|^2$

$$\mathcal{E} = \sum_{k=1}^{C} \|\vec{t}_k - \Phi \vec{w}_k\|_2^2$$

$$\frac{\partial \mathcal{E}}{\partial w_m^k} = \frac{\partial}{\partial w_m^c}\left(\sum_{n=1}^{N}\sum_{k=1}^{C}\left|t_k^n - \sum_{j=1}^{M} w_j^k \phi_j^n\right|^2\right)$$

$$= -\sum_{n=1}^{N}\left(t_c^n - \sum_{j=1}^{M} w_j^c \phi_j^n\right)\phi_c^n = 0$$

$$\sum_{n=1}^{N} t_c^n \phi_c^n = \sum_{n=1}^{N}\sum_{j=1}^{M} w_j^c \phi_j^n \phi_c^n$$

$$T^T \Phi = W \Phi^T \Phi$$

$$T^T \Phi (\Phi^T \Phi)^{-1} = W \qquad \boxed{\begin{aligned} W^T &= (\Phi^T \Phi)^{-1} \Phi^T T \\ &= \Phi^{+} T \end{aligned}}$$

TRY AGAIN

$$\mathcal{E} = \sum_{k=1}^{C} \|\vec{t}_k - \Phi \vec{w}_k\|_2^2$$

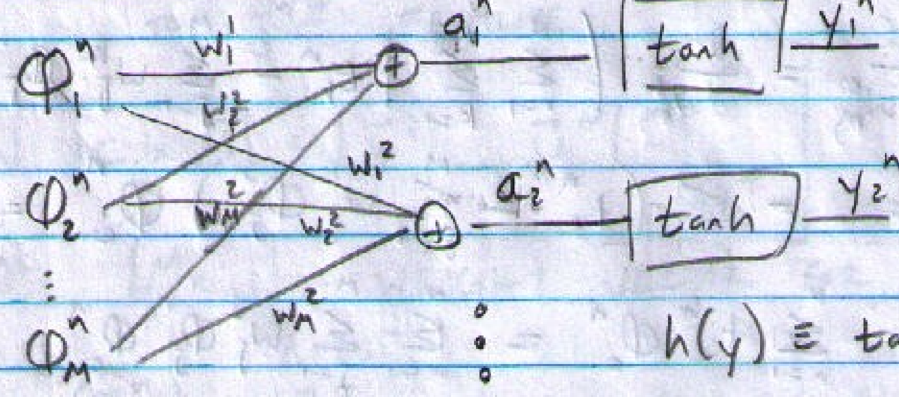
$$\underbrace{\nabla_{\vec{w}_k} \mathcal{E}}_{[M \times 1]} = -2 \underbrace{\Phi^T}_{[M \times N]} \underbrace{(\vec{t}_k - \Phi \vec{w}_k)}_{[N \times 1]} = \vec{0}$$

$$\Phi^T \vec{t}_k = \Phi^T \Phi \vec{w}_k$$

$$\boxed{\vec{w}_k = (\Phi^T \Phi)^{-1} \Phi^T \vec{t}_k}$$

SAME ANSWER!

II. NONLINEAR MSE



$h(y) \equiv \tanh(y)$

$$\mathcal{E} = \sum_{n=1}^{N} \sum_{k=1}^{C} \left| t_k^n - y_k^n \right|^2$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{C} \left| t_k^n - h\left(\sum_{j=1}^{M} \Phi_j^n w_j^k \right) \right|^2$$

$$= \sum_{n=1}^{N} \sum_{k=1}^{C} \left| t_k^n - h\left(\vec{\Phi}^{nT} \vec{w}^k \right) \right|^2$$

$$\frac{\partial \mathcal{E}}{\partial w_m^c} = -\sum_{n=1}^{N} \left(t_c^n - h\left(\sum_{j=1}^{M} \Phi_j^n w_j^c \right) \right)$$

$$\times \frac{\partial}{\partial w_m^c} h\left(\sum_{j=1}^{M} \Phi_j^n w_j^c \right)$$

$$= -\sum_{n=1}^{N} \left(t_c^n - y_c^n \right) \cdot h'(a_c^n) \cdot \frac{\partial a_c^n}{\partial w_m^c}$$

$$= -\sum_{n=1}^{N} \left(t_c^n - y_c^n \right) \cdot h'(a_c^n) \cdot \Phi_m^n$$

REMEMBER $h'(a_c^n) = 1 - h^2(a_c^n)$

$$= 1 - (y_c^n)^2$$

GRADIENT DESCENT

① START WITH RANDOM W

② FOR SMALL LEARNING RATE η,

$$W_m^c = W_m^c - \eta \frac{\partial E}{\partial w_m^c}$$

$$= w_m^c + \eta \sum_{n=1}^{N} \Phi_m^n \underbrace{(t_c^n - y_c^n)\, h'(a_c^n)}_{\text{CALL THIS } \delta_c^n}$$

Δ $[N \times C]$

Φ $[N \times M]$

SO $W = W + \eta \Delta^T \Phi$

PROPAGATION BACK TO m^{TH} COMPONENT OF INPUT FEATURES

GRADIENT OF ANN OUTPUTS