Machine Learning - Assignment 1

Νικόλαος Σμυρνιούδης - 3170148

1 Μαθηματικοί τύπου για τις παραγώγους του $W^{(1)}$

Η συνάρτηση κόστους είναι:

$$W^{(1)} \in \mathrm{R}^{Mx(D+1)}$$

$$E(w) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log(y_{nk}) - \frac{\lambda}{2} \|\mathbf{w}\|^2$$

Επίσης ο τύπος για υπολογισμό πιθανοτήτων είναι:

$$y_{nk} = \frac{e^{w_k^{(2)T} z_n}}{\sum_{j=1}^{K} e^{w_j^{(2)T} z_n}}$$

Και οι έξοδοι του πρώτου επιπέδου:

$$z_{nj} = h(w_j^{(2)T} x_n)$$

Για ευκολία έστω:

$$E'(w) = \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log(y_{nk})$$

και

$$R'(w) = -\frac{\lambda}{2} \|\mathbf{w}\|^2$$

Θα ισχύει

$$\frac{\partial E}{\partial W^{(1)}} = \frac{\partial E'}{\partial W^{(1)}} + \frac{\partial R'}{\partial W^{(1)}}$$

$$\frac{\partial - \frac{\lambda}{2} \|\mathbf{w}\|^2}{\partial w_{md}^{(1)}} = \frac{\partial - \frac{\lambda}{2} \sum_{i=1}^{M} \sum_{j=1}^{D+1} w_{ij}^2}{\partial w_{md}^{(1)}} = -\lambda w_{md}^{(1)}$$

Αρα για τον τελεστη κανονικοποίησης:

$$\frac{\partial R'}{\partial W^{(1)}} = -\lambda * W^{(1)}$$

Όσον αφορά το E' θα βρούμε την παράγωγο του του E' ως προς τις μεταβλητές z_{ij} που είναι οι έξοδοι του πρώτου επιπέδου. Έστω ένα z_i διάνυσμα που είναι η i-οστή γραμμή του πίνακα Z.

$$\frac{\partial E'}{\partial z_i} = \frac{\partial \sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log(y_{nk})}{\partial z_i} = \frac{\partial \sum_{n=1}^{N} t_{nk(n)} \log(y_{nk(n)})}{\partial z_i} \tag{1}$$

$$= \frac{\partial t_{ik(n)} \log(y_{ik(n)})}{\partial z_i} = \frac{t_{ik(n)}}{y_{ik(n)}} \frac{\partial y_{ik(n)}}{\partial z_i} \tag{2}$$

Στην παραπάνω ισότητα θεωρούμε πως $k(n)$ αντιπροσωπεύει τον αριθμό της σωστής κλάσσης για το παράδειγμα εκπαίδευσης n. Το άθροισμα συνεπώς για $k \neq k(n)$ είναι 0 αφού για αυτά τα k ισχύει $t_{nk} = 0$

Μελετώντας τις παραγώγους των πιθανοτήτων εξόδου ως πρός τα z

$$\frac{\partial y_{nk}}{\partial z_n} = \frac{w_k^{(2)} e^{w_k^{(2)T} z_n} \sum_{j=1}^{K} e^{w_j^{(2)T} z_n} - (\sum_{j=1}^{K} e^{w_j^{(2)T} z_n} w_j^{(2)}) e^{w_k^{(2)T} z_n}}{(\sum_{j=1}^{K} e^{w_j^{(2)T} z_n})^2} \tag{3}$$

$$= w_k^{(2)} y_{nk} - (\sum_{j'=1}^{K} \frac{e^{w_{j'}^{(2)T} z_n} w_j^{(2)}}{\sum_{j=1}^{K} e^{w_j^{(2)T} z_n}}) y_{nk} \tag{4}$$

$$= y_{nk}(w_k^{(2)} - \sum_{j'=1}^{K} w_j^{(2)} y_{nj}) \tag{5}$$

$$= y_{nk}(W^{(2)T} t_n - W^{(2)T} y_n) \tag{6}$$

$$= y_{nk} W^{(2)T} (t_n - y_n) \tag{7}$$

Συνδυάζοντας την (7) με την (2) έχουμε:

$$\frac{\partial E'}{\partial z_i} = W^{(2)T} (t_i - y_i) \tag{8}$$

Θα χρησιμοποιήσουμε $[\cdot]_j$ για συμβολίζουμε το j-οστό στοιχείο ενός διανύσματος. Χρησιμοποιοώντας τον κανόνα αλυσίδας:

$$\frac{\partial E'}{\partial w_j^{(1)}} = \sum_{n=1}^{N} \sum_{j'=0}^{M} \frac{\partial E'}{\partial z_{nj'}} \frac{\partial z_{nj'}}{\partial w_j^{(1)}} \tag{9}$$

Ωστόσο για $j' \neq j$ το οποίο ισχύει επίσης για $j' = 0$, δηλαδή για τα biases του δεύτερου επιπέδου, ισχύει

$$\frac{\partial z_{nj'}}{\partial w_j^{(1)}} = 0$$

Γιατι το $z_{nj'}$ δεν εξαρτάται από το $w_j^{(1)}$. Συνεπώς:

$$\frac{\partial E'}{\partial w_j^{(1)}} = \sum_{n=1}^{N} \frac{\partial E'}{\partial z_{nj}} \frac{\partial z_{nj}}{\partial w_j^{(1)}} \tag{10}$$

Χρησιμοποιώντας τον τύπο για τα z:

$$\frac{\partial z_{nj'}}{\partial w_j^{(1)}} = \frac{\partial h(w_j^{(1)T} x_n)}{\partial w_j^{(1)}} = h'(w_j^{(1)T} x_n) x_n \tag{11}$$

Συνδιάζοντας τις (11) και (10):

$$\begin{aligned} \frac{\partial E'}{\partial w_j^{(1)}} &= \sum_{n=1}^{N} \frac{\partial E'}{\partial z_{nj}} \frac{\partial z_{nj}}{\partial w_j^{(1)}} && (12) \\ &= \sum_{n=1}^{N} [W^{(2)T}(t_n - y_n)]_j h'(w_j^{(1)T} x_n) x_n && (13) \end{aligned}$$

Έστω

$$\begin{aligned} c_{nj} &= [W^{(2)T}(t_n - y_n)]_j \\ a_{nj} &= h'(w_j^{(1)T} x_n) \end{aligned}$$

και οι πίνακες που τα περιέχουν αντίστοιχα, $A \in \mathbf{R}^{Nb \times M}$ και $C \in \mathbf{R}^{Nb \times M}$.

Απο τα στοιχεία c, δεν χρησιμοποιούμε ποτέ τα στοιχεία με δεύτερο δείκτη 0 (υπάρχουν μόνο M γραμμές του πίνακα $W^{(1)}$). Οπότε για τα c_{nj} δεν χρησιμοποιείται η στήλη 0 του πίνακα $W^{(2)}$. Άρα στον τύπο του c μπορούμε αντικαταστήσουμε το $W^{(2)} \in \mathbf{R}^{K \times (M+1)}$ με $W'^{(2)} \in \mathbf{R}^{K \times M}$ που δεν έχει την στήλη 0.

Για τον A ισχύει πως το στοιχείο n, j θα είναι το εσωτερικό γινόμενο της n γραμμής του πίνακα X με την j γραμμή του πίνακα $W^(2)$, στο οποίο εφαρμόζεται η παράγωγος της activation h. Άρα

$$A = h'(XW^{(1)T}) \tag{14}$$

Επίσης για τον C μπορούμε να κάνουμε τις εξής απλοποιήσεις:

$$\begin{aligned} C &= \left[\begin{pmatrix} W'^{(2)T}(t_1 - y_1) \end{pmatrix} \quad \dots \quad \begin{pmatrix} W'^{(2)T}(t_N - y_N) \end{pmatrix} \right]^T \\ &= (W'^{(2)T} \left[\begin{pmatrix} t_1 - y_1 \end{pmatrix} \quad \dots \quad \begin{pmatrix} t_N - y_N \end{pmatrix} \right])^T \\ &= (W'^{(2)T}(T - Y)^T)^T \\ &= (T - Y)W'^{(2)} \end{aligned}$$

Πίνακας 1: Τα εύρη των υπερπαραμέτρων που δοκιμάστηκαν

Υπερπαράμετρος	Εύρος τιμών
η	$[10^{-2}\ 10^{-3}\ 10^{-4}]$
λ	$[10^{-4}\ 10^{-5}\ 10^{-6}]$
M	[100 200 300]
activation	$[logexp(x)\ tanh(x)\ cos(x)]$

Τώρα μπορούμε να γράψουμε την (13):

$$\frac{\partial E'}{\partial w_j^{(1)}} = \sum_{n=1}^{N} c_{nj} a_{nj} x_n \tag{15}$$

$$= X^T \begin{bmatrix} c_{1j} a_{1j} \\ \vdots \\ c_{Nj} a_{Nj} \end{bmatrix} \tag{16}$$

Τέλος συνδυάζοντας τις παραγώγους της κάθε γραμμής του $W^{(1)}$

$$\begin{aligned}
\frac{\partial E'}{\partial W^{(1)}} &= \left[X^T \begin{pmatrix} c_{11} a_{11} \\ \vdots \\ c_{N1} a_{N1} \end{pmatrix} \quad \dots \quad X^T \begin{pmatrix} c_{1M} a_{1M} \\ \vdots \\ c_{NM} a_{NM} \end{pmatrix} \right]^T \\
&= X^T \left[\begin{pmatrix} c_{11} a_{11} \\ \vdots \\ c_{N1} a_{N1} \end{pmatrix} \quad \dots \quad \begin{pmatrix} c_{1M} a_{1M} \\ \vdots \\ c_{NM} a_{NM} \end{pmatrix} \right]^T \\
&= (X^T (C \odot A))^T \\
&= (C^T \odot A^T) X \\
&= (W'^{(2)T} (T - Y)^T \odot h'(W^{(1)} X^T)) X
\end{aligned}$$

2 Αποτελέσματα

Οι παράγωγοι για τους δύο πίνακες δοκιμάστηκαν επιτυχώς στην συνάρτηση gradcheck με $\epsilon = 10^{-6}$ και tolerance $= 10^{-4}$ Τα δύο dataset εκπαίδευσης τών MNIST και CIFAR-10 έγιναν split σε train και dev

Το η είναι η παράμετρος μάθησης (learning rate), το λ είναι η παράμετρος κανονικοποίησης, το M είναι το μέγεθος του κρυφού επιπέδου και activation είναι η συνάρτηση ενεργοποίησης. Η κάθε δοκιμή υπερπαραμέτρων έγινε με batch size ίσο με 200 και για 50 εποχές.

Πίνακας 2: Τα εύρη των υπερπαραμέτρων που δοκιμάστηκαν

Υπερπαράμετρος	Εύρος τιμών
η	$[10^{-2}\ 10^{-3}\ 10^{-4}]$
λ	$[10^{-4}\ 10^{-5}\ 10^{-6}]$
M	[100 200 300]
activation	$[logexp(x)\ tanh(x)\ cos(x)]$

Πίνακας 3: Τα ποτελέσματα των καλύτερων υπερπαραμέτρων στα δεδομένα ελέγχου τών MNIST και CIFAR-10

ΜΝΙΣΤ	῝ΙΦΑΡ-10
2.18%	50%

Για το MNIST το καλύτερο σετ υπερπαραμέτρων ήταν

$$h = 10^{-3}$$
$$l = 10^{-6}$$
$$M = 300$$
$$\text{activation} = cos(x)$$

Και πέτυχε σφάλμα 2.22 % στα δεδομένα *dev*. Για το CIFAR-10 το καλύτερο σετ υπερπαραμέτρων ήταν

$$h = 10^{-4}$$
$$l = 10^{-5}$$
$$M = 300$$
$$\text{activation} = cos(x)$$

Και πέτυχε σφάλμα 50.5 % στα δεδομένα *dev*. Τέλος, δοκιμάστηκαν οι καλύτερες υπερπαράμετροι για το εκάστοτε μοντέλο στα δεδομένα ελέγχου.

Πίνακας 4: Το σφάλμα του μοντέλου με τις καλύτερες υπερπαραμέτρους στα δεδομένα ελέγχου τών MNIST και CIFAR-10

MNIST	CIFAR-10
2.18 %	50.40 %