Μηχανική Μάθηση

Πρώτη Προγραμματιστική Άσκηση

Ονοματεπώνυμο	**ΑΜ**
Βανακάρης Άγγελος	3150006

Contents

Εισαγωγή

Στην εργασία αυτή υλοποιήθηκε ένα νευρωνικό δίκτυο με ένα κρυφό επίπεδο και δυναμικό πλήθος νευρώνων σε κάθε επίπεδο. Κάθε επίπεδο συνδέεται πλήρως με το επόμενό του, με κάθε σύνδεση να έχει το δικό της βάρος. Επίσης, το επίπεδο εισόδου και το κρυφό επίπεδο έχουν από ένα bias το οποίο αρχικοποιείται με την τιμή 1. Η υλοποίηση έγινε με τη βοήθεια των πινάκων της βιβλιοθήκης numpy, επομένως όλες οι διαδικασίες έχουν αναχθεί σε πράξεις μεταξύ πινάκων.
Οι βασικές διαδικασίες που χρησιμοποιούνται είναι η πρόβλεψη μίας κατηγορίας δοθέντος κάποιου παραδείγματος εισόδου *(feed forward)* και η εκπαίδευση του μοντέλου δοθέντων κάποιων παραδειγμάτων εκπαίδευσης *(με τη χρήση της διαδικασίας back propagation)*.
Σαν παραδείγματα εφαρμογής του μοντέλου χρησιμοποιήθηκαν το σύνολο δεδομένων MNIST για κατηγοριοποίηση χειρόγραφων αριθμών και το σύνολο δεδομένων CIFAR-10 για κατηγοριοποίηση φωτογραφιών με διάφορα αντικείμενα και ζώα.

Προετοιμασία Δεδομένων

MNIST

Αρχικά, γίνεται το διάβασμα των δεδομένων από τα αρχεία τύπου text του dataset, τα οποία είναι ομαδοποιημένα ως εξής:
Κάθε ομάδα παραδειγμάτων εκπαίδευσης βρίσκεται σε ένα αρχείο με όνομα *train%d.txt* με το *%d* να παίρνει την τιμή της κατηγορίας στην οποία ανήκουν τα παραδείγματα που εμπεριέχονται σε αυτό. Αντίστοιχα, οι ομάδες των παραδειγμάτων αξιολόγησης βρίσκονται σε αρχεία με όνομα *test%d.txt* ακολουθώντας την ίδια κωδικοποίηση. Κάθε παράδειγμα είναι μία εικόνα και έχει 784 χαρακτηριστικά. Κάθε τέτοιο χαρακτηριστικό είναι μία αριθμητική τιμή του αντίστοιχου pixel της εικόνας. Κατά το διάβασμα του dataset καταχωρείται η κατηγορία κάθε παραδείγματος σε έναν πίνακα 10 θέσεων με όλα τα στοιχεία του να είναι 0 εκτός από εκείνο που βρίσκεται στη θέση *%d* του πίνακα και το οποίο έχει την τιμή 1 *(one-hot vector)*. Μπορούμε να αποθηκεύσουμε αυτήν την πληροφορία για όλα τα παραδείγματα σε έναν πίνακα $\boldsymbol{T}$ με διαστάσεις $(\boldsymbol{N_b} \times \boldsymbol{K})$ όπου $\boldsymbol{N_b}$ το πλήθος των παραδειγμάτων.

CIFAR-10

Αρχικά, γίνεται το διάβασμα των δεδομένων από τα αρχεία του dataset, τα οποία έχουν την εξής μορφή: Τα δεδομένα εκπαίδευσης βρίσκονται στα αρχεία με όνομα *data_batch_%d.bin*. Κάθε παράδειγμα είναι μία εικόνα και έχει 3072 χαρακτηριστικά. Τα χαρακτηριστικά αυτά είναι χωρισμένα στα τρία κανάλια χρώματος (R, G, B), επομένως τα πρώτα 1024 αφορούν στο κόκκινο κανάλι χρώματος, τα επόμενα 1024 στο πράσινο και τα τελευταία 1024 στο μπλε. Κάθε ένα από αυτά τα χαρακτηριστικά είναι μία αριθμητική τιμή του αντίστοιχου pixel της εικόνας για το αντίστοιχο κανάλι χρώματος. Εκτός των 3072 χαρακτηριστικών του, κάθε παράδειγμα περιέχει και την κατηγορία του, η οποία περιγράφεται από ένα διάνυσμα one-hot *(όπως συμβαίνει και στο σύνολο δεδομένων MNIST)*. Μπορούμε να αποθηκεύσουμε αυτήν την πληροφορία για όλα τα παραδείγματα σε έναν πίνακα $\boldsymbol{T}$ με διαστάσεις $(\boldsymbol{N_b} \times \boldsymbol{K})$ όπου $\boldsymbol{N_b}$ το πλήθος των παραδειγμάτων.

Πρόβλεψη

Το επίπεδο εισόδου του νευρωνικού δικτύου έχει $(\boldsymbol{D}+\mathbf{1})$ νευρώνες, όπου $\boldsymbol{D}$ η διάσταση ενός παραδείγματος εκπαίδευσης (για το σύνολο MNIST έχουμε $\boldsymbol{D}=\mathbf{784}$, άρα $\boldsymbol{D}+\mathbf{1}=\mathbf{785}$). Επομένως, κάθε χαρακτηριστικό κάθε παραδείγματος θα αντιστοιχεί σε έναν νευρώνα και θα υπάρχει ένας επιπλέον νευρώνας με τιμή 1 που αναπαριστά το bias του επιπέδου.

Το κρυφό επίπεδο του νευρωνικού δικτύου έχει $(\boldsymbol{M}+\mathbf{1})$ νευρώνες, όπου $\boldsymbol{M}$ μία μεταβλητή που ορίζει ο χρήστης. Κάθε νευρώνας του κρυφού επιπέδου συνδέεται με όλους τους νευρώνες του επιπέδου εισόδου και κάθε σύνδεση έχει το δικό της βάρος. Ο ένας επιπλέον νευρώνας του κρυφού επιπέδου έχει τιμή 1 και αναπαριστά το bias του επιπέδου *(ο νευρώνας αυτός δε συνδέεται με τους νευρώνες του επιπέδου εισόδου)*.

Οι $\boldsymbol{M}$ νευρώνες θα παίρνουν σαν είσοδο την τιμή:

$$\sum_{\boldsymbol{d}=\mathbf{1}}^{\boldsymbol{D}+\mathbf{1}} \boldsymbol{x_d w_d^{(1)}}$$

όπου $\boldsymbol{x_d}$ είναι το *d*-οστό χαρακτηριστικό του παραδείγματος και $\boldsymbol{w_d^{(1)}}$ είναι το βάρος της σύνδεσης που συνδέει το νευρώνα του κρυφού επιπέδου με τον *d*-οστό νευρώνα του επιπέδου εισόδου. Επομένως, κάθε νευρώνας του κρυφού επιπέδου έχει $(\boldsymbol{D}+\mathbf{1})$ συνδέσεις που φτάνουν σε αυτόν. Μας ενδιαφέρουν τα βάρη των συνδέσεων, επομένως μπορούμε να κρατήσουμε σε ένα διάνυσμα $\boldsymbol{w_m^{(1)}}$ τα βάρη των συνδέσεων που φτάνουν στον *m*-οστό νευρώνα του κρυφού επιπέδου. Στη συνέχεια, μπορούμε να ορίσουμε έναν πίνακα $\boldsymbol{W^{(1)}}$ με διαστάσεις $\boldsymbol{M}\times(\boldsymbol{D}+\mathbf{1})$, ο οποίος θα αποθηκεύει τα βάρη όλων των συνδέσεων του επιπέδου εισόδου με το κρυφό επίπεδο. Κάθε γραμμή του $\boldsymbol{W^{(1)}}$ θα είναι το αντίστοιχο διάνυσμα $\boldsymbol{w_m^{(1)}}$ που περιγράφηκε παραπάνω.

Για τις διαδικασίες της πρόβλεψης και της εκπαίδευσης χρησιμοποιούνται batches από παραδείγματα, επομένως μπορούμε και αυτά να τα αποθηκεύσουμε σε μορφή πίνακα.
Έστω ότι έχουμε $\boldsymbol{N_b}$ παραδείγματα εισόδου.
Κάθε παράδειγμα έχει $(\boldsymbol{D}+\mathbf{1})$ χαρακτηριστικά, επομένως μπορούμε να αποθηκεύσουμε όλα τα παραδείγματα εισόδου σε έναν πίνακα $\boldsymbol{X}$ με διαστάσεις $\boldsymbol{N_b}\times(\boldsymbol{D}+\mathbf{1})$, κάθε γραμμή του οποίου θα αντιστοιχεί σε ένα παράδειγμα εισόδου.

Έχοντας αυτά ως δεδομένα, μπορούμε να χρησιμοποιήσουμε το εσωτερικό γινόμενο των πινάκων $\boldsymbol{X}$ και $\left(\boldsymbol{W^{(1)}}\right)^{\boldsymbol{T}}$ για να υπολογίσουμε τις τιμές που θα έχουν οι νευρώνες του κρυφού επιπέδου όταν δοθεί σαν είσοδο στο νευρωνικό δίκτυο ένα batch δεδομένων εισόδου. Τις τιμές αυτές μπορούμε να τις κρατάμε σε έναν πίνακα $\boldsymbol{I}$ με διαστάσεις $\boldsymbol{N_b}\times\boldsymbol{M}$.

Οπότε, για την έξοδο του επιπέδου εισόδου *(είσοδος του κρυφού επιπέδου)* θα έχουμε:

$$\boldsymbol{I}=\boldsymbol{X}\cdot\left(\boldsymbol{W^{(1)}}\right)^{\boldsymbol{T}}$$

Στον πίνακα $\boldsymbol{I}$ προσθέτουμε μία επιπλέον στήλη με άσους που αναπαριστά την τιμή του επιπλέον νευρώνα για το bias του κρυφού επιπέδου.

Αφού υπολογιστούν οι τιμές των νευρώνων του κρυφού επιπέδου, μπορούμε να αποθηκεύσουμε σε έναν άλλον πίνακα $\boldsymbol{Z}$ τις τιμές εξόδου του κρυφού επιπέδου. Οι τιμές αυτές προκύπτουν με την ενεργοποίηση των νευρώνων, επομένως είναι οι τιμές που θα επιστρέψει η συνάρτηση ενεργοποίησης $\boldsymbol{h}$ αν της δώσουμε ως είσοδο τις τιμές του πίνακα $\boldsymbol{I}$.

Επομένως, έχουμε:

$$\boldsymbol{Z = h(I)}$$

Το κρυφό επίπεδο του νευρωνικού δικτύου έχει $\boldsymbol{K}$ νευρώνες, όπου $\boldsymbol{K}$ μία μεταβλητή που ορίζει ο χρήστης και δείχνει το πλήθος των κατηγοριών που έχουμε. Κάθε νευρώνας του επιπέδου εξόδου συνδέεται με όλους τους νευρώνες του κρυφού επιπέδου και κάθε σύνδεση έχει το δικό της βάρος.

Οι $\boldsymbol{K}$ νευρώνες θα παίρνουν σαν είσοδο την τιμή:

$$\sum_{\boldsymbol{m=1}}^{\boldsymbol{M+1}} \boldsymbol{z_m w_m^{(2)}}$$

όπου $\boldsymbol{z_m}$ είναι η τιμή εξόδου του *m*-οστού νευρώνα του κρυφού επιπέδου και $\boldsymbol{w_m^{(2)}}$ είναι το βάρος της σύνδεσης που συνδέει το νευρώνα του επιπέδου εξόδου με τον *m*-οστό νευρώνα του κρυφού επιπέδου. Επομένως, κάθε νευρώνας του επιπέδου εξόδου έχει $\boldsymbol{(M+1)}$ συνδέσεις που φτάνουν σε αυτόν. Μας ενδιαφέρουν τα βάρη των συνδέσεων, επομένως μπορούμε να κρατήσουμε σε ένα διάνυσμα $\boldsymbol{w_k^{(2)}}$ τα βάρη των συνδέσεων που φτάνουν στον *k*-οστό νευρώνα του επιπέδου εξόδου. Στη συνέχεια, μπορούμε να ορίσουμε έναν πίνακα $\boldsymbol{W^{(2)}}$ με διαστάσεις $\boldsymbol{K \times (M+1)}$, ο οποίος θα αποθηκεύει τα βάρη όλων των συνδέσεων του κρυφού επιπέδου με το επίπεδο εξόδου. Κάθε γραμμή του $\boldsymbol{W^{(2)}}$ θα είναι το αντίστοιχο διάνυσμα $\boldsymbol{w_k^{(2)}}$ που περιγράφηκε παραπάνω.

Όπως και προηγουμένως, μπορούμε να χρησιμοποιήσουμε το εσωτερικό γινόμενο των πινάκων $\boldsymbol{Z}$ και $\boldsymbol{\left(W^{(2)}\right)^T}$ για να υπολογίσουμε τις τιμές που θα έχουν οι νευρώνες του επιπέδου εξόδου όταν δοθεί σαν είσοδο στο νευρωνικό δίκτυο ένα batch δεδομένων εισόδου. Τις τιμές αυτές μπορούμε να τις κρατάμε σε έναν πίνακα $\boldsymbol{O}$ με διαστάσεις $\boldsymbol{N_b \times K}$.

Οπότε, για την έξοδο του κρυφού επιπέδου *(είσοδος του επιπέδου εξόδου)* θα έχουμε:

$$\boldsymbol{O = Z \cdot \left(W^{(2)}\right)^T}$$

Αφού υπολογιστούν οι τιμές των νευρώνων του επιπέδου εξόδου, μπορούμε να αποθηκεύσουμε σε έναν άλλον πίνακα $\boldsymbol{Y}$ τις τιμές εξόδου του επιπέδου εξόδου. Οι τιμές αυτές προκύπτουν με την ενεργοποίηση των νευρώνων, επομένως είναι οι τιμές που θα επιστρέψει η συνάρτηση ενεργοποίησης $\boldsymbol{softmax}$ αν της δώσουμε ως είσοδο τις τιμές του πίνακα $\boldsymbol{O}$.

Επομένως, έχουμε:

$$\boldsymbol{Y = softmax(O)}$$

Οι τιμές του πίνακα $\boldsymbol{Y}$ είναι και η τελική έξοδος του νευρωνικού δικτύου για τα παραδείγματα που περιέχονται στο batch και του δόθηκαν ως είσοδος.

Εκπαίδευση

Για την αξιολόγηση του μοντέλου χρησιμοποιούμε την παρακάτω συνάρτηση κόστους:

$$\boldsymbol{E(w) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk} \log y_{nk} - \frac{\lambda}{2}\left(\left\|W^{(1)}\right\|^2 + \left\|W^{(2)}\right\|^2\right)},$$

όπου τα $\boldsymbol{t_{nk}}$ και $\boldsymbol{y_{nk}}$ είναι τα στοιχεία στη θέση $\boldsymbol{(n,k)}$ των πινάκων $\boldsymbol{T}$ και $\boldsymbol{Y}$ αντίστοιχα.

Κατά την εκπαίδευση του μοντέλου προσπαθούμε να μεγιστοποιήσουμε την τιμή της συνάρτησης κόστους, ενημερώνοντας τα βάρη όλων των συνδέσεων του νευρωνικού δικτύου. Για να το πετύχουμε αυτό αρκεί να μεταβάλλουμε τις τιμές των βαρών των συνδέσεων προς την κατεύθυνση εκείνη η οποία μας οδηγεί στη μέγιστη τιμή της συνάρτησης κόστους. Το διάνυσμα που μας δείχνει την κατεύθυνση αυτή υπολογίζεται με τη βοήθεια των μερικών παραγώγων της συνάρτησης κόστους ως προς το βάρος της κάθε σύνδεσης.
Επομένως, για το μοντέλο μας χρειαζόμαστε τα 2 παρακάτω διανύσματα:

- Ένα που θα αποτελείται από τις μερικές παραγώγους της συνάρτησης κόστους ως προς τα βάρη των συνδέσεων του κρυφού επιπέδου με το επίπεδο εξόδου, δηλαδή:

$$\frac{\partial E(w)}{\partial W^{(2)}} = \begin{bmatrix} \frac{\partial E(w)}{\partial w_{11}^{(2)}} & \cdots & \frac{\partial E(w)}{\partial w_{1(M+1)}^{(2)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial E(w)}{\partial w_{K1}^{(2)}} & \cdots & \frac{\partial E(w)}{\partial w_{K(M+1)}^{(2)}} \end{bmatrix} \qquad (A)$$

- Ένα που θα αποτελείται από τις μερικές παραγώγους της συνάρτησης κόστους ως προς τα βάρη των συνδέσεων του επιπέδου εισόδου με το κρυφό επίπεδο, δηλαδή:

$$\frac{\partial E(w)}{\partial W^{(1)}} = \begin{bmatrix} \frac{\partial E(w)}{\partial w_{11}^{(1)}} & \cdots & \frac{\partial E(w)}{\partial w_{1(D+1)}^{(1)}} \\ \vdots & \ddots & \vdots \\ \frac{\partial E(w)}{\partial w_{M1}^{(1)}} & \cdots & \frac{\partial E(w)}{\partial w_{M(D+1)}^{(1)}} \end{bmatrix} \qquad (B)$$

Για τον υπολογισμό του τύπου (A) θα χρησιμοποιήσουμε τον κανόνα της αλυσίδας ως εξής:

$$(A) \Rightarrow \frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{W}^{(2)}} = \frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{O}} \frac{\partial \boldsymbol{O}}{\partial \boldsymbol{W}^{(2)}} \qquad (1)$$

$$\frac{\partial \boldsymbol{O}}{\partial \boldsymbol{W}^{(2)}} = \frac{\partial \boldsymbol{Z} \cdot \left(\boldsymbol{W}^{(2)}\right)^{\boldsymbol{T}}}{\partial \boldsymbol{W}^{(2)}} = \boldsymbol{Z} \qquad (2)$$

$$\frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{O}} = \begin{bmatrix} \frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{o}_{11}} & \cdots & \frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{o}_{1K}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{o}_{N_b 1}} & \cdots & \frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{o}_{N_b K}} \end{bmatrix} \qquad (3)$$

Εξετάζοντας το πρώτο στοιχείο του παραπάνω πίνακα έχουμε:

$$\frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{o}_{11}} = \frac{\partial \left(\sum_{n=1}^{N} \sum_{k=1}^{K} \boldsymbol{t}_{nk} \log \boldsymbol{y}_{nk} - \frac{\boldsymbol{\lambda}}{\boldsymbol{2}} \left(\left\| \boldsymbol{W}^{(1)} \right\|^2 + \left\| \boldsymbol{W}^{(2)} \right\|^2 \right) \right)}{\partial \boldsymbol{o}_{11}}$$

$$= \frac{\partial \left(\sum_{n=1}^{N} \sum_{k=1}^{K} \boldsymbol{t}_{nk} \log \boldsymbol{y}_{nk} \right)}{\partial \boldsymbol{o}_{11}} - \frac{\partial \frac{\boldsymbol{\lambda}}{\boldsymbol{2}} \left(\left\| \boldsymbol{W}^{(1)} \right\|^2 + \left\| \boldsymbol{W}^{(2)} \right\|^2 \right)}{\partial \boldsymbol{o}_{11}}$$

Το δεξί κλάσμα μηδενίζεται, καθώς ο αριθμητής είναι σταθερός όρος. Στο αριστερό κλάσμα, ο όρος $\boldsymbol{o}_{11}$ *στον αριθμητή εμφανίζεται μόνο για* $\boldsymbol{n} = \boldsymbol{k} = \boldsymbol{1}$*, ενώ οι υπόλοιποι όροι του αθροίσματος είναι σταθεροί, οπότε μηδενίζονται.*

$$= \frac{\partial \boldsymbol{t}_{11} \log \boldsymbol{y}_{11}}{\partial \boldsymbol{o}_{11}} = \boldsymbol{t}_{11} \frac{\partial \log \frac{\boldsymbol{e}^{\boldsymbol{o}_{11}}}{\sum_{j=1}^{K} \boldsymbol{e}^{\boldsymbol{o}_{1j}}}}{\partial \boldsymbol{o}_{11}} = \boldsymbol{t}_{11} \frac{\partial \left(\log \boldsymbol{e}^{\boldsymbol{o}_{11}} - \log \sum_{j=1}^{K} \boldsymbol{e}^{\boldsymbol{o}_{1j}} \right)}{\partial \boldsymbol{o}_{11}}$$

$$= \boldsymbol{t}_{11} \left(\frac{\partial (\log \boldsymbol{e}^{\boldsymbol{o}_{11}})}{\partial \boldsymbol{o}_{11}} - \frac{\partial \log \sum_{j=1}^{K} \boldsymbol{e}^{\boldsymbol{o}_{1j}}}{\partial \boldsymbol{o}_{11}} \right) = \boldsymbol{t}_{11} \left(\frac{\partial (\boldsymbol{o}_{11} \log \boldsymbol{e})}{\partial \boldsymbol{o}_{11}} - \frac{\partial \log \sum_{j=1}^{K} \boldsymbol{e}^{\boldsymbol{o}_{1j}}}{\partial \boldsymbol{o}_{11}} \right)$$

$$= \boldsymbol{t}_{11} \left(\log \boldsymbol{e} - \frac{\boldsymbol{1}}{\sum_{j=1}^{K} \boldsymbol{e}^{\boldsymbol{o}_{1j}}} \frac{\partial \sum_{j=1}^{K} \boldsymbol{e}^{\boldsymbol{o}_{1j}}}{\partial \boldsymbol{o}_{11}} \right) = \boldsymbol{t}_{11} \left(\boldsymbol{1} - \frac{\boldsymbol{1}}{\sum_{j=1}^{K} \boldsymbol{e}^{\boldsymbol{o}_{1j}}} \boldsymbol{e}^{\boldsymbol{o}_{11}} \right)$$

$$= \boldsymbol{t}_{11} (\boldsymbol{1} - \boldsymbol{y}_{11}) = \boldsymbol{t}_{11} - \boldsymbol{t}_{11} \boldsymbol{y}_{11}$$

Το γινόμενο $\boldsymbol{t}_{11} \boldsymbol{y}_{11}$ μπορεί να είναι είτε 0 είτε $\boldsymbol{y}_{11}$, καθώς το $\boldsymbol{t}_{11}$ θα είναι είτε 0 είτε 1 ανάλογα με την κατηγορία του παραδείγματος. Εμείς θέλουμε άσχετα με την κατηγορία του παραδείγματος να αφαιρούμε αυτό που προέβλεψε το μοντέλο μας, έτσι ώστε να φαίνεται η διαφορά της πρόβλεψης από την πραγματική τιμή.

Οπότε, τελικά κρατάμε:

$$\frac{\partial E(w)}{\partial o_{11}} = t_{11} - y_{11}$$

Όμοια, προκύπτει ο εξής τύπος:

$$\frac{\partial E(w)}{\partial o_{ij}} = t_{ij} - y_{ij}, \qquad \begin{matrix} i = 1, \dots, N_b \\ j = 1, \dots, K \end{matrix} \qquad (4)$$

Άρα, για τη σχέση (3) τελικά έχουμε:

$$(3) \overset{(4)}{\Longrightarrow} \frac{\partial E(w)}{\partial O} = \begin{bmatrix} t_{11} - y_{11} & \cdots & t_{1K} - y_{1K} \\ \vdots & \ddots & \vdots \\ t_{N_b 1} - y_{N_b 1} & \cdots & t_{N_b K} - y_{N_b K} \end{bmatrix} = T - Y \qquad (5)$$

Με το συνδυσμό των παραπάνω τύπων προκύπτει το εξής:

$$(1) \overset{(2)\ \&\ (5)}{\Longrightarrow} \frac{\partial E(w)}{\partial W^{(2)}} = (T - Y)^T \cdot Z$$

Επομένως, κατά την εκπαίδευση, ο πίνακας βαρών των συνδέσεων του κρυφού επιπέδου με το επίπεδο εξόδου ενημερώνεται ως εξής:

$$W^{(2)} = W^{(2)} + lr * \left((T - Y)^T \cdot Z - \lambda * W^{(2)}\right)$$

όπου ο όρος $\lambda * W^{(2)}$ έχει προστεθεί για να περιορίσει την ύπαρξη ακραίων τιμών στα βάρη *(regularization)* και η μεταβλητή lr είναι το βήμα της εκπαίδευσης.

Για τον υπολογισμό του τύπου (B) θα χρησιμοποιήσουμε τον κανόνα της αλυσίδας ως εξής:

$$(B) \Rightarrow \frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{W}^{(1)}} = \frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{I}} \frac{\partial \boldsymbol{I}}{\partial \boldsymbol{W}^{(1)}} \qquad (6)$$

$$\frac{\partial \boldsymbol{I}}{\partial \boldsymbol{W}^{(1)}} = \frac{\partial \boldsymbol{X} \cdot \left(\boldsymbol{W}^{(1)}\right)^{\boldsymbol{T}}}{\partial \boldsymbol{W}^{(1)}} = \boldsymbol{X} \qquad (7)$$

$$\frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{I}} = \begin{bmatrix} \frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{i}_{11}} & \cdots & \frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{i}_{1M}} \\ \vdots & \ddots & \vdots \\ \frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{i}_{N_b 1}} & \cdots & \frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{i}_{N_b M}} \end{bmatrix} \qquad (8)$$

Εξετάζοντας το πρώτο στοιχείο του παραπάνω πίνακα έχουμε:

$$\frac{\partial \boldsymbol{E}(\boldsymbol{w})}{\partial \boldsymbol{i}_{11}} = \frac{\partial \left(\sum_{n=1}^{N} \sum_{k=1}^{K} \boldsymbol{t}_{nk} \log \boldsymbol{y}_{nk} - \frac{\boldsymbol{\lambda}}{\boldsymbol{2}} \left(\left\| \boldsymbol{W}^{(1)} \right\|^2 + \left\| \boldsymbol{W}^{(2)} \right\|^2 \right) \right)}{\partial \boldsymbol{i}_{11}}$$

$$= \frac{\partial \left(\sum_{n=1}^{N} \sum_{k=1}^{K} \boldsymbol{t}_{nk} \log \boldsymbol{y}_{nk} \right)}{\partial \boldsymbol{i}_{11}} - \frac{\partial \frac{\boldsymbol{\lambda}}{\boldsymbol{2}} \left(\left\| \boldsymbol{W}^{(1)} \right\|^2 + \left\| \boldsymbol{W}^{(2)} \right\|^2 \right)}{\partial \boldsymbol{i}_{11}}$$

Το δεξί κλάσμα μηδενίζεται, καθώς ο αριθμητής είναι σταθερός όρος. Στο αριστερό κλάσμα μπορούμε να εφαρμόσουμε τον κανόνα της αλυσίδας προκειμένου να απλοποιηθεί η παράσταση.

$$= \frac{\partial \left(\sum_{n=1}^{N} \sum_{k=1}^{K} \boldsymbol{t}_{nk} \log \boldsymbol{y}_{nk} \right)}{\partial \boldsymbol{o}_{11}} \frac{\partial \boldsymbol{o}_{11}}{\partial \boldsymbol{i}_{11}}$$

Το αριστερό κλάσμα έχει ήδη υπολογιστεί στην προηγούμενη διαδικασία και προκύπτει από τον τύπο (4)*, οπότε κάνουμε αντικατάσταση.*

$$= (\boldsymbol{t}_{11} - \boldsymbol{y}_{11}) \frac{\partial \boldsymbol{o}_{11}}{\partial \boldsymbol{i}_{11}} = (\boldsymbol{t}_{11} - \boldsymbol{y}_{11}) \frac{\partial \sum_{m=1}^{M+1} \boldsymbol{h}(\boldsymbol{i}_{1m}) \left(\boldsymbol{w}_{m1}^{(2)} \right)^{\boldsymbol{T}}}{\partial \boldsymbol{i}_{11}}$$

Στο κλάσμα, ο όρος $\boldsymbol{i}_{11}$ *στον αριθμητή εμφανίζεται μόνο για* $\boldsymbol{m} = \boldsymbol{1}$*, ενώ οι υπόλοιποι όροι του αθροίσματος είναι σταθεροί, οπότε μηδενίζονται.*

$$= (\boldsymbol{t}_{11} - \boldsymbol{y}_{11}) \frac{\partial \boldsymbol{h}(\boldsymbol{i}_{11}) \left(\boldsymbol{w}_{11}^{(2)} \right)^{\boldsymbol{T}}}{\partial \boldsymbol{i}_{11}} = (\boldsymbol{t}_{11} - \boldsymbol{y}_{11}) \left(\boldsymbol{w}_{11}^{(2)} \right)^{\boldsymbol{T}} \frac{\partial \boldsymbol{h}(\boldsymbol{i}_{11})}{\partial \boldsymbol{i}_{11}}$$

Έστω $\boldsymbol{h\_der(a)}$ η παράγωγος της συνάρτησης ενεργοποίησης. Τότε, έχουμε ότι:

$$\frac{\partial \boldsymbol{h(i_{11})}}{\partial \boldsymbol{i_{11}}} = \boldsymbol{h\_der(i_{11})}$$

Οπότε, τελικά έχουμε:

$$\frac{\partial \boldsymbol{E(w)}}{\partial \boldsymbol{i_{11}}} = \boldsymbol{(t_{11} - y_{11})}\left(\boldsymbol{w_{11}^{(2)}}\right)^{\boldsymbol{T}} \boldsymbol{h\_der(i_{11})}$$

Όμοια, προκύπτει ο εξής τύπος:

$$\frac{\partial \boldsymbol{E(w)}}{\partial \boldsymbol{i_{ij}}} = \boldsymbol{(t_{ij} - y_{ij})}\left(\boldsymbol{w_{ij}^{(2)}}\right)^{\boldsymbol{T}} \boldsymbol{h\_der(i_{ij})}, \qquad \begin{matrix} \boldsymbol{i = 1, \dots, N_b} \\ \boldsymbol{j = 1, \dots, M} \end{matrix}$$

Άρα, για τη σχέση (8) τελικά έχουμε:

$$(8) \Rightarrow \frac{\partial \boldsymbol{E(w)}}{\partial \boldsymbol{I}} = \boldsymbol{(T - Y)} \cdot \boldsymbol{W^{(2)}} * \boldsymbol{h\_der(I)} \quad (9)$$

*Όπου με την πράξη * δηλώνουμε πολλαπλασιασμό 2 πινάκων στοιχείο προς στοιχείο.*

Με το συνδυσμό των παραπάνω τύπων προκύπτει το εξής:

$$(6) \xrightarrow{(7)\ \&\ (9)} \frac{\partial \boldsymbol{E(w)}}{\partial \boldsymbol{W^{(1)}}} = \left(\boldsymbol{(T - Y)} \cdot \boldsymbol{W^{(2)}} * \boldsymbol{h\_der(I)}\right)^{\boldsymbol{T}} \cdot \boldsymbol{X}$$

Επομένως, κατά την εκπαίδευση, ο πίνακας βαρών των συνδέσεων του επιπέδου εισόδου με το κρυφό επίπεδο ενημερώνεται ως εξής:

$$\boldsymbol{W^{(1)} = W^{(1)} + lr} * \left(\left(\boldsymbol{(T - Y)} \cdot \boldsymbol{W^{(2)}} * \boldsymbol{h\_der(I)}\right)^{\boldsymbol{T}} \cdot \boldsymbol{X} - \boldsymbol{\lambda} * \boldsymbol{W^{(1)}}\right)$$

Κατά την εκπαίδευση του μοντέλου η παράμετρος $\boldsymbol{lr}$ *μειώνεται στην περίπτωση που η εκπαίδευση «κάνει κύκλους», δηλαδή το εκάστοτε κοντινότερο τοπικό μέγιστο είναι σε απόσταση τέτοια όπου με ένα βήμα* $\boldsymbol{lr}$ *προς την κατεύθυνσή του θα το προσπεράσει και θα βρεθεί σε χειρότερη κατάσταση. Με τη μείωση του βήματος σε αυτές τις περπτώσεις δίνουμε στο μοντέλο τη δυνατότητα να προσεγγίσει καλύτερα το τοπικό μέγιστο.*

Έλεγχος Κλίσης (Gradient Checking)

Προκειμένου να χρησιμοποιήσουμε τις κλίσεις που υπολογίσαμε στην προηγούμενη ενότητα και να εκπαιδεύσουμε το μοντέλο μας, πρέπει να σιγουρευτούμε πως οι κλίσεις αυτές υπολογίζονται σωστά. Για να το κάνουμε αυτό χρησιμοποιούμε την αριθμητική μέθοδο *Gradient check*, η οποία θα μας δώσει μία προσέγγιση για τις τιμές της κλίσης. Στη συνέχεια, συγκρίνουμε τις προσεγγίσεις του *Gradient check* με τις τιμές που δίνουν οι παραπάνω υπολογισμοί των κλίσεων.

Οι κλίσεις που υπολογίστηκαν παραπάνω είναι σωστές, καθώς η διαφορά των προσεγγίσεων από τις υπολογισμένες είναι της τάξης του $\mathbf{4} * \mathbf{10^{-7}}$ για τον πίνακα βαρών $\boldsymbol{W^{(1)}}$ και της τάξης του $\mathbf{2} * \mathbf{10^{-7}}$ για τον πίνακα βαρών $\boldsymbol{W^{(2)}}$.

Αποτελέσματα

MNIST

Παρακάτω φαίνονται τα αποτελέσματα που έδωσε το μοντέλο στα δεδομένα ελέγχου του συνόλου δεδομένων **MNIST** μετά την εκπαίδευση για τις διάφορες τιμές των παραμέτρων ***lr*** *(initial learning rate)*, ***M***, ***λ*** και για κάθε μία από τις τρεις διαφορετικές συναρτήσεις ενεργοποίησης.
Σε όλες τις περιπτώσεις ο αλγόριθμος εκτέλεσε 50 επαναλήψεις και χρησιμοποιήθηκε **minibatch** μεγέθους 100.

Συνάρτηση: **softplus**

		λ			
		0.1		0.01	
		lr		**lr**	
		0.0001	0.00001	0.0001	0.00001
M	100	0.9202	0.898	0.9123	0.898
	200	0.9065	0.899	0.9063	0.899
	300	0.9072	0.9004	0.9065	0.899

Συνάρτηση: **tanh**

		λ			
		0.1		0.01	
		lr		**lr**	
		0.01	0.001	0.001	0.0001
M	100	0.9663	0.9617	0.9457	0.9116
	200	0.9478	0.965	0.9499	0.9359
	300	0.9338	0.9667	0.9449	0.9277

Συνάρτηση: **cos**

		λ			
		0.1		0.01	
		lr		**lr**	
		0.01	0.001	0.01	0.001
M	100	0.9728	0.971	0.9403	0.9667
	200	0.9738	0.9737	0.9774	0.9697
	300	0.9704	0.9736	0.9804	0.9726

CIFAR-10

Παρακάτω φαίνονται τα αποτελέσματα που έδωσε το μοντέλο στα δεδομένα ελέγχου του συνόλου δεδομένων **CIFAR-10** μετά την εκπαίδευση για τις διάφορες τιμές των παραμέτρων $\boldsymbol{lr}$ *(initial learning rate)*, $\boldsymbol{M}$ και για κάθε μία από τις τρεις διαφορετικές συναρτήσεις ενεργοποίησης.
Σε όλες τις περιπτώσεις ο αλγόριθμος εκτέλεσε 100 επαναλήψεις, χρησιμοποιήθηκε $\boldsymbol{\lambda = 0.1}$ και μέγεθος **minibatch** 100.

Συνάρτηση: **softplus**

		lr
		0.000001
M	100	0.2117
	200	0.1564
	300	0.1548

Συνάρτηση: **tanh**

		lr	
		0.01	0.001
M	100	0.4405	0.4365
	200	0.4287	0.3759
	300	0.3831	0.321

Συνάρτηση: **cos**

		lr	
		0.01	0.001
M	100	0.22	0.3084
	200	0.2629	0.3612
	300	0.3133	0.3764