ΜΗΧΑΝΙΚΗ ΜΑΘΗΣΗ
PROJECT: 1. ΥΛΟΠΟΙΗΣΗ ΝΕΥΡΩΝΙΚΟΥ ΔΙΚΤΥΟΥ
ΧΡΗΣΤΟΣ ΚΟΡΜΑΡΗΣ
ΑΜ: ΕΥ1617

ΗΜΕΡΟΜΗΝΙΑ ΠΑΡΑΔΟΣΗΣ: 11/06/2017

Η εργασία αφορά την κατασκευή ενός νευρωνικού δικτύου με M κρυμμένα επίπεδα. Έχει υλοποιηθεί ο αλγόριθμος ανοδικής κλήσης, για μεγιστοποίηση της συνάρτησης κόστους E(w). Σε κάθε επανάληψη η λογαριθμική πιθανοφάνεια αυξάνεται, όπως είναι αναμενόμενο.
Συμπεριλαμβάνονται παραδείγματα εκτέλεσης για διάφορες τιμές του M, για καθεμία από τις 3 συναρτήσεις ενεργοποίησης, με διάφορες τιμές της παραμέτρου κανονικοποίησης λ και για διάφορες τιμές αριθμού μέγιστων επαναλήψεων.

Το πρόγραμμα αρχικά ζητάει από το χρήστη να εισάγει τον αριθμό των κρυμμένων επιπέδων M, να επιλέξει τη συνάρτηση ενεργοποίησης h1, h2 ή h3, την παράμετρο κανονικοποίησης λ και τον μέγιστο αριθμό επαναλήψεων. Στη συνέχεια το πρόγραμμα ζητάει από το χρήστη να επιλέξει να γίνει εκτέλεση με όλα τα train και test δεδομένα ή ένα μικρό υποσύνολο των δεδομένων για λόγους επιτάχυνσης. Σχετικά με την επιλογή για μικρό δείγμα δεδομένων, η εκπαίδευση (train) του νευρωνικού δικτύου γίνεται πάνω σε ένα δείγμα **6000** δεδομένων, που είναι **τυχαίες** γραμμές του πίνακα X (μεγέθους 60000 γραμμών), 600 δεδομένα για κάθε κατηγορία. Όσον αφορά τα δεδομένα δοκιμής (test data) μικρού δείγματος, επιλέγονται ως **1000 τυχαίες** γραμμές του πίνακα Xtest (μεγέθους 10000 γραμμών), 100 δεδομένα για κάθε κατηγορία.
Στον τελικό πίνακα με τα δεδομένα εκπαίδευσης X και στον πίνακα με τα δεδομένα δοκιμής, Xtest, γίνεται κανονικοποίηση κάνοντας διαίρεση με το εύρος τιμών των pixel: **εύρος = 255 - 0 = 255**.
Υπάρχουν συνολικά 10 κατηγορίες, που αντιστοιχούν στα ψηφία του δεκαδικού συστήματος.
Τα αποτελέσματα καταγράφονται σε αρχείο με ονομασία ανάλογα με την τιμή του M, τη συνάρτηση ενεργοποίησης, το λ και το μέγιστο αριθμό επαναλήψεων.
Γίνεται φόρτωση του συνόλου MNIST από τα αρχεία του φακέλου **“mnisttxt”** που βρίσκονται στο eclass και χρήση της συνάρτησης **softmax**.
Ο αλγόριθμος ανοδική κλίσης έχει υλοποιηθεί στο αρχείο “**ml_softmaxTrain.m**”. Ο υπολογισμός της συνάρτησης κόστους (λογαριθμική πιθανοφάνεια) έχει υλοποιηθεί μέσα στο αρχείο “**costgrad_softmaxNN.m”**. Ο έλεγχος ορθότητας των μερικών παραγώγων έχει υλοποιηθεί μέσα στο αρχείο “**gradcheck_softmaxNN.m**”. Η τελευταία συνάρτηση για τον έλεγχο των μερικών παραγώγων καλείται μόνο στην αρχή της εκτέλεσης του αλγορίθμου και πριν την εκπαίδευση για λόγους ταχύτητας.
Τα αποτελέσματα εκτέλεσης της συνάρτησης “**gradcheck_softmaxNN.m**” για M=2, συνάρτηση ενεργοποίησης h1, και λ = 0.5 ήταν τα εξής:

```
Gradcheck for parameters W1, W2.
The maximum absolute norm for parameter W1 in the gradcheck is: 1.4665e-07
The maximum absolute norm for parameter W2 in the gradcheck is: 1.6046e-05
```

Περαιτέρω έχει υλοποιηθεί η κάθε συνάρτηση ενεργοποίησης στο αρχείο **h1.m, h2.m και h3.m** αντίστοιχα. Η κάθε συνάρτηση δέχεται μία τιμή a και επιστρέφει την h(a), ή και την h'(a) προαιρετικά, σαν δεύτερη έξοδο. Οι τιμές h'(a) χρησιμοποιούνται για τον υπολογισμό της μερικής παραγώγου της συνάρτησης κόστους E(w) ως προς την παράμετρο $W^{(1)}$.

ΑΠΟΤΕΛΕΣΜΑΤΑ ΕΚΤΕΛΕΣΗΣ

- **maxiter=1000, συνάρτηση ενεργοποίησης h1**

M	**2**	**100**	**200**	**300**	**400**	**500**
λ	0	0.5	0.5	0.5	0.5	1
σφάλμα	41.17%	4.23%	5.07%	5.39%	5.09%	5.59%
MLE	-69782.692765	-8390.646882	-9934.185910	-10528.902568	-10177.406990	-11415.571013

- **maxiter=1000, συνάρτηση ενεργοποίησης h2**

M	**2**	**100**	**200**	**300**	**400**	**500**
λ	0	0.5	0.5	0.5	0.5	1
σφάλμα	48.53%	4.38%	4.95%	4.89%	5.03%	5.41%
MLE	-77867.698127	-8554.169011	-9709.737894	-9923.486137	-10071.123425	-11200.693085

- **maxiter=1000, συνάρτηση ενεργοποίησης h3**

M	**2**	**100**	**200**	**300**	**400**	**500**
λ	0	0.5	0.5	0.5	0.5	1
σφάλμα	40.69%	4.39%	4.75%	5.04%	4,86%	5.98%
MLE	-77389.543324	-8524.959369	-9370.249835	-10285.668998	-9858.084165	-12746.456739

Παρατηρείται ότι με μεγαλύτερο αριθμό κρυφών επιπέδων M, το μοντέλο αποδίδει καλύτερα στα δεδομένα δοκιμής. Το μικρότερο σφάλμα κατηγοριοποίησης των δεδομένων δοκιμής που βρέθηκε είναι 4.23% με MLE (maximum likelihood estimate) στην τελευταία επανάληψη -8390.646882 για M = 100, συνάρτηση ενεργοποίησης h2, λ = 0.5 και αριθμό μέγιστων επαναλήψεων = 1000.

Σημείωση: Ο έλεγχος ορθότητας της μερικής παραγώγου στα παραδείγματα εκτέλεσης έχει απενεργοποιηθεί για λόγους επιτάχυνσης.

<u>Απόδειξη με συνάρτηση ενεργοποίησης h1 = log(1 + e^a)</u>

Δεδομένα

Η συνάρτηση κόστους:

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

όπου:

$$E_n(w)=\sum_{k=1}^{K} t_{nk}\cdot \log y_{nk} \quad , \quad \|w\|=\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{D+1}\left(W^{(1)}.*W^{(1)}\right)_{ij}+\sum_{i=1}^{K}\sum_{j=1}^{M+1}\left(W^{(2)}.*W^{(2)}\right)_{ij}} \quad , \quad y_{nk}=\frac{e^{\left(w_k^{(2)}\right)^T\cdot z_n}}{\sum_{j=1}^{K} e^{\left(w_j^{(2)}\right)^T\cdot z_n}}$$

Το y_{nk} υπολογίζει την softmax πιθανότητα. Επίσης ισχύουν:

$a_{nj}=\left(w_j^{(1)}\right)^T\cdot x_n$, j = 1, . . . , M, n = 1, . . . , N

$h_1(a)=\log\left(1+e^a\right)$, η παράγωγος της h: $h_1{}'(a)=\left(\log\left(1+e^a\right)\right)'=\frac{1}{1+e^a}\cdot e^a=\frac{e^a\cdot\frac{1}{e^a}}{\left(1+e^a\right)\cdot\frac{1}{e^a}}=\frac{1}{e^{-a}+1}$

$z_{nj}=h\left(a_{nj}\right)$, j = 1, . . . , M, n = 1, . . . , N

$\nabla z_{nj}=h'\left(a_{nj}\right)$, j = 1, . . . , M, n = 1, . . . , N

$s_{nk}=y_{nk}$, n= 1, . . . , N, k = 1, . . . , K

υπολογίζεται ότι:

$\delta_{nk}=\left(t_{nk}-s_{nk}\right)^T=\left(T-S\right)^T$, n= 1, . . . , N, k = 1, . . . , K

$$\frac{\frac{\lambda}{2}\cdot\|w\|^2}{\partial W^{(2)}}=\lambda\cdot W^{(2)}$$

Η μερική παράγωγος της συνάρτησης κόστους E(W) ως προς την παράμετρο $W^{(2)}$ είναι:

$$\frac{\partial E(w)}{\partial W^{(2)}}=\frac{\partial\sum_{n=1}^{N} E_n(w)-\frac{\lambda}{2}\cdot\|w\|^2}{\partial W^{(2)}}=\sum_{n=1}^{N}\frac{\partial E_n(w)}{\partial W^{(2)}}-\frac{\frac{\lambda}{2}\|w\|^2}{\partial W^{(2)}}=\sum_{n=1}^{N}\frac{\partial E_n(w)}{\partial W^{(2)}}-\lambda\cdot W^{(2)}$$

τελικά:

$$\frac{\partial E(w)}{\partial W^{(2)}}=\delta_{nk}\cdot x_{nd}-\lambda\cdot W^{(2)}=\left(T-S\right)^T\cdot X-\lambda\cdot W^{(2)}$$

Ψάχνουμε να βρούμε την μερική παράγωγος της συνάρτησης κόστους ως προς την παράμετρο $W^{(1)}$.

Πηγαίνουμε ξανά στη συνάρτηση κόστους:

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

όπου:

$$E_n(w)=\sum_{k=1}^{K} t_{nk}\cdot\log y_{nk}$$

Παραγωγίζουμε την E(w) ως προς $W^{(1)}$.

$$\frac{\partial E(w)}{\partial W^{(1)}}=\frac{\partial\sum_{n=1}^{N} E_n(w)-\frac{\lambda}{2}\cdot\|w\|^2}{\partial W^{(1)}}=\frac{\sum_{n=1}^{N}\partial E_n(w)}{\partial W^{(1)}}-\frac{\frac{\lambda}{2}\|w\|^2}{\partial W^{(1)}}=\frac{\sum_{n=1}^{N}\partial E_n(w)}{\partial W^{(1)}}-\lambda\cdot W^{(1)}$$

όπου:

$$\frac{\frac{\lambda}{2}\cdot\|w\|^2}{\partial W^{(1)}}=\lambda\cdot W^{(1)}$$

Αρκεί να υπολογίσουμε την παράγωγο της E_n(w) ως προς $W^{(1)}$.

$$\frac{\partial E_n(w)}{\partial W^{(1)}}=\frac{\sum_{k=1}^{K} t_{nk}\cdot\log y_{nk}}{\partial W^{(1)}}$$

Αντικαθιστούμε το y_{nk} στην E_n(w).

$$E_n(w)=\sum_{k=1}^{K} t_{nk}\cdot\log\frac{e^{\left(w_k^{(2)}\right)^T\cdot z_n}}{\sum_{j=1}^{K} e^{\left(w_j^{(2)}\right)^T\cdot z_n}}=\sum_{k=1}^{K} t_{nk}\cdot\log e^{\left(w_k^{(2)}\right)^T\cdot z_n}-t_{nk}\cdot\log\sum_{j=1}^{K} e^{\left(w_j^{(2)}\right)^T\cdot z_n}\quad =>$$

$$E_n(w)=\sum_{k=1}^{K} t_{nk}\cdot\left(w_k^{(2)}\right)^T\cdot z_n\cdot\log e-t_{nk}\cdot\log\sum_{j=1}^{K} e^{\left(w_j^{(2)}\right)^T\cdot z_n}=\sum_{k=1}^{K} t_{nk}\cdot\left(w_k^{(2)}\right)^T\cdot z_n-\log\sum_{j=1}^{K} e^{\left(w_j^{(2)}\right)^T\cdot z_n}$$

όπου:

$$log e=1\ \ ,\ \ \sum_{k=1}^{K} t_{nk}=1$$

Κανόνας της αλυσίδας:

$$\frac{\partial E_n(w)}{\partial W^{(1)}} = \frac{\partial E_n(w)}{\partial (w_j^{(2)})^T \cdot z_n} \cdot \frac{\partial (w_j^{(2)})^T \cdot z_n}{\partial z_n} \cdot \frac{\partial z_n}{\partial (w_j^{(1)})^T \cdot x_n} \cdot \frac{\partial (w_j^{(1)})^T \cdot x_n}{\partial W^{(1)}}$$

όπου:

- $$\frac{\partial E_n(w)}{\partial (w_j^{(2)})^T \cdot z_n} = \frac{\partial \sum_{k=1}^{K} t_{nk} \cdot (w_k^{(2)})^T \cdot z_n - \log \sum_{j=1}^{K} e^{(w_j^{(2)})^T \cdot z_n}}{\partial (w_j^{(2)})^T \cdot z_n} = \frac{\partial \sum_{k=1}^{K} t_{nk} \cdot (w_k^{(2)})^T \cdot z_n}{\partial (w_j^{(2)})^T \cdot z_n} - \frac{\partial \log \sum_{j=1}^{K} e^{(w_j^{(2)})^T \cdot z_n}}{\partial (w_j^{(2)})^T \cdot z_n} \quad =>$$

  $$\frac{\partial E_n(w)}{\partial (w_j^{(2)})^T \cdot z_n} = t_{nk} - \frac{e^{(w_k^{(2)})^T \cdot z_n}}{\sum_{j=1}^{K} e^{(w_j^{(2)})^T \cdot z_n}} = t_{nk} - y_{nk} = T - Y$$

- $$\frac{\partial (w_j^{(2)})^T \cdot z_n}{\partial z_n} = W^{(2)}$$

- $$\frac{\partial z_n}{\partial (w_j^{(1)})^T \cdot x_n} = h_1'(a) = \frac{1}{e^{-a}+1} = \frac{1}{e^{-(w_j^{(1)})^T \cdot x_n}+1}$$

- $$\frac{\partial (w_j^{(1)})^T \cdot x_n}{\partial W^{(1)}} = X$$

αντικαθιστώντας:

$$\frac{\partial E_n(w)}{\partial W^{(1)}} = ((T-Y) * W^{(2)} .* h'((w_j^{(1)})^T \cdot x_n))^T * X$$

θέτουμε $\delta_{jn} = ((T-Y) * W^{(2)} .* h'((w_j^{(1)})^T \cdot x_n))^T$

άρα:

$$\frac{\partial E_n(w)}{\partial W^{(1)}} = \delta_{jn} * X$$

Τελικά:

$$\frac{\partial E(W)}{\partial W^{(1)}} = \delta_{jn} \cdot x_{nd} - \lambda \cdot W^{(1)} = \delta_{jn} \cdot X - \lambda \cdot W^{(1)}$$