Μηχανική Μάθηση
Προγραμματιστική άσκηση
Διδάσκων: Μ. Τίτσιας
Παράδοση 8 Δεκεμβρίου

Έστω η συνάρτηση κόστους (λογαριθμική πιθανοφάνεια συν όρος κανονικοποίησης) την οποία θέλουμε να μεγιστοποιήσουμε για το πρόβλημα κατηγοριοποίησης K κατηγοριών

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

όπου κάθε πιθανότητα y_{nk} για το $(D+1)$-διάστατο δεδομένο εισόδου $\mathbf{x}_n$ ορίζεται από ένα νευρωνικό δίκτυο με ένα κρυμμένο επίπεδο το οποίο έχει M κρυμμένες υπολογιστικές μονάδες (hidden units). Συγκεκριμένα το y_{nk} δίνεται από την σχέση

$$y_{nk} = \frac{e^{(\mathbf{w}_k^{(2)})^T \mathbf{z}_n}}{\sum_{j=1}^{K} e^{(\mathbf{w}_j^{(2)})^T \mathbf{z}_n}},$$

όπου το $(M+1)$-διάστατο διάνυσμα $\mathbf{z}_n$ αποθηκεύει τις εξόδους του κρυμμένου επιπέδου (activation function values) τέτοιες ώστε $z_{n0} = 1$ και

$$z_{nj} = h\left((\mathbf{w}_j^{(1)})^T \mathbf{x}_n\right), \quad j = 1, \ldots, M,$$

όπου η συνάρτηση ενεργοποίησης $h(\cdot)$ της υπολογιστικής μονάδας j θα μπορεί να έχει μια από τις εξής μορφές

$$h(\alpha) = \log(1 + e^{\alpha}).$$

$$h(\alpha) = \frac{e^{\alpha} - e^{-\alpha}}{e^{\alpha} + e^{-\alpha}}$$

$$h(\alpha) = \cos(\alpha)$$

έτσι ώστε ο χρήστης θα μπορεί να επιλέγει μια εκ των παραπάνω συναρτήσεων. Όλοι οι παράμετροι $\mathbf{w}$ του νευρωνικού δικτύου μπορούν να αναπαρασταθούν από δύο πίνακες: τον $K \times M + 1$ πίνακα $W^{(2)}$ (όπου η k γραμμή του αποθηκεύει το διάνυσμα $\mathbf{w}_k^{(2)}$) καθώς και τον $M \times (D + 1)$ πίνακα $W^{(1)}$ (όπου η j γραμμή του αποθηκεύει το διάνυσμα $\mathbf{w}_j^{(1)}$).

Ο σκοπός της προγραμματικής άσκησης είναι να υλοποιήσετε σε Python τον αλγόριθμο της στοχαστικής ανοδικής κλίσης (stochastic gradient ascent) για την μεγιστοποίηση της παραπάνω συνάρτησης κόστους και της εκτίμησης των παραμέτρων $W^{(2)}$ και $W^{(1)}$. Για το σκόπο αυτό θα πρέπει να υλοποιήσετε μια συνάρτηση που επιστρέφει την τιμή του κόστους καθώς και όλες τις μερικές παραγώγους για τους δύο πίνακες παραμέτρων (λόγω του ότι απαιτούνται στον αλγόριθμο stochastic gradient ascent). Οι μερικές παραγώγοι για τις παραμέτρους $W^{(2)}$ έχουν όμοια μορφή με αυτή της απλής γραμμικής λογιστικής παλινδρόμησης πολλών κατηγοριών και δίνονται από την σχέση

$$(T - Y)^T Z - \lambda W^{(2)},$$

όπου T είναι ένας $N_b \times K$ πίνακας με όλα τα δεδομένα εξόδου για το minibatch δεδομένων μεγέθους N_b, δηλ. τέτοιος ώστε $[T]_{nk} = t_{nk}$, Y είναι ο αντίστοιχος $N_b \times K$ πίνακας που αποθηκεύει τις τιμές των softmax πιθανοτήτων, δηλ. $[Y]_{nk} = y_{nk}$ και Z είναι ο $N_b \times (M + 1)$ πίνακας στον αποίο αποθηκεύουμε τα διανύσματα $\mathbf{z}_n$ των εξόδων του κρυμμένου επιπέδου. Χρησιμοποιώντας τον κανόνα αλυσίδας θα πρέπει να βρείτε μια αντίστοιχη σχέση που θα εκφράζει τις μερικές παράγωγους για τον πίνακα παραμέτρων $W^{(1)}$. Έπειτα θα είστε σε θέση να προγραμματίσετε την σύναρτηση/μέθοδο που θα επιστρέφει την τιμή του κόστους καθώς και όλες τις μερικές παραγώγους και για τους δύο πίνακες.

Ως έλεγχο ορθότητας των μερικών παραγώγων, θα πρέπει να κατασκευάσετε και μια αντίστοιχη συνάρτηση gradcheck που θα συγκρίνει τις αναλυτικές παραγώγους με αριθμητικές διαφορές ακριβώς όπως παρουσιάζεται στις διαφάνειες που υπάρχουν στο e-class και συγκεκριμένα στο σύνολο διαφανειών με τίτλο Non-linear models and neural networks (part II).

Στην συνέχεια θα χρησιμοποιήσετε την υλοποίησή σας για την κατασκευή ενός συστήματος κατηγοριοποίησης για το σύνολο MNIST που υπάρχει στο e-class καθώς και για το σύνολο CIFAR-10 (https://www.cs.toronto.edu/~kriz/cifar.html). Στο αλγόριθμο του stochastic gradient ascent μπορείτε να χρησιμοποιήσετε minibatches μεγέθους 100 ή 200.

Ως παραδοτέο θα πρέπει να ανεβάσετε στο e-class ένα αρχείο τύπου zip με τον κώδικα σας που επίσης θα περιέχει ένα αρχείο τύπου pdf με 1) τις μαθηματικές σχέσεις των μερικών παραγώγων που χρησιμοποιήσατε για τον πίνακα $W^{(1)}$ και 2) παράδειγματα εφαρμογής του συστήματός σας και για τα δύο προβλήματα (MNIST,CIFAR-10) για διαφορετικές τιμές M (π.χ. $M = 100, 200, 300$), για τις τρεις διαφορετικές συνάρτησεις ενεργοποίησης $h(\cdot)$ και όπου σε κάθε περίπτωση θα ανάφερετε το σφάλμα στα δεδομένα ελέγχου, τον αριθμό επαναλήψεων του αλγόριθμου, την τιμή του learning rate η καθώς και την τιμή του λ.