Εργασία: Μηχανική Μάθηση
Τμήμα Πληροφορικής ΟΠΑ

25 Νοεμβρίου 2022

Σκοπός της προγραμματιστικής εργασίας σε Python είναι να κατασκευάσετε δύο δυαδικούς ταξινομητές και να τους εκπαιδεύσετε/αξιολογήσετε σε ένα υποσύνολο του συνόλου δεδομένων MNIST (εύκολα διαθέσιμου μέσω της βιβλιοθήκης Tensorflow Datasets).

ΜΕΡΟΣ Α (Προεπεξεργασία)

α) Κατεβάστε τα δεδομένα των κλάσεων «5» και «6», τα οποία είναι ήδη χωρισμένα σε σύνολο εκπαίδευσης και σύνολο ελέγχου. Στη συνέχεια, διασπάστε το δοθέν σύνολο εκπαίδευσης σε αναλογία 80%/20%. Χρησιμοποιήστε το μικρότερο υποσύνολο ως σύνολο επικύρωσης.

β) Προεπεξεργαστείτε όλες τις εικόνες (εκπαίδευσης, επικύρωσης και ελέγχου) με τον εξής τρόπο:

- Μετατρέψτε κάθε εικόνα διάστασης 28×28 pixel σε ένα διάνυσμα 784 στοιχείων.
- Μετατρέψτε όλα τα στοιχεία των εν λόγω διανυσμάτων, από ακεραίους στο διάστημα $[0, 255]$ σε πραγματικούς αριθμούς στο διάστημα $[0, 1]$.

ΜΕΡΟΣ Β (Λογιστική Παλινδρόμηση)

γ) Υλοποιήστε ένα μοντέλο λογιστικής παλινδρόμησης για δυαδική ταξινόμηση των κλάσεων «5» και «6», το οποίο θα εκπαιδεύεται με άνοδο κλίσης (gradient ascent). Ο βελτιστοποιητής (optimizer) πρέπει να γραφεί από το μηδέν.

δ) Εκπαιδεύστε τον ταξινομητή στο σύνολο εκπαίδευσης για πεπερασμένο πλήθος επαναλήψεων. Στη συνέχεια, χρησιμοποιήστε το προκύπτον μοντέλο για να ταξινομήσετε διαδοχικά τα πρότυπα ελέγχου. Υπολογίστε το ποσοστό των σωστά ταξινομούμενων προτύπων ελέγχου και καταγράψτε αυτή την ακρίβεια[1].

ε) Σκεφτείτε πώς να προσθέσετε κανονικοποίηση L_2 (L_2 regularization) στον κώδικα της εκπαίδευσης. Εκπαιδεύστε 100 κανονικοποιημένες εκδοχές του ταξινομητή, η καθεμία με διαφορετική τιμή του βαθμωτού συντελεστή λ (υπερπαράμετρος κανονικοποίησης L_2). Οι διαφορετικές τιμές του λ πρέπει να καλύπτουν ένα εύρος από 10^{-4} έως 10. Σκεφτείτε πώς να εξαπλώσετε όσο το δυνατόν περισσότερο τις 100 τιμές του λ μέσα σε αυτό το εύρος. Αξιολογείστε ξεχωριστά την ακρίβεια ταξινόμησης κάθε προκύπτοντος μοντέλου (από τα εν

[1] Χρησιμοποιήστε 2 ή 3 δεκαδικά ψηφία.

λόγω 100) στο σύνολο επικύρωσης και επιλέξτε αυτό με το μικρότερο σφάλμα επικύρωσης. Στη συνέχεια, αξιολογείστε την ακρίβειά του στο σύνολο ελέγχου. Τι παρατηρείτε; Δικαιολογήσετε τα αποτελέσματα, σε σύγκριση με το ερώτημα δ).

ΜΕΡΟΣ Γ(Νευρωνικό Δίκτυο)

στ) Κατασκευάστε ένα MLP για δυαδική ταξινόμηση. Πρέπει να έχει ένα επίπεδο εισόδου με 784 νευρώνες εισόδου, ένα κρυμμένο επίπεδο των M νευρώνων και ένα επίπεδο εξόδου με έναν μόνο νευρώνα. Αυτός, κατά τη λειτουργία του δικτύου, θα προβλέπει την πιθανότητα το εκάστοτε πρότυπο εισόδου να ανήκει στην κλάση «6». Χρησιμοποιείστε τη λογιστική σιγμοειδή ως συνάρτηση ενεργοποίησης στους κρυμμένους νευρώνες και στον νευρώνα εξόδου. Σκεφτείτε πώς πρέπει να οργανώσετε τον πίνακα βαρών (weight matrix) και το διάνυσμα πολώσεων (bias vector), με τα οποία αναπαρίσταται κάθε επίπεδο. Πρέπει να υλοποιήσετε χειροκίνητα τον κώδικα της εκπαίδευσης με κάθοδο κλίσης (gradient descent) και οπισθοδιάδοση σφάλματος (error back-propagation), αποθηκεύοντας το τελικό βέλτιστο μοντέλο.

ζ) Υπολογίστε τον αναλυτικό τύπο της παραγώγου της Δυαδικής Διασταυρούμενης Εντροπίας (Binary Cross-Entropy) ως προς την τελική έξοδο του δικτύου, και συμπεριλάβετε στην τελική αναφορά σας την απόδειξη. Για επαλήθευση της ορθότητας του αναλυτικού τύπου, ενσωματώστε στον κώδικά σας το βοηθητικό σκριπτάκι ελέγχου των παραγώγων (gradient checking) από το εργαστήριο του μαθήματος[2]. Εκπαιδεύστε on-line τον ταξινομητή στο σύνολο εκπαίδευσης, με χρήση της Δυαδικής Διασταυρούμενης Εντροπίας ως συνάρτησης κόστους και του πρόωρου σταματήματος (early stopping) ως κριτηρίου τερματισμού. Δηλαδή: i) υπολογίζετε μετά από κάθε εποχή το μέσο κόστος του εκάστοτε τρέχοντος μοντέλου στο σύνολο επικύρωσης, και ii) τερματίζετε την εκπαίδευση αν αυτό το μέσο κόστος επικύρωσης δεν παρουσιάζει καμία μείωση (ή αυξάνεται) για 5 συνεχόμενες εποχές. Στο τέλος, αποθηκεύστε ως βέλτιστο το μοντέλο της εποχής με το ελάχιστο μέσο κόστος επικύρωσης. Σε πόσες εποχές (E) προέκυψε το βέλτιστο μοντέλο;

η) Επαναλάβετε την ανωτέρω διαδικασία για διαφορετικές τιμές υπερπαραμέτρων (ρυθμός μάθησης η της καθόδου κλίσης, πλήθος M κρυμμένων νευρώνων). Δοκιμάστε: i) 10 διαφορετικές τιμές του η σε ένα εύρος από 10^{-5} έως 0.5. Σκεφτείτε πώς να εξαπλώσετε βέλτιστα τις 10 τιμές του η μέσα σε αυτό το εύρος. ii) 10 διαφορετικές τιμές του M, ξεκινώντας από 2 και διπλασιάζοντας κάθε φορά. Εν τέλει, μετά την ανωτέρω χειροκίνητη βελτιστοποίηση υπερπαραμέτρων στο σύνολο επικύρωσης, κρατήστε ως βέλτιστο μόνο το μοντέλο με το συνολικά ελάχιστο μέσο κόστος επικύρωσης και καταγράψτε τις τιμές υπερπαραμέτρων του (η, M, E). Τι παρατηρείτε για την τιμή του E ως συνάρτηση του η; Πώς το εξηγείτε;

θ) Αξιοποιήστε επαναληπτικά, με έναν βρόχο, το αποθηκευμένο συνολικά βέλτιστο εκ-

[2]Τον κώδικα ελέγχου παραγώγων εφαρμόστε τον μία φορά, για να επαληθεύσετε την ορθότητα του αναλυτικού τύπου, και στη συνέχεια απενεργοποιήστε τον. Ο κώδικας αυτός πρέπει να εφαρμοστεί σε κάθε πίνακα βαρών του νευρωνικού δικτύου.

παιδευμένο μοντέλο ώστε να ταξινομήσετε διαδοχικά τα πρότυπα ελέγχου. Υπολογίστε το ποσοστό των σωστά ταξινομούμενων προτύπων ελέγχου, συγκρίνοντας για το καθένα την πρόβλεψη του μοντέλου με την αντίστοιχη ετικέτα (ακρίβεια ταξινόμησης). Καταγράψτε αυτή την ακρίβεια.

ι) Επαναλάβετε τις διαδικασίες των ερωτημάτων ζ) - θ), υλοποιώντας τώρα στοχαστική κάθοδο κλίσης με μέγεθος μικρής δέσμης (mini-batch) ίσο με B. Για να το πετύχετε, τροποποιήστε τον κώδικα εκπαίδευσης ώστε η είσοδος του δικτύου να μην είναι διάνυσμα, αλλά ένας πίνακας $\mathbf{X}$ με B στήλες (κάθε στήλη του $\mathbf{X}$ είναι ένα πρότυπο εκπαίδευσης). Σκεφτείτε πώς πρέπει να τροποποιηθεί το ευθύ και το αντίστροφο πέρασμα, κατά την εκπαίδευση. Αφού υλοποιήσετε την εν λόγω παραλλαγή, δοκιμάστε 8 διαφορετικές τιμές του B, ξεκινώντας από $B = 2$ και διπλασιάζοντας κάθε φορά. Κρατήστε το συνολικά βέλτιστο μοντέλο (με βάση το μέσο κόστος επικύρωσης) και υπολογίστε την ακρίβεια ταξινόμησής του στο σύνολο ελέγχου. Καταγράψτε αυτήν και το βέλτιστο B. Συγκρίνετε τις επιδόσεις του MLP με αυτές του μοντέλου λογιστικής παλινδρόμησης.

Δίνεται η πρότυπη λογιστική σιγμοειδής συνάρτηση και η παράγωγός της:
$\sigma(z) = \frac{1}{1+e^{-z}}$,
$\sigma'(z) = \sigma(z)\,(1 - \sigma(z))$.

ΟΔΗΓΙΕΣ

Μαζί με τον κώδικα, θα παραδώσετε και μία γραπτή αναφορά σε μορφή αρχείου pdf. Στην αρχή της θα γράφετε ποια ερωτήματα έχετε απαντήσει/υλοποιήσει και ποια όχι. Στη συνέχεια, ξεχωριστά για καθένα από τα ερωτήματα δ) - ι): i) θα απαντάτε στα ζητούμενα του κάθε ερωτήματος, και ii) θα εκθέτετε το σκεπτικό σας για κάθε σχεδιαστική σας επιλογή. Στην απάντηση του ερωτήματος ζ), ενσωματώστε με PrintScreen την έξοδο του κώδικα ελέγχου παραγώγων. **Η προθεσμία παράδοσης της εργασίας είναι η ημερομηνία γραπτής εξέτασης του μαθήματος στη χειμερινή εξεταστική της περιόδου 2022 - 2023**.
ΠΡΟΤΕΙΝΕΤΑΙ:

- Να χρησιμοποιήσετε κατάλληλο κώδικα από τα εργαστήρια-φροντιστήρια του μαθήματος, όπου πιστεύετε ότι μπορεί να φανεί χρήσιμος.
- Να αναζητήσετε βοήθεια, ιδέες ή διευκρινήσεις περί της υλοποίησης σε **αξιόπιστες** πηγές (βιβλία ή ιστοσελίδες).

ΔΕΝ ΕΠΙΤΡΕΠΕΤΑΙ:

- Να αντιγράψετε έτοιμο κώδικα από βιβλία ή ιστοσελίδες.
- Να συνεργαστείτε σε ομάδες (η εργασία είναι ατομική).
- Να αντιγράψετε από συμφοιτητή σας κώδικα ή απαντήσεις στα επιμέρους ερωτήματα. Εάν διαπιστωθεί η οποιαδήποτε αντιγραφή, οι εργασίες αυτές μηδενίζονται αυτομάτως.