Υπολογιστική Νοημοσύνη
Εργαστηριακές Ασκήσεις ακ. έτους 2021-22

Να κατασκευάσετε σύνολα δεδομένων Σ1 και Σ2 για τα ακόλουθα προβλήματα:
Σ1) **Πρόβλημα ταξινόμησης τεσσάρων κατηγοριών**: Θα δημιουργήσετε **τυχαία** 8000 παραδείγματα (σημεία (x_1,x_2) στο επίπεδο) μέσα στο τετράγωνο [-1,1]x[-1,1] (4000 για το σύνολο εκπαίδευσης και 4000 για το σύνολο ελέγχου). Στη συνέχεια θα κατατάξετε κάθε παράδειγμα (x_1,x_2) (από τα 8000 παραδείγματα) σε μια κατηγορία από τέσσερεις κατηγορίες ως εξής:
1) εάν $(x_1 – 0.5)^2 + (x_2 – 0.5)^2 <0.16$, τότε το (x_1,x_2) κατατάσσεται στην κατηγορία C1,
2) εάν $(x_1 + 0.5)^2 + (x_2 + 0.5)^2 <0.16$, τότε το (x_1,x_2) κατατάσσεται στην κατηγορία C1,
3) εάν $(x_1 – 0.5)^2 + (x_2 + 0.5)^2 <0.16$, τότε το (x_1,x_2) κατατάσσεται στην κατηγορία C2,
4) εάν $(x_1 + 0.5)^2 + (x_2 – 0.5)^2 <0.16$, τότε το (x_1,x_2) κατατάσσεται στην κατηγορία C2,
Εαν δεν ισχύει κάποια από τις παραπάνω συνθήκες τότε:
5) εάν το (x_1,x_2) ανήκει στο 1ο ή στο 3ο τεταρτημόριο κατατάσσεται στην κατηγορία C3,
6) εάν το (x_1,x_2) ανήκει στο 2ο ή στο 4ο τεταρτημόριο κατατάσσεται στην κατηγορία C4.

Στη συνέχεια **προσθέτουμε θόρυβο μόνο στο σύνολο εκπαίδευσης** ως εξής: για κάθε παράδειγμα του συνόλου εκπαίδευσης με πιθανότητα 0.1 του αλλάζουμε κατηγορία και το αναθέτουμε σε κάποια τυχαία άλλη κατηγορία.

Σ2) **Πρόβλημα ομαδοποίησης εννέα ομάδων** (1200 παραδείγματα): δημιουργούμε **τυχαία** σημεία (x_1,x_2) στο επίπεδο ως εξής: 1) 150 σημεία στο τετράγωνο [0.75,1.25]x[0.75,1.25], 2) 150 σημεία στο τετράγωνο [0, 0.5]x[0,0.5], 3) 150 σημεία στο τετράγωνο [0,0.5] x [1.5,2], 4) 150 σημεία στο τετράγωνο [1.5,2]x[0,0.5], 5) 150 σημεία στο τετράγωνο [1.5,2] x [1.5,2], 6) 75 σημεία στο τετράγωνο [0.6,0.8]x[0,0.4], 7) 75 σημεία στο τετράγωνο [0.6,0.8] x [1.6,2], 8) 75 σημεία στο τετράγωνο [1.2,1.4]x[0,0.4], 9) 75 σημεία στο τετράγωνο [1.2,1.4] x [1.6,2], 10) 150 σημεία στο τετράγωνο [0,2]x[0,2].

Ασκηση 1: Να κατασκευάσετε **προγράμματα ταξινόμησης** βασισμένα στο **πολυεπίπεδο perceptron** (MLP). Το πρόγραμμα **Π1** υλοποιεί MLP με **δύο κρυμμένα επίπεδα** και το πρόγραμμα **Π2** υλοποιεί MLP με **τρία κρυμμένα επίπεδα**.
Και στα δύο προγράμματα οι **συναρτήσεις ενεργοποίησης ορίζονται ως εξής:** i) στα κρυμμένα επίπεδα την υπερβολική εφαπτομένη (tanh(u)) ή την relu και ii) για το επίπεδο εξόδου θα ορίσετε εσείς τη συνάρτηση ενεργοποίησης που απαιτείται για το συγκεκριμένο πρόβλημα.
Το πρόγραμμα θα πρέπει να αποτελείται από τις ακόλουθες μονάδες:
1) Με χρήση της εντολής define, καθορισμός αριθμού εισόδων (d), αριθμού κατηγοριών (K), αριθμού νευρώνων στο πρώτο κρυμμένο επίπεδο (H1), αριθμού νευρώνων στο δεύτερο κρυμμένο επίπεδο (H2), αριθμού νευρώνων στο τρίτο κρυμμένο (H3) (μόνο για το Π2) και είδος συνάρτησης ενεργοποίησης (tanh ή relu) για τα κρυμμένα επίπεδα.
2) Φόρτωμα των συνόλων εκπαίδευσης και ελέγχου (από αντίστοιχα αρχεία) και κωδικοποίηση των κατηγοριών (ορισμός των επιθυμητών εξόδων για κάθε κατηγορία).
3) Καθορισμός της αρχιτεκτονικής του δικτύου MLP. Ορισμός των απαιτούμενων πινάκων και άλλων δομών ως καθολικών μεταβλητών. Καθορισμός του ρυθμού μάθησης και του κατωφλίου τερματισμού. Τυχαία αρχικοποίηση των βαρών/πολώσεων στο διάστημα (-1,1).
4) Υλοποίηση της συνάρτησης forward-pass (float *x, int d, float *y, int K) η οποία υπολογίζει το διάνυσμα εξόδου y (διάστασης K) του MLP δοθέντος του διανύσματος εισόδου x (διάστασης d).
5) Υλοποίηση της συνάρτησης backprop(float *x, int d, float *t, int K) η οποία λαμβάνει τα διανύσματα x διάστασης d (είσοδος) και t διάστασης K (επιθυμητή έξοδος) και υπολογίζει τις παραγώγους του σφάλματος ως προς οποιαδήποτε παράμετρο (βάρος ή πόλωση) του δικτύου ενημερώνοντας τους αντίστοιχους πίνακες.
6) Χρησιμοποιώντας τα παραπάνω να υλοποιήσετε τον **αλγόριθμο εκπαίδευσης gradient descent και ενημέρωση των βαρών ανά ομάδες των Β παραδειγμάτων (mini-batches) θεωρώντας τα Ν**

παραδείγματα του συνόλου εκπαίδευσης (όπου το Β διαιρέτης του Ν και ορίζεται στην αρχή του προγράμματος). Σημειώστε ότι εάν Β=1 έχουμε σειριακή ενημέρωση, ενώ εάν Β=Ν έχουμε ομαδική ενημέρωση.

Στο τέλος κάθε εποχής θα πρέπει υποχρεωτικά να υπολογίζετε και να τυπώνετε την τιμή του συνολικού σφάλματος εκπαίδευσης. Τερματίζουμε όταν η διαφορά της τιμής του σφάλματος εκπαίδευσης μεταξύ δύο εποχών γίνει μικρότερη από κάποιο κατώφλι, αφού όμως ο αλγόριθμος έχει τρέξει για τουλάχιστον 700 εποχές.

7) Αφού τερματιστεί η εκπαίδευση του δικτύου να γίνεται υπολογισμός και εκτύπωση της **ικανότητας γενίκευσης** του δικτύου που προκύπτει, υπολογίζοντας το **ποσοστό σωστών αποφάσεων στο σύνολο ελέγχου**.

Χρησιμοποιώντας τα προγράμματα Π1 και Π2 να μελετήσετε το πρόβλημα Σ1.

Να εξετάσετε και να καταγράψετε πώς μεταβάλλεται η γενικευτική ικανότητα του δικτύου (ποσοστό επιτυχίας στο σύνολο ελέγχου) θεωρώντας:

α) Διάφορους συνδυασμούς τιμών για τα Η1, Η2 (πρόγραμμα Π1) και για τα Η1, Η2, Η3 (πρόγραμμα Π2).

β) Συνάρτηση ενεργοποίησης στους κρυμμένους νευρώνες την υπερβολική εφαπτομένη ή την relu και

γ) Β= Ν/10 ή Ν/100.

Για το δίκτυο με την καλύτερη γενικευτική ικανότητα που θα βρείτε, να τυπώσετε τα παραδείγματα του συνόλου ελέγχου χρησιμοποιώντας διαφορετικό στυλ (πχ + και -) ανάλογα με το αν το παράδειγμα ταξινομείται από το δίκτυο στη σωστή κατηγορία ή όχι.

Προκύπτει κάποιο όφελος από τη χρήση του τρίτου κρυμμένου επιπέδου;

Ασκηση 2: Να κατασκευάσετε **πρόγραμμα ομαδοποίησης** (Π3) **με Μ ομάδες** (το Μ θα ορίζεται με την εντολή #define) **βασισμένο στον αλγόριθμο k-means**. Το πρόγραμμα θα φορτώνει το αρχείο με τα παραδείγματα, θα εκτελεί τον αλγόριθμο k-means με Μ κέντρα και στο τέλος θα αποθηκεύει τις συντεταγμένες των κέντρων των ομάδων. Η αρχική θέση κάθε κέντρου να γίνεται επιλέγοντας τυχαία κάποιο από τα παραδείγματα. Επίσης θα πρέπει στο τέλος να υπολογίζεται **και να τυπώνεται το σφάλμα ομαδοποίησης ως εξής: για κάθε παράδειγμα x_i υπολογίζουμε την Ευκλείδεια απόσταση $||x_i-\mu_k||^2$ από το κέντρο μ_k της ομάδας στην οποία ανήκει και αθροίζουμε τις αποστάσεις για όλα τα παραδείγματα x_i.**

Να εκτελέσετε το πρόγραμμα ομαδοποίησης (Π3) στο σύνολο δεδομένων (Σ2) για Μ=3,5,7,9,11,13 ομάδες. Για **κάθε τιμή του Μ** να κάνετε τα εξής:

α) Να εκτελέσετε 20 τρεξίματα του προγράμματος με διαφορετικά (τυχαία επιλεγμένα αρχικά κέντρα) και να κρατήσετε τη **λύση με το μικρότερο σφάλμα ομαδοποίησης**.

β) Στη συνέχεια να εμφανίσετε (plot) στο ίδιο σχήμα τόσο τα παραδείγματα (π.χ. με '+') όσο και τις θέσεις των κέντρων που βρήκατε (π.χ. με '*').

Βάσει των αποτελεσμάτων να φτιάξετε ένα διάγραμμα που να δείχνει πώς μεταβάλλεται το σφάλμα ομαδοποίησης με τον αριθμό των ομάδων; Μπορεί να χρησιμοποιηθεί το σφάλμα ομαδοποίησης για να εκτιμήσουμε τον πραγματικό αριθμό ομάδων; (στο σύνολο Σ2 ο πραγματικός αριθμός των ομάδων είναι 9).