ΠΑΝΕΠΙΣΤΗΜΙΟ ΙΩΑΝΝΙΝΩΝ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

Μεταπυχιακό Μάθημα

ΒΕΛΤΙΣΤΟΠΟΙΗΣΗ

Ακαδημαϊκό Έτος 2020-21

- ΕΡΓΑΣΙΑ 2 -

<u>Παράδοση μέχρι</u>: 31 Δεκεμβρίου 2020

Τα τεχνητά νευρωνικά δίκτυα (τ.ν.δ.) είναι ευφυή συστήματα υπολογισμού αποτελούμενα από μονάδες επεξεργασίας, τους τεχνητούς νευρώνες, που προσομοιώνουν τη λειτουργία των βιολογικών νευρώνων του εγκεφάλου. Όπως οι συνάψεις σε έναν βιολογικό εγκέφαλο μπορούν να μεταδώσουν σήμα σε άλλους νευρώνες, έτσι κι ένας τεχνητός νευρώνας που λαμβάνει ένα σήμα το επεξεργάζεται και το μεταδίδει σε άλλους νευρώνες με τους οποίους συνδέεται. Το σήμα στα τεχνητά νευρωνικά δίκτυα είναι ένας πραγματικός αριθμός και η έξοδος κάθε νευρώνα υπολογίζεται μέσω κάποιας μη γραμμικής συνάρτησης των εισόδων του.

Τα τ.ν.δ. εφαρμόζονται σε πληθώρα προβλημάτων όπως μοντελοποίηση και προσέγγιση συναρτήσεων, αναγνώριση προτύπων, ταξινόμηση, ανάλυση δεδομένων, ομαδοποίηση κλπ. Ωστόσο, η απόδοσή τους βασίζεται στην κατάλληλη *εκπαίδευσή* τους, δηλαδή στη διαδικασία *βελτιστοποίησης* των παραμέτρων (και πιθανώς της μορφής) του τ.ν.δ. επί ενός συνόλου εκπαίδευσης. Το σύνολο αυτό περιέχει παραδείγματα και τις σωστές απαντήσεις που πρέπει να δώσει το τ.ν.δ. (π.χ. διανύσματα και την κλάση στην οποία ταξινομούνται). Ακολούθως, η απόδοση του τ.ν.δ. αξιολογείται σε άγνωστα παραδείγματα. Η ικανότητα του δικτύου να απαντά σωστά σε άγνωστα παραδείγματα καλείται *γενίκευση*.

Στο παρεχόμενο αρχείο *PNN-paper.pdf* δίνεται μια δημοσιευμένη ερευνητική εργασία με αντικείμενο την εκπαίδευση ενός ιδιαίτερου τύπου τ.ν.δ., το οποίο ονομάζεται ***πιθανοτικό νευρωνικό δίκτυο*** (π.ν.δ. - probabilistic neural network). Αυτός ο τύπος δικτύου χαρακτηρίζεται από μια μοναδική παράμετρο (ομοσκεδαστικό π.ν.δ.) ή ένα σύνολο παραμέτρων (ετεροσκεδαστικό π.ν.δ.) που καθορίζουν τις τυπικές αποκλίσεις ενός συνόλου Gaussian κατανομών, με τις οποίες το π.ν.δ. κάνει ταξινόμηση των δεδομένων σε κατηγορίες. Η εκπαίδευση του π.ν.δ. έγκειται στον προσδιορισμό βέλτιστων τιμών αυτών των τυπικών αποκλίσεων.

Στη δοθείσα εργασία θα δείτε μια εφαρμογή ενός αλγορίθμου νοημοσύνης σμηνών (της particle swarm optimization) για την εκπαίδευση π.ν.δ.. Τα προβλήματα που μελετώνται πειραματικά αφορούν σε ταξινόμηση στα ακόλουθα 4 σύνολα δεδομένων:

1. E.coli dataset (https://archive.ics.uci.edu/ml/datasets/Ecoli)
2. Breast Cancer dataset (https://archive.ics.uci.edu/ml/datasets/Breast+Cancer)
3. Yeast dataset (https://archive.ics.uci.edu/ml/datasets/Yeast)
4. Pima Indians Diabetes dataset (δεν υπάρχει διαθέσιμο)

Επίσης, για την εκπαίδευση και αξιολόγηση του π.ν.δ. απαιτείται ο διαχωρισμός του συνόλου δεδομένων σε σύνολο εκπαίδευσης (training set) και σε σύνολο αξιολόγησης (test set). Ο διαχωρισμός μπορεί να γίνει με διάφορους τρόπους. Στη συγκεκριμένη εργασία χρησιμοποιούνται 3 εναλλακτικές μέθοδοι:

A) Stratified Random Sampling
B) λ-Fold Cross-Validation
C) Train-Validation-Test Partitioning

Οι παραπάνω μέθοδοι περιγράφονται επαρκώς στη δοθείσα εργασία.

Ο στόχος της παρούσας εργασίας είναι να εφαρμοστούν στο συγκεκριμένο πρόβλημα εκπαίδευσης π.ν.δ. οι ακόλουθες μέθοδοι:

I. Nelder-Mead
II. Τυχαία αναζήτηση με χρήση Gaussian κατανομής
III. Τυχαίος περίπατος με μεταβλητό βήμα
IV. Simulated annealing

για ένα από τα παραπάνω διαθέσιμα datasets (1-3) και για έναν από τους παραπάνω τρόπους διαχωρισμού (A-C) που θα επιλέξετε εσείς. Στα datasets οι λεκτικές μεταβλητές μπορούν να αντικατασταθούν από αριθμούς (π.χ. “yes”=1, “no”=0 κ.λ.π.). Αφού καθορίσετε το dataset που θα χρησιμοποιήσετε και την αντίστοιχη μέθοδο διαχωρισμού, ακολουθήστε επακριβώς την αντίστοιχη παραμετροποίηση που προτείνεται στη δοθείσα εργασία (εκτός από το πλήθος των συναρτησιακών υπολογισμών των αλγορίθμων βελτιστοποίησης). Ακολούθως, θα πρέπει να μελετήσετε ενδελεχώς τους ζητούμενους αλγορίθμους. Μια πλήρης μελέτη θα πρέπει να περιλαμβάνει για κάθε αλγόριθμο τα ακόλουθα:

- Πολλαπλά πειράματα (με διαφορετική αρχική συνθήκη) για διαφορετικές παραμέτρους του αλγορίθμου και διαφορετικές τιμές των μέγιστων συναρτησιακών υπολογισμών (computational budget, k_{max}).
- Στατιστική ανάλυση των αποτελεσμάτων (δηλαδή του *accuracy percentage* του π.ν.δ.) και ανάδειξη των καλύτερων εκδοχών του αλγορίθμου για κάθε k_{max} (με έλεγχο στατιστικής σημαντικότητας).
- Σύγκριση της καλύτερης εκδοχής κάθε αλγορίθμου με τους υπόλοιπους αλγορίθμους για κάθε διαφορετικό k_{max}.
- Ενδεικτική σύγκριση με τα δημοσιευμένα αποτελέσματα.
- Παρουσίαση των αποτελεσμάτων διαμέσου πινάκων, γραφημάτων, boxplots κ.λ.π. και σχολιασμός των συμπερασμάτων.

Τα πλήρη αποτελέσματα της μελέτης θα πρέπει να αναλυθούν και να δοθούν σε μια ολοκληρωμένη αναφορά. Επιπλέον, θα πρέπει να περιγράφεται η λειτουργία του κώδικα, ο οποίος θα πρέπει να παραδοθεί μαζί με την αναφορά και θα πρέπει να παράγει τα ίδια αποτελέσματα με τα αναφερθέντα στην αναφορά. Για τον λόγο αυτό, θα πρέπει να διατηρηθεί σε αρχείο το seed της γεννήτριας τυχαίων αριθμών πριν από την κάθε εκτέλεση του αλγορίθμου (π.χ. αν κάνετε 30 πειράματα για μια συγκεκριμένη περίπτωση, να υπάρχει το αρχείο με τα 30 seeds που βάλατε στη γεννήτρια τυχαίων αριθμών).

Οι υλοποιήσεις μπορούν να γίνουν σε γλώσσα προγραμματισμού της επιλογής από τις ακόλουθες: C/ C++, Java, Fortran, Matlab/Octave, Python, R.

Ημερομηνία παράδοσης: **31 Δεκεμβρίου 2020**

Η αξιολόγηση των εργασιών θα βασίζεται στην ποιότητα της αναφοράς, του κώδικα και στην αξιοποίηση γνώσεων από το μάθημα.