ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ Η/Υ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΥΠΟΛΟΓΙΣΤΙΚΗ ΝΟΗΜΟΣΥΝΗ

Διδάσκοντες: Δ. Κουτσομητρόπουλος

Ακαδημαϊκό Έτος 2020-2021

Εργαστηριακή Άσκηση
Μέρος Β΄

Β. Επιλογή Χαρακτηριστικών στο πρόβλημα Αναγνώρισης Χειρόγραφων Ψηφίων με χρήση ΓΑ

Η επιλογή χαρακτηριστικών (feature selection) είναι η διαδικασία μείωσης του αριθμού των εισόδων κατά τον σχεδιασμό και την εφαρμογή ενός αλγορίθμου μηχανικής μάθησης. Ένας αυξημένος αριθμός χαρακτηριστικών οδηγεί σε αύξηση του χρόνου εκπαίδευσης, ενώ παράλληλα δυσχεραίνει τη μάθηση και μπορεί να οδηγήσει σε υπερεκπαίδευση (κατάρα της διαστατικότητας). Γνωστές μέθοδοι για επιλογή χαρακτηριστικών αποτελούν τα ίδια τα ΤΝΔ, η συσχέτιση Pearson, η Ανάλυση Κύριων Συνιστωσών (PCA), η ομαλοποίηση, το dropout κ.α. Στην εργασία αυτή σας ζητείται να προτείνετε και να υλοποιήσετε **Γενετικό Αλγόριθμο** που θα χρησιμοποιηθεί για τη μείωση των χαρακτηριστικών κατά την αναγνώριση χειρόγραφων ψηφίων από ένα ΤΝΔ (Μέρος Α).
Σκοπός του αλγορίθμου είναι να εντοπίσει ποιες από τις 784 εισόδους, που αναπαριστούν μια εικόνα 28Χ28 pixels μπορούν να παραμείνουν και ποιες να παραλειφθούν σε ένα ΤΝΔ που έχει εκπαιδευτεί για αναγνώριση ψηφίων, έτσι ώστε να επιτυγχάνεται μείωση της διαστατικότητας, αλλά και η ακρίβεια του δικτύου να είναι η καλύτερη δυνατή. Για τις ανάγκες τις άσκησης, θα θεωρήσετε το βέλτιστο ΤΝΔ στο οποίο καταλήξατε στο μέρος Α.

Β1. Σχεδιασμός ΓΑ *[30 μονάδες]*

α) Κωδικοποίηση: Να προτείνετε μια κωδικοποίηση για τα άτομα του πληθυσμού. Λάβετε υπόψη τα παρακάτω:

i. Ένα άτομο αναπαριστά το ίδιο νευρωνικό δίκτυο, αλλά με διαφορετικές εισόδους κάθε φορά.
ii. Η ύπαρξη ή μη μιας εισόδου είναι δυαδική (0 ή 1) και υπάρχουν συνολικά 784 είσοδοι.

β) Αρχικός πληθυσμός: Περιγράψτε μια διαδικασία για τη δημιουργία αρχικού πληθυσμού ατόμων. Τα άτομα του πληθυσμού αναπαριστούν ΤΝΔ με διαφορετικές εισόδους.

γ) Συνάρτηση καταλληλότητας: Ένα άτομο είναι πιο κατάλληλο από άλλα, εφόσον:

i. Το ΤΝΔ που αναπαριστά έχει μικρότερο σφάλμα στο σύνολο ελέγχου.
ii. Έχει μικρότερο αριθμό εισόδων.

Επομένως η συνάρτηση καταλληλότητας θα πρέπει να συνδυάζει αυτά τα δύο ανταγωνιστικά κριτήρια. Για το i. μπορείτε να χρησιμοποιήσετε την ελαχιστοποίηση του σφάλματος στο σύνολο ελέγχου με κριτήριο το cross entropy loss. Εναλλακτικά, τη μετρική της *ακρίβειας* (accuracy), που υπολογίζει το ποσοστό των εικόνων που ταξινομούνται σωστά από το δίκτυο. Για το ii. θα πρέπει να εφαρμόζετε μια ποινή στα άτομα που έχουν υψηλό αριθμό εισόδων. Η

ποινή αυτή θα πρέπει να είναι σε κλίμακα ανάλογη με το σφάλμα στο i., έτσι ώστε αφενός να μην κυριαρχεί κατά την αξιολόγηση ενός ατόμου, αφετέρου να παίζει ικανό ρόλο κατά την επιλογή. Να αιτιολογήσετε επαρκώς την συνάρτηση καταλληλότητας στην οποία καταλήξατε.

Ορθότερο φυσικά είναι το κάθε άτομο (ΤΝΔ) να εκπαιδευτεί εκ νέου και να αξιολογηθεί με CV, ώστε να αποφευχθεί πιθανή πόλωση στο συγκεκριμένο σύνολο ελέγχου. Παρόλα αυτά, για λόγους οικονομίας χρόνου, προτείνεται τα βάρη να παραμείνουν σταθερά, όπως προέκυψαν από την εκπαίδευση του καλύτερου μοντέλου στο Α πάνω σε όλο το σύνολο εκπαίδευσης. Στη συνέχεια η αξιολόγηση των ατόμων θα γίνεται με βάση το σφάλμα στο σύνολο ελέγχου. Αυτό έχει λογική βάση, διότι το νευρωνικό αναμένεται να έχει ήδη υποεκτιμήσει τα βάρη των εισόδων που είναι λιγότερο σημαντικές και τα βάρη των σημαντικών εισόδων δεν θα αλλάξουν κατά πολύ σε ενδεχόμενο εκ νέου εκπαίδευσης.

δ) Γενετικοί Τελεστές: Με βάση την κωδικοποίηση που επιλέξατε να προτείνετε τους τελεστές επιλογής, διασταύρωσης και μετάλλαξης που θα χρησιμοποιήσετε.

i. Ειδικά για την επιλογή, να αξιολογήσετε τη χρήση *ρουλέτας με βάση το κόστος, με βάση την κατάταξη* και *τουρνουά.*
ii. Ειδικά για τη διασταύρωση, να αξιολογήσετε την καταλληλότητα των ακόλουθων τελεστών: *Διασταύρωση μονού σημείου, διασταύρωση πολλαπλού σημείου, ομοιόμορφη διασταύρωση*.
iii. Ειδικά για τη μετάλλαξη, να αξιολογήσετε τη χρήση *ελιτισμού*.

Β2. Υλοποίηση ΓΑ *[30 μονάδες]*
Να γράψετε ένα πρόγραμμα, σε οποιοδήποτε περιβάλλον ή γλώσσα προγραμματισμού, που να υλοποιεί τον γενετικό αλγόριθμο που σχεδιάσετε.

Β3. Αξιολόγηση και Επίδραση Παραμέτρων *[30 μονάδες]*
α) Να τρέξετε τον αλγόριθμο για τις τιμές των παραμέτρων που φαίνονται στον παρακάτω πίνακα και να τον συμπληρώσετε. Ο αλγόριθμος θα τερματίζει όταν πληρούνται ένα ή περισσότερα από τα *κριτήρια τερματισμού*, δηλαδή όταν:

i. το καλύτερο άτομο της κάθε γενιάς πάψει να βελτιώνεται για ορισμένο αριθμό γενεών ή
ii. βελτιώνεται κάτω από ένα ποσοστό (<1%) ή
iii. έχει ξεπεραστεί ένας προκαθορισμένος αριθμός γενεών (π.χ. 1000)

Α/Α	ΜΕΓΕΘΟΣ ΠΛΗΘΥΣΜΟΥ	ΠΙΘΑΝΟΤΗΤΑ ΔΙΑΣΤΑΥΡΩΣΗΣ	ΠΙΘΑΝΟΤΗΤΑ ΜΕΤΑΛΛΑΞΗΣ	ΜΕΣΗ ΤΙΜΗ ΒΕΛΤΙΣΤΟΥ	ΜΕΣΟΣ ΑΡΙΘΜΟΣ ΓΕΝΕΩΝ
1	20	0.6	0.00		
2	20	0.6	0.01		
3	20	0.6	0.10		
4	20	0.9	0.01		
5	20	0.1	0.01		
6	200	0.6	0.00		
7	200	0.6	0.01		
8	200	0.6	0.10		
9	200	0.9	0.01		
10	200	0.1	0.01		

Προσοχή: *Επειδή οι ΓΑ είναι στοχαστικοί αλγόριθμοι και συνεπώς δεν εξασφαλίζουν την ίδια απόδοση σε κάθε εκτέλεσή τους, θα πρέπει να εκτελέσετε τον αλγόριθμο τουλάχιστον δέκα φορές για κάθε περίπτωση. Στον πίνακα να σημειώσετε το μέσο όρο της απόδοσης της καλύτερης λύσης σε κάθε τρέξιμο.*

β) Για κάθε περίπτωση του παραπάνω πίνακα να σχεδιάστε την καμπύλη εξέλιξης (απόδοση/αριθμό γενιών) της καλύτερης λύσης (της μέσης τιμής αυτής, σε κάθε τρέξιμο).

γ) Με βάση αυτές τις καμπύλες, αλλά και τα αποτελέσματα του παραπάνω πίνακα, να διατυπώσετε αναλυτικά τα συμπεράσματά σας σχετικά με την επίδραση της κάθε παραμέτρου (μέγεθος πληθυσμού, πιθανότητα διασταύρωσης, πιθανότητα μετάλλαξης) στη σύγκλιση του αλγορίθμου.

Β4. Αξιολόγηση ΤΝΔ *[10 μονάδες]*
α) Να συγκρίνετε την απόδοση του βέλτιστου ΤΝΔ που βρήκατε από τα παραπάνω πειράματα με αυτή του βέλτιστου ΤΝΔ που προέκυψε από το Α και να διατυπώσετε τα συμπεράσματά σας ως προς:

i. την γενικευτική ικανότητα των δύο δικτύων.
ii. την επίδραση της επιλογής (μείωσης) των χαρακτηριστικών στην απόδοση του δικτύου. Υπάρχει κάποια σχέση ανάμεσα στα χαρακτηριστικά που αποκόπτονται και στη θέση των pixels στην εικόνα;
iii. το ενδεχόμενο υπερπροσαρμογής στα δεδομένα ελέγχου.

β) Επανεκπαιδεύστε το ΤΝΔ που προέκυψε από τον ΓΑ με όλο το σύνολο εκπαίδευσης και επαναλάβετε το α. Εξηγείστε αν υπάρχει τώρα διαφοροποίηση.

Παραδοτέα

Η αναφορά που θα παραδώσετε θα πρέπει να περιέχει εκτενή σχολιασμό των πειραμάτων σας, καθώς και πλήρη καταγραφή των αποτελεσμάτων και των συμπερασμάτων σας, ανά υπο-ερώτημα. Επίσης, πρέπει να συμπεριλάβετε στην αρχή της αναφοράς σας ένα link προς τον κώδικα που έχετε χρησιμοποιήσει (σε κάποια file sharing υπηρεσία ή code repo).
Μην ξεχάσετε να συμπληρώσετε τα στοιχεία σας στην αρχή της 1ης σελίδας.

Αξιολόγηση

Η απάντηση των ερωτημάτων Α και Β έχει βαρύτητα 20% στον τελικό βαθμό του μαθήματος (το σύνολο και των δύο μερών της εργασίας έχει βαρύτητα 40%). Ο βαθμός του Bonus (10%) προστίθεται στο παραπάνω ποσοστό 40%.

Παρατηρήσεις

1. Η αναφορά, σε ηλεκτρονική μορφή, πρέπει να αναρτηθεί στο e-class μέχρι τη Δευτέρα, 7/6/2021, στις 23:59.
2. Για οποιαδήποτε διευκρίνηση / ερώτηση μπορείτε να χρησιμοποιείτε το σχετικό forum στο eclass του μαθήματος.