

Εργαστηριακή Άσκηση Μέρος Β' 2020-2021.

Ανδρούτσος Λάμπρος

A.M : 1054396

Η εργασία αυτή δημιουργήθηκε στα πλαίσια του μαθήματος :

Υπολογιστική Νοημοσύνη



Τμήμα Μηχανικών Ηλεκτρονικών Υπολογιστών και Πληροφορικής

Πανεπιστήμιο Πατρών

Ελλάδα

2021

Περιεχόμενα

B1. Σχεδιασμός ΓΑ	3
A) Κωδικοποίηση	3
B) Αρχικός πληθυσμός	4
Γ) Συνάρτηση καταλληλότητας	4
Δ) Γενετικοί Τελεστές.....	4
B2. Υλοποίηση ΓΑ	6
B3. Αξιολόγηση και Επίδραση Παραμέτρων	6
B4. Αξιολόγηση ΤΝΔ	13

Ο κώδικας βρίσκεται **στο παρακάτω link** στο προσωπικό μου github λογαριασμό:

<https://github.com/labrosandroutsos/Genetic-Algorithm-assignment>

Στο repository αυτό θα βρείτε το **καλύτερο μοντέλο** από την πρώτη εργαστηριακή άσκηση, τον **κώδικα του Γενετικού Αλγορίθμου** και την **αναφορά της άσκησης**.

B1. Σχεδιασμός ΓΑ

Σκοπός της δεύτερης εργαστηριακής άσκησης του μαθήματος είναι ο σχεδιασμός και η υλοποίηση ενός Γενετικού Αλγορίθμου, ο οποίος θα χρησιμεύει στην επιλογή χαρακτηριστικών (feature selection). Feature selection είναι η διαδικασία μείωσης του αριθμού των εισόδων κατά τον σχεδιασμό και εφαρμογή ενός αλγορίθμου μάθησης, και συγκεκριμένα σε αυτό το πρότζεκτ θα υλοποιηθεί το feature selection κατά την αναγνώριση χειρόγραφων ψηφίων μέσω ενός Γενετικού Αλγορίθμου. Ουσιαστικά, το πρότζεκτ αυτό αποτελεί την συνέχεια της πρώτη εργαστηριακής άσκησης, στην οποία υλοποιήσαμε και κάναμε fine tuning σε ΤΝΔ πάνω στο ίδιο dataset.

Το dataset αυτό αποτελείται από εικόνες με 28x28 pixels, οι οποίες αντιπροσωπεύουν αριθμούς από το 0 έως το 9, χειρόγραφα γραμμένους. Το dataset αυτό θα χρησιμοποιηθεί όπως και στην 1^η άσκηση, όπου τα δεδομένα είναι Flatten (μετατροπή matrix N διαστάσεων σε one dimensional matrix), δηλαδή θα είναι 28x28 = 784 πια είσοδοι. Κάθε είσοδος από τις 784 αποτελεί ένα pixel και στην παρούσα άσκηση θα προσπαθήσω να ξεχωρίσω ποια pixels αποτελούν τα σημαντικότερα για το πρόβλημα αναγνώρισης χειρόγραφων ψηφίων και ποια πρέπει να παραλειφθούν.

Ένας απλός Γενετικός Αλγόριθμος έχει συγκεκριμένη αλληλουχία εργασιών, η οποία είναι: Κωδικοποίηση, Αρχικοποίηση, Αποκωδικοποίηση, Υπολογισμός συνάρτησης καταλληλότητας (Fitness calculation), Επιλογή, Αναπαραγωγή και τέλος Μετάλλαξη.

Στην συνέχεια λοιπόν θα αναλυθούν οι επιλογές μου σχετικά με τις παραπάνω φάσεις του γενετικού αλγορίθμου για το πρόβλημα το οποίο εξετάζουμε.

A) Κωδικοποίηση

Πρώτο βήμα ενός Γενετικού Αλγορίθμου αποτελεί η **Κωδικοποίηση**. Οι κλασικοί Γενετικοί Αλγόριθμοι χρησιμοποιούν δυαδική κωδικοποίηση, δηλαδή κάθε χρωμόσωμα-individual θα αποτελείται μόνο από bit strings. Γενικά, η δυαδική κωδικοποίηση στους γενετικούς αλγορίθμους έχει αρκετά πλεονεκτήματα. Αρχικά, είναι η πιο εύκολη μορφή κωδικοποίησης όσον αφορά την υλοποίηση της, αφού απλά ορίζεις 0 ή 1 τυχαία για κάθε element του χρωμοσώματος-individual και είναι η πιο συχνά χρησιμοποιούμενη κωδικοποίηση, οπότε υπάρχει και αρκετή βιβλιογραφία σχετικά με αυτήν ώστε να ανατρέξει ο καθένας για πληροφορίες. Είναι αρκετά space-efficient, που είναι και ένας από τους κύριους λόγους που ήταν ίσως η πιο συχνή κωδικοποίηση στις αρχές των γενετικών αλγορίθμων, όπου δεν υπήρχε αρκετή δυνατότητα για μεγάλο όγκο μνήμης. Επίσης, εφόσον θα βρίσκονται τα individuals μας

σε 0-1 κωδικοποίηση, θα είναι αρκετά πιο εύκολο να χειριστούμε τις διαδικασίες της διασταύρωσης (cross over) και της μετάλλαξης (mutation).

Στο πρόβλημα που εξετάζουμε έχουμε 784 pixels για κάθε χρωμόσωμα-individual. Όπως γνωρίζουμε κάθε pixel αποθηκεύεται σαν ένα bit, δηλαδή 0 ή 1, πάντα μιλώντας για greyscale εικόνες όπως του προβλήματος που εξετάζουμε. (0 = μαύρο, 1 = λευκό). Σε τέτοιου είδους προβλήματα η τιμή 0 για ένα pixel σημαίνει την ανενεργή κατάσταση του pixel, ενώ τιμή 1 αντίστοιχα σημαίνει ενεργή κατάσταση.

Οπότε, για το συγκεκριμένο πρόβλημα θα επιλέξω την δυαδική κωδικοποίηση για τους παραπάνω λόγους.

B) Αρχικός πληθυσμός

Κάθε άτομο του πληθυσμού θα έχει με τυχαίο τρόπο 0 ή 1 σε κάθε ένα από τα 784 pixels. Αυτός είναι ο προτιμότερος τρόπος αρχικοποίησης του πληθυσμού για το συγκεκριμένο πρόβλημα.

Γ) Συνάρτηση καταλληλότητας

Για να γίνει η επιλογή μεταξύ των ατόμων, πρέπει να καθοριστεί μία συνάρτηση καταλληλότητας (fitness function) μέσω της οποίας θα βρίσκουμε μία τιμή για κάθε άτομο και μέσω αυτών των τιμών θα βρίσκουμε τα καταλληλότερα. Από την εκφώνηση προτείνεται η εύρεση δύο τιμών καταλληλότητας για κάθε άτομο: την μετρική της ακρίβειας (ποσοστό των εικόνων που ταξινομούνται σωστά) για κάθε άτομο και τον αριθμό των εισόδων. Για την ακρίβεια θέλουμε την μεγαλύτερη ακρίβεια μεταξύ των ατόμων, ενώ επιθυμούμε τον ελάχιστο δυνατό αριθμό εισόδων για κάθε άτομο. Για να επιτευχθεί η μείωση των εισόδων θα εφαρμόσουμε μία ποινή στα άτομα με υψηλό αριθμό εισόδων.

Δ) Γενετικοί Τελεστές

Έχω επιλέξει δυαδική κωδικοποίηση και τώρα είμαστε στο στάδιο επιλογής των τελεστών επιλογής, διασταύρωσης και μετάλλαξης που θα χρησιμοποιηθούν για τα άτομα του πληθυσμού.

I. **Επιλογή (Selection) :** Θα αξιολογήσω τη χρήση των μεθόδων *ρουλέτας με βάση το κόστος, ρουλέτα με βάση την κατάταξη και τουρνουά*.

Ρουλέτα με βάση το κόστος: Γενικά, η πιο συχνή μέθοδος επιλογής ατόμων αποτελεί η μέθοδος εξαναγκασμένης ρουλέτας. Ουσιαστικά, χρησιμοποιείται μια ρουλέτα με σχισμές, όπου το μέγεθος κάθε σχισμής αναλογεί στην απόδοση του ατόμου που αντιστοιχεί στην σχισμή αυτή. Άτομα τα οποία είναι πιο κατάλληλα για επιλογή (μεγαλύτερο fitness value) έχουν μεγαλύτερο μερίδιο στην ρουλέτα και είναι πιο πιθανό να επιλεγθούν. Δεν επιλέγω αυτή την μέθοδο, καθώς έτσι αποφεύγω την εκτεταμένη επανεπιλογή κάποιας υπερβολικά καλής λύσης.

Ρουλέτα με βάση την κατάταξη: Χρησιμοποιείται όταν τα άτομα του πληθυσμού έχουν πολύ κοντινές τιμές καταλληλότητας, άρα σχεδόν ίδια

πιθανότητα να επιλεγθούν και σχεδόν ίδιο μέγεθος σχισμής στην ρουλέτα. Η μέθοδος αυτή έχει το μειονέκτημα πως η επιλογή των γονέων δεν επικεντρώνεται στα άτομα τα οποία έχουν μεγαλύτερα fitness values, με αποτέλεσμα να γίνονται κακές επιλογές γονέων. Αυτό συμβαίνει γιατί η μέθοδος αυτή επικεντρώνεται στην κατάταξη των γονέων ως προς την καταλληλότητα και όχι στην τιμή της συνάρτησης καταλληλότητας. Δεν επιλέγω αυτή τη μέθοδο καθώς είναι προτιμότερη η επόμενη μέθοδος για το πρόβλημα μας.

Τουρνουά: Στην επιλογή μέσω τουρνουά επιλέγονται τυχαία K άτομα από τον πληθυσμό και έπειτα από αυτά τα K άτομα επιλέγονται τα καλύτερα για να γίνουν γονείς. Αυτό επαναλαμβάνεται για την επιλογή κάθε γονέα. Ένα ιδιαίτερο πλεονέκτημα της μεθόδου αυτής είναι ότι δουλεύει και με αρνητικές τιμές καταλληλότητας (fitness values). **Αυτή η μέθοδος αποτελεί την καταλληλότερη** για το πρόβλημα μας, καθώς αυτή η τυχαία επιλογή των ατόμων δίνει την ευκαιρία σε περισσότερες περιπτώσεις διασταύρωσης και βελτίωση των λύσεων. Επίσης, με την μέθοδο αυτή δεν θα επιλέγονται ποτέ τα χειρότερα άτομα από αυτά τα K τυχαία και έτσι ο αλγόριθμος θα συγκλίνει τελικά πιο γρήγορα, αφού θα έχουμε καλές λύσεις πάντα.

II. **Διασταύρωση(crossover) :** Θα αξιολογήσω τη χρήση των μεθόδων *Διασταύρωση μονού σημείου, διασταύρωση πολλαπλού σημείου και ομοιόμορφη διασταύρωση.*

Διασταύρωση μονού σημείου: Στην μέθοδο αυτή, όπως φαίνεται και από το όνομά της, επιλέγεται ένα τυχαίο σημείο διασταύρωσης και στους δύο γονείς. Τα δυαδικά ψηφία στα δεξιά του σημείου αυτού εναλλάσσονται μεταξύ των δύο γονέων. Από την διασταύρωση αυτή προκύπτουν δύο παιδιά που έχουν γενετική πληροφορία και από τους δύο γονείς. Στην μέθοδο αυτή είναι δύσκολο ως αδύνατο να δημιουργηθούν σχήματα υψηλής απόδοσης, με αποτέλεσμα να χάνονται πολύτιμες πληροφορίες από τους γονείς.

Διασταύρωση πολλαπλού σημείου: Στη μέθοδο αυτή επιλέγονται τυχαία K σημεία διασταύρωσης και στους δύο γονείς. Τα δυαδικά ψηφία ανάμεσα στα σημεία αυτά εναλλάσσονται μεταξύ των δύο γονέων. Η μέθοδος αυτή διορθώνει το πρόβλημα της μεθόδου μονού σημείου και επιτρέπει τον συνδυασμό δύο σχημάτων των γονέων. Θα χρησιμοποιήσουμε αυτή την μέθοδο , και συγκεκριμένα με $K=2$, καθώς είναι προτιμότερη από την μονού σημείου και από την ομοιόμορφη, η οποία συγκλίνει ιδιαίτερα αργά.

Ομοιόμορφη διασταύρωση: Στην μέθοδο αυτή κάθε δυαδικό ψηφίο των δύο γονέων έχει ίση πιθανότητα να επιλεγθεί. Συνήθως, κάθε δυαδικό ψηφίο έχει πιθανότητα 50% για το εάν θα υπάρχει στον απόγονο του γονέα. Υπάρχει μία μάσκα διασταύρωσης από δυαδικά ψηφία που αποφασίζει ποιος γονέας συνεισφέρει κάθε δυαδικό ψηφίο στους απογόνους τους. Βέβαια, υπάρχουν και παραλλαγές που προσδίδουν μεγαλύτερη πιθανότητα σε έναν από τους δύο γονείς, ώστε το παιδί να έχει περισσότερη γενετική πληροφορία από το έναν

γονέα. Ουσιαστικά αποτελεί την γενίκευση των άλλων δύο μεθόδων που εξετάσαμε. Συγκλίνει πιο αργά από τις άλλες δύο μεθόδους καθώς ασχολείται με όλα τα δυαδικά ψηφία των γονέων.

III. Μετάλλαξη (mutation) : Θα αξιολογήσω τη χρήση της μεθόδου του ελιτισμού. Γενικά, η μετάλλαξη βοηθάει στην αύξηση της ποικιλομορφία του πληθυσμού. Συνήθως, αποτελεί μία μικρή τυχαία αλλαγή που προκύπτει με μικρή πιθανότητα. Στην δυαδική κωδικοποίηση που εξετάζουμε η πιο συχνή μέθοδος μετάλλαξης αποτελεί η αντιστροφή δυαδικών ψηφίων (τυχαία επιλέγεται κάποιο bit και αντιστρέφεται) . Επειδή στην μετάλλαξη δεν γνωρίζουμε αν θα βοηθήσει για μία καλή λύση το άτομο που θα προκύψει, είναι συνετό να χρησιμοποιούμε τεχνικές που μας επιτρέπουν να καθοδηγήσουμε αυτή την διαδικασία ώστε να προκύψουν άτομα που θα οδηγήσουν σε καλές λύσεις. Εδώ εφαρμόζεται η μέθοδος του **Ελιτισμού**. Ο ελιτισμός μεταφέρει αμέσως στην επόμενη γενιά K άτομα με τις K καλύτερες τιμές της συνάρτησης καταλληλότητας. Με αυτόν τον τρόπο διασφαλίζει ότι το καταλληλότερο άτομο του τωρινού πληθυσμού δεν θα αντικατασταθεί από κάποιο απόγονο με χειρότερη τιμή στην συνάρτηση καταλληλότητας. Βοηθάει στην πιο γρήγορη σύγκλιση του αλγορίθμου, **όμως** στην περίπτωση μας έχουμε μία δυαδική κωδικοποίηση του πληθυσμού, οπότε θα παραμείνουμε σε μία αντίστοιχη μέθοδο μετάλλαξης, την **αντιστροφή δυαδικών ψηφίων**, προσπαθώντας να διατηρήσουμε και την ποικιλομορφία του πληθυσμού.

B2. Υλοποίηση ΓΑ

Η υλοποίηση του ΓΑ πραγματοποιήθηκε με γλώσσα Python και έκδοση 3.7.9. Οι κύριες βιβλιοθήκες που χρησιμοποιήθηκαν ήταν η βιβλιοθήκη **DEAP**, μέσω της οποίας κατασκευάστηκε ο γενετικός αλγόριθμος και βιβλιοθήκες όπως pandas, numpy, matplotlib και tensorflow για τα διάφορα plots και μετρήσεις που ήταν ζητούμενα της άσκησης. Επίσης, χρησιμοποιήθηκε το καλύτερο νευρωνικό δίκτυο της πρώτης εργασίας

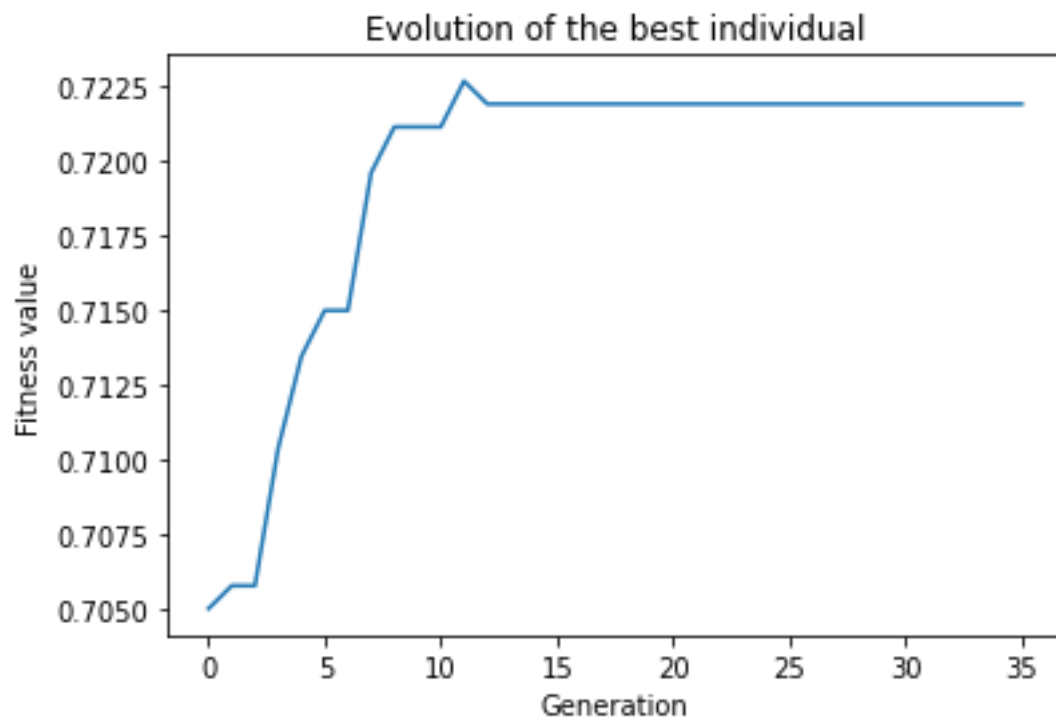
B3. Αξιολόγηση και Επίδραση Παραμέτρων

α) Ακολουθεί ο πίνακας αποτελεσμάτων για την αξιολόγηση των διαφόρων τιμών των παραμέτρων του γενετικού αλγόριθμου:

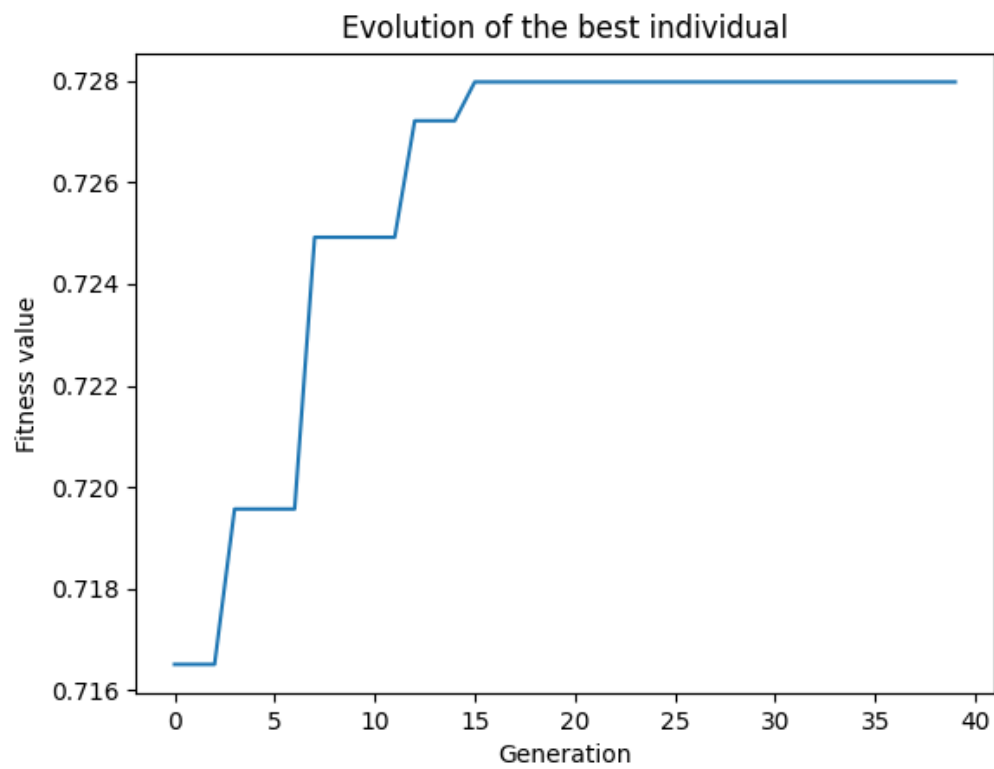
A/A	ΜΕΓΕΘΟΣ ΠΛΗΘΥΣΜΟΥ	ΠΙΘΑΝΟΤΗΤΑ ΔΙΑΣΤΑΥΡΩΣΗΣ	ΠΙΘΑΝΟΤΗΤΑ ΜΕΤΑΛΛΑΞΗΣ	ΜΕΣΗ ΤΙΜΗ ΒΕΛΤΙΣΤΟΥ	ΜΕΣΟΣ ΑΡΙΘΜΟΣ ΓΕΝΕΩΝ
1	20	0.6	0.00	0.7226265	32
2	20	0.6	0.01	0.7291316	65
3	20	0.6	0.10	0.7639530	83
4	20	0.9	0.01	0.7446340	79
5	20	0.1	0.01	0.7122948	85

6	200	0.6	0.00	0.8357275	95
7	200	0.6	0.01	0.8423512	133
8	200	0.6	0.10	0.8212364	96
9	200	0.9	0.01	0.8754213	253
10	200	0.1	0.01	0.8138745	190

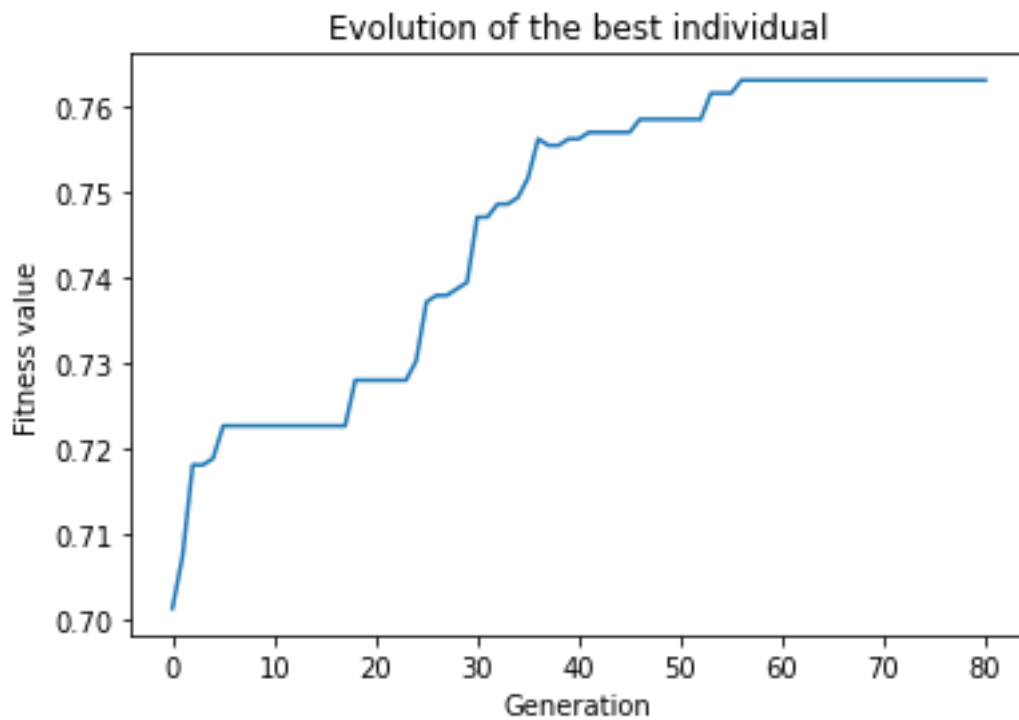
β) Κάθε πείραμα έτρεξε δέκα φορές, καθώς οι γενετικοί αλγόριθμοι είναι στοχαστικοί αλγόριθμοι και σε κάθε εκτέλεση τους δεν είναι σίγουρο ότι θα καταλήξουν στην ίδια απόδοση. Στις εικόνες που ακολουθούν παρουσιάζεται η καμπύλη εξέλιξης της καλύτερης λύσης (της μέσης τιμής της).



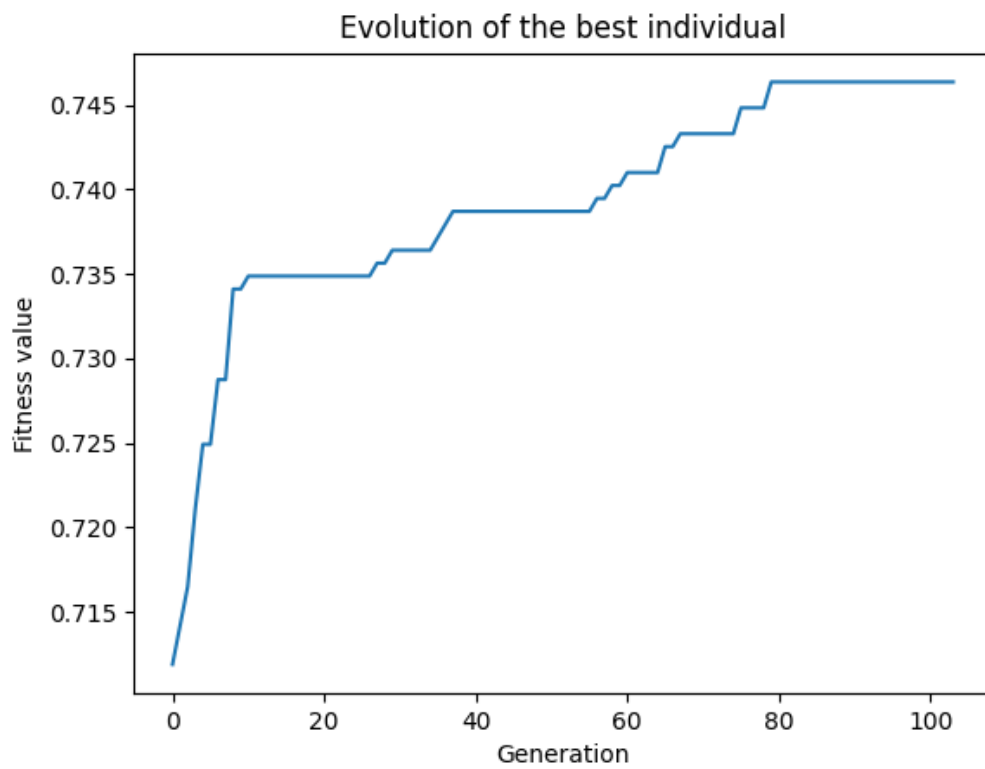
Εικόνα 1/ Population=20, Crossover rate=0.6, Mutation rate=0.0



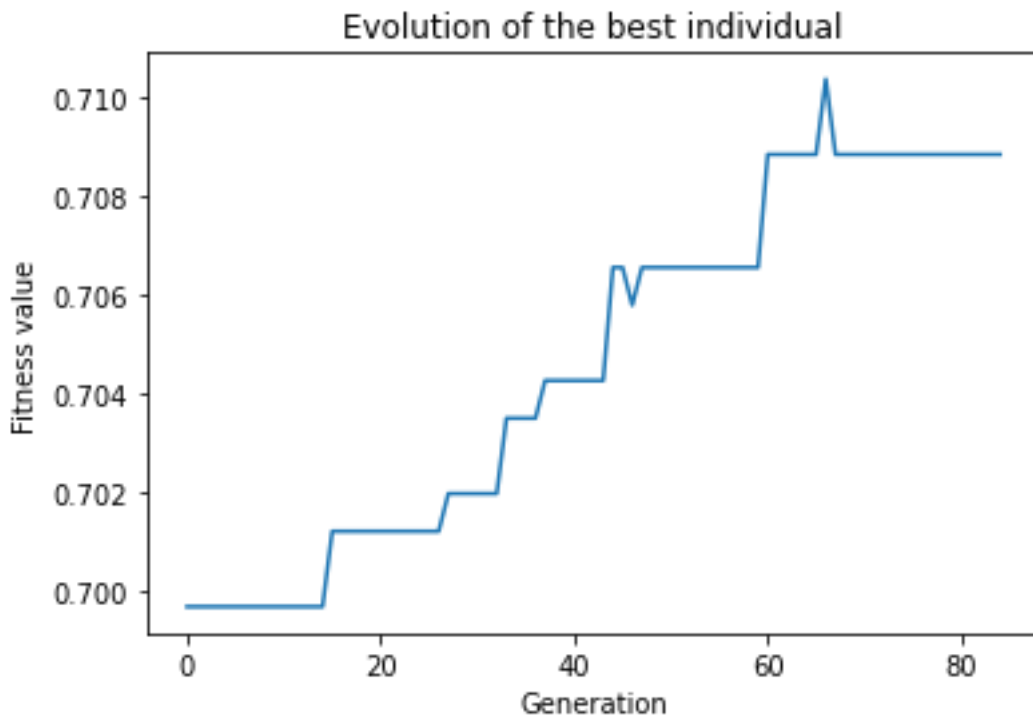
Εικόνα 2/ Population=20, Crossover rate=0.6, Mutation rate=0.01



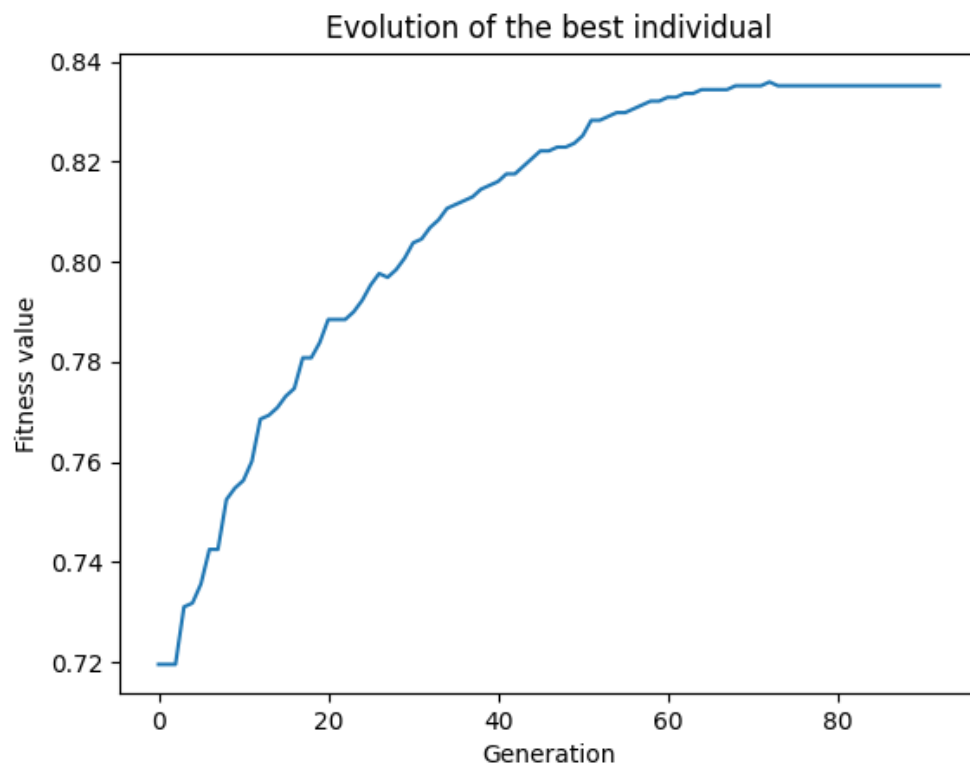
Εικόνα 3/ Population=20, Crossover rate=0.6, Mutation rate=0.1



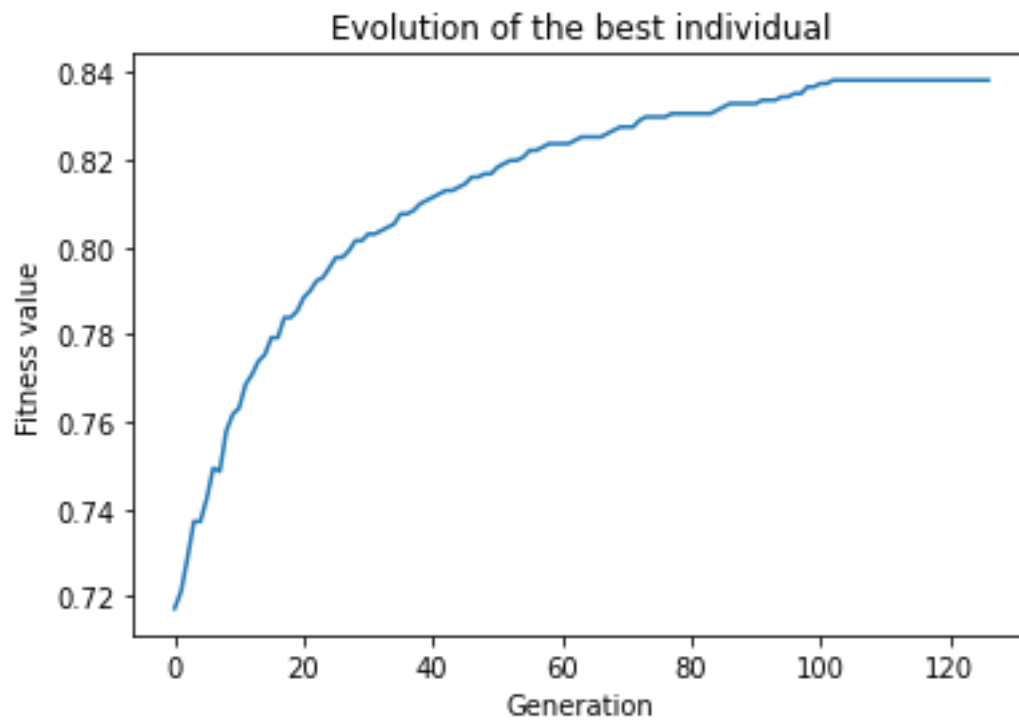
Εικόνα 4/Population=20, Crossover rate=0.9, Mutation rate=0.01



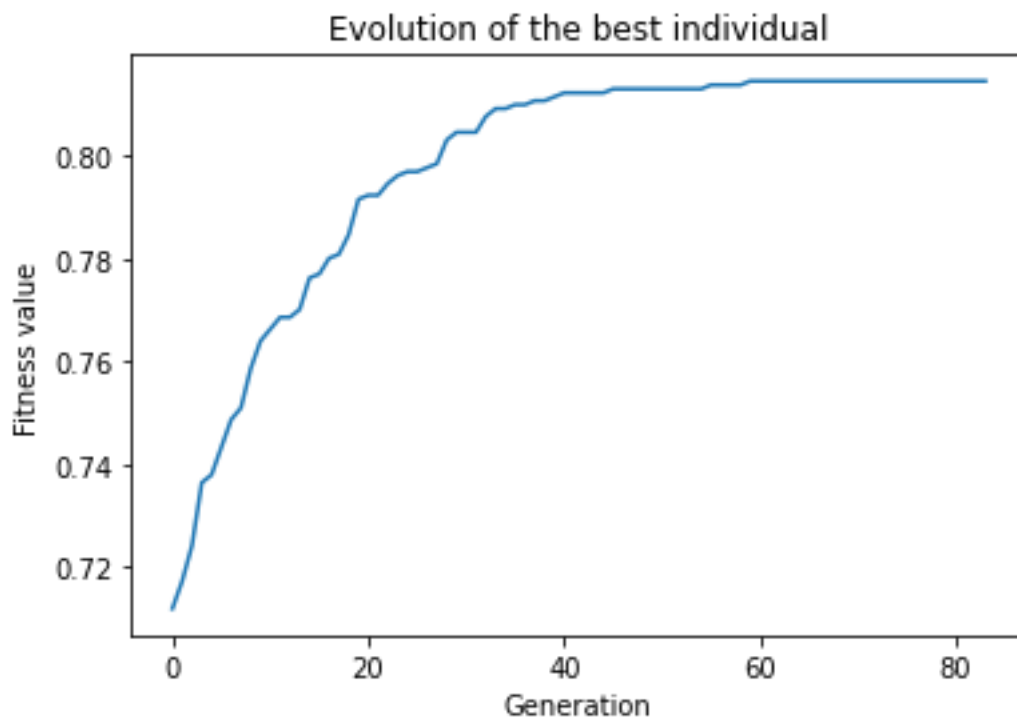
Εικόνα 5/Population=20, Crossover rate=0.1, Mutation rate=0.01



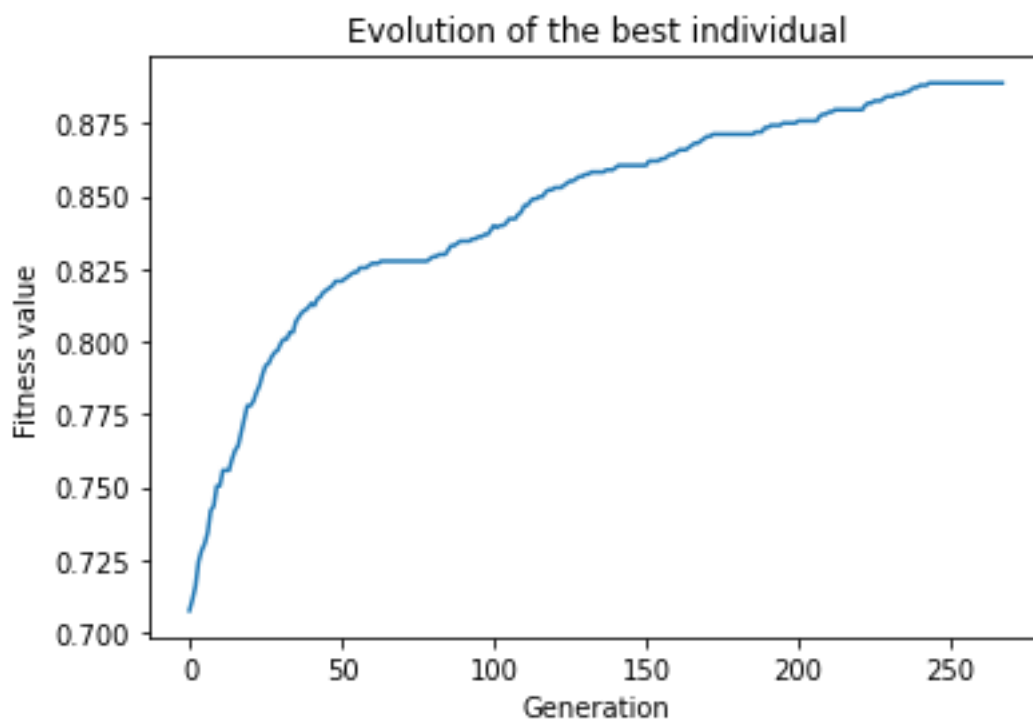
Εικόνα 6/Population=200, Crossover rate=0.6, Mutation rate=0.00



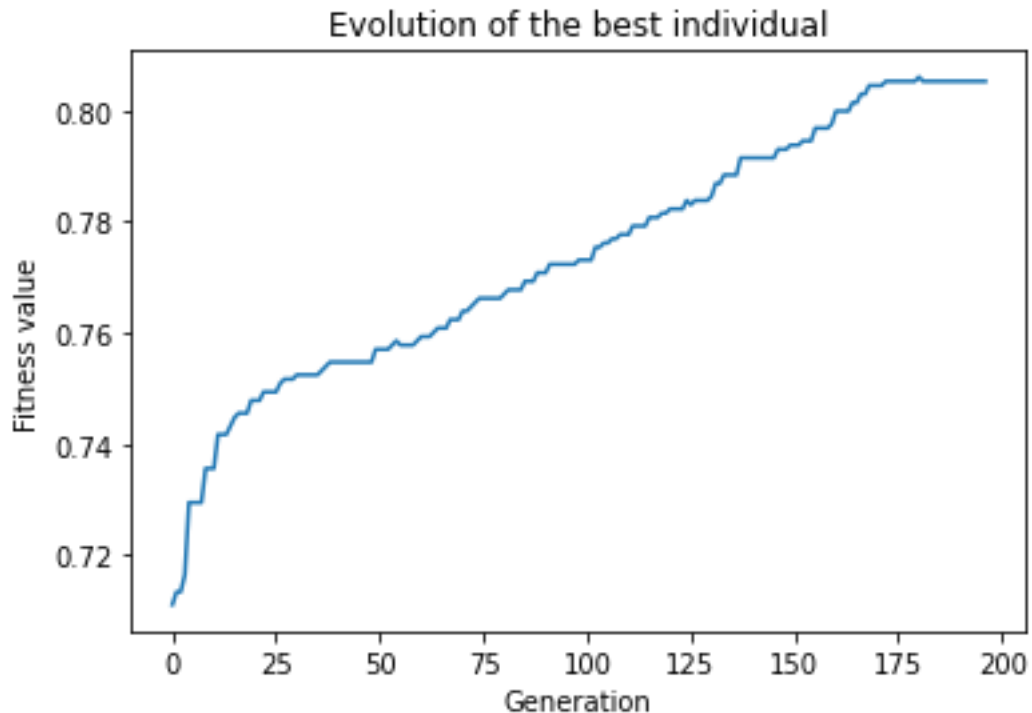
Εικόνα 7/Population=200, Crossover rate=0.6, Mutation rate=0.01



Εικόνα 8/Population=200, Crossover rate=0.6, Mutation rate=0.10



Εικόνα 9/Population=200, Crossover=0.9, Mutation rate=0.01



Εικόνα 10|Population=200, Crossover rate=0.1, Mutation rate=0.01

γ) Ο λόγος που παρατηρούμε τις αποδόσεις των καλύτερων λύσεων και την καμπύλη εξέλιξης της καλύτερης λύσης είναι για να κατανοήσουμε το πόσο επιδράει κάθε παράμετρος (μέγεθος πληθυσμού, πιθανότητα διασταύρωσης και πιθανότητα μετάλλαξης) στη σύγκλιση του γενετικού αλγορίθμου.

Παρατηρώ πως το μέγεθος του πληθυσμού έχει άμεση επίδραση στην σύγκλιση του ΓΑ. Με 200 άτομα πληθυσμό η εξέλιξη παίρνει αρκετά περισσότερες γενιές, που σημαίνει πιο αργή σύγκλιση. Με 200 άτομα όμως εξετάζεται μεγαλύτερος μέρος του χώρου καταστάσεων. Αυτή η πιο αργή και ομαλή σύγκλιση όμως οδηγεί σε καλύτερα αποτελέσματα στην συνάρτηση καταλληλότητας (fitness function). Άρα προτιμάμε το **μεγαλύτερο μέγεθος πληθυσμού** και ας παίρνει σημαντικά περισσότερο χρόνο και υπολογιστική ισχύ για την ολοκλήρωση του ΓΑ.

Όσον αφορά την πιθανότητα μετάλλαξης, παρατηρώ πως όσο πιο μικρή είναι αυτή η πιθανότητα, τόσο μικρότερη είναι η τιμή της συνάρτησης καταλληλότητας και πιο γρήγορη η σύγκλιση. Στην περίπτωση με την μεγαλύτερη πιθανότητα (0.1) μετάλλαξης, παρατηρούνται οι καλύτερες τιμές καταλληλότητας, κάτι που δείχνει πως η μετάλλαξη του πληθυσμού οδήγησε τον ΓΑ στην δημιουργία καινούριων καλών λύσεων (ατόμων), διαφορετικών από των προηγούμενων γενιών. Αυτή η μεγαλύτερη πιθανότητα μετάλλαξης βοηθάει τον ΓΑ να ξεφύγει από τοπικά ακρότητα. Άρα θα την προτιμούσαμε στο πρόβλημα αυτό, αν δεν είχε χειρότερα αποτελέσματα στο μέγεθος πληθυσμού 200. Για τον λόγο αυτό θα επιλέξουμε το mutation rate 0.01 το οποίο έχει καλά αποτελέσματα και στα δύο μεγέθη πληθυσμού.

Τέλος, για την πιθανότητα διασταύρωσης παρατηρείται κάτι αντίστοιχο με την πιθανότητα μετάλλαξης. Η μικρότερη τιμή (0.1) οδηγεί στα χειρότερα αποτελέσματα

στην συνάρτηση καταλληλότητας. Αυτό είναι λογικό, καθώς με τόσο μικρή πιθανότητα διασταύρωσης, δεν δημιουργούνται αρκετά νέα καλά άτομα και καταλήγουμε με άτομα σχεδόν ίδια ή και ίδια με τους γονείς τους. Με 0.6 και 0.9 πιθανότητα διασταύρωσης έχουμε πολύ καλά αποτελέσματα, με το 0.9 να υπερτερεί του 0.6 σε συνδυασμό με το μέγεθος πληθυσμού 200. Μπορεί να χρειάζεται σχεδόν διπλάσιες εποχές από την 0.6 πιθανότητα διασταύρωσης, αλλά καταφέρει καλύτερα αποτελέσματα στην συνάρτηση καταλληλότητας.

Για αυτό θα επιλέξουμε τον συνδυασμό 200, 0.9, 0.01 ώστε να συγκρίνουμε την απόδοση του βέλτιστου ΤΝΔ με το βέλτιστο ΤΝΔ της Α εργαστηριακής άσκησης.

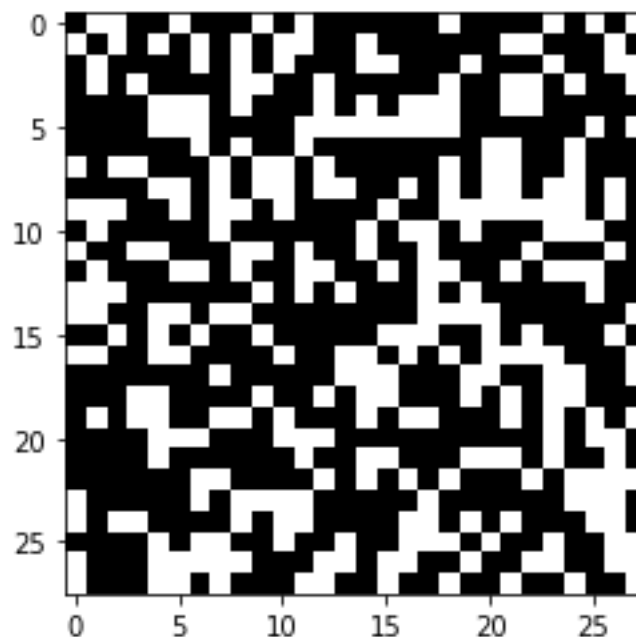
B4. Αξιολόγηση ΤΝΔ

α) Όσον αφορά το βέλτιστο ΤΝΔ του Α πρότζεκτ, αποδείξαμε πως η γενικευτική του ικανότητα ήταν αρκετά καλή και δεν προέκυπτε πρόβλημα υπερπροσαρμογής (overfitting) στα δεδομένα ελέγχου. Η απόδοση του όσον αφορά την ακρίβεια ήταν γύρω στο 98%.

Το βέλτιστο ΤΝΔ που προκύπτει από τα παραπάνω πειράματα καταφέρει απόδοση 90%, το οποίο δείχνει πως η μείωση των χαρακτηριστικών είχε σημαντική επίδραση στην απόδοση του δικτύου.

Η εικόνα είναι η παρακάτω και παρατηρούμε πως τα pixels τα οποία αποκόπτονται είναι κυρίως από το περίγραμμα της εικόνας. Αυτό είναι λογικό καθώς στο dataset μας τα χειρόγραφα ψηφία είναι σχεδιασμένα συνήθως στο κέντρο της εικόνας.

Τέλος, δεν μπορούμε να καταλάβουμε αν υπάρχει overfitting στα δεδομένα ελέγχου μόνο μέσω αυτού του ελέγχου.



β) Αφού το δίκτυο εκπαιδεύτηκε ξανά με όλο το σύνολο εκπαίδευσης και έγινε ξανά η σύγκριση του α ερωτήματος, προέκυψε ακρίβεια 97%, δηλαδή πάρα πολύ κοντά στο

βέλτιστο ΤΝΔ του Α πρότζεκτ. Αυτό μας δείχνει πως το νέο ΤΝΔ που προέκυψε από τον ΓΑ έχει σχεδόν την ίδια απόδοση, αλλά με λιγότερα ενεργοποιημένα χαρακτηριστικά και άρα χρειάζεται και μικρότερη υπολογιστική ισχύ και χρόνο για την εκτέλεση του. Η γενικευτική του ικανότητα είναι εφάμιλλη με το άλλο βέλτιστο ΤΝΔ. Αντίστοιχα, overfitting δεν παρουσιάζεται ούτε πια με τα αρκετά λιγότερα χαρακτηριστικά.