



ΠΑΝΕΠΙΣΤΗΜΙΟ
ΠΑΤΡΩΝ
UNIVERSITY OF PATRAS

**ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΗΛΕΚΤΡΟΝΙΚΩΝ
ΥΠΟΛΟΓΙΣΤΩΝ ΚΑΙ ΠΛΗΡΟΦΟΡΙΚΗΣ**

**Αναφορά Εργαστηριακής Άσκησης
Μέρος Β΄
Υπολογιστική Νοημοσύνη**

Κωνσταντίνος Τσάκωνας
Α.Μ.: 1059666

Ακαδημαϊκό έτος 2020-21
Εαρινό Εξάμηνο

Περιεχόμενα

Repository Κώδικα

Για την ανάπτυξη της άσκησης χρησιμοποιήθηκε η γλώσσα Python με τις βιβλιοθήκες Tensorflow, numpy, pandas, dear και matplotlib. Παρακάτω υπάρχει το repository του κώδικα στο github

[github link](#)

B1. Σχεδιασμός ΓΑ

1. Τα άτομα του αρχικού πληθυσμού θα αναπαραστηθούν ως δυαδικές συμβολοσειρές. Ο λόγος που θα ακολουθηθεί αυτή η κωδικοποίηση προέρχεται από το σκεπτικό ότι, αυτό που θέλουμε να κάνουμε είναι να μειώσουμε το εισόδοι από τα 784 pixels, δηλαδή να μηδενισούμε πολλές από αυτές. Δημιουργώντας λοιπόν άτομα του πληθυσμού ως πίνακες 784 pixels που περιέχουν δυαδικά στοιχεία, μπορούμε με ένα πολλαπλασιασμό element-wise να κρατήσουμε τις εισόδους που θέλουμε.
2. Ο αρχικός πληθυσμός θα είναι N τυχαίοι πίνακες (τυχαία δειγματοληψία) 28x28 και οι τιμές που θα περιέχουν θα είναι 0 και 1.
3. Η συνάρτηση καταλληλότητας που επιλέχθηκε είναι οι εξής:
 - κάθε φορά που θα γίνεται έλεγχος στο νευρωνικό σύμφωνα με τις εισόδους που προκύπτουν από κάθε άτομο του πληθυσμού θα κάνουμε ταξινόμηση των πρώτων 10.000 εικόνων και θα συγκρίνουμε την ταξινόμηση αυτή με βάση τα labels για να δούμε πόσο ακριβής είναι. Το αποτέλεσμα που θα προκύπτει από αυτό θα είναι μια τιμή μεταξύ του διαστήματος $[0, 1]$ και στόχος του γενετικού αλγορίθμου θα είναι να πλησιάσει όσο το δυνατό πιο κοντά στο 1.
 - Επίσης θα επιβάλεται μία ποινή σε άτομα του πληθυσμού που έχουν μεγάλο αριθμό εισόδων, συγκεκριμένα άτομα που έχουν περισσότερες από 300 εισόδους. Η τιμή η οποία θα αφαιρείται είναι ανάλογη με το σφάλμα της ταξινόμησης δια εκατό επί το πλήθος των παραπάνω εισόδων που έχει σε σχέση με αυτές που έχουμε θέσει ως επιθυμητό άνω όριο. Δηλαδή:

$$f(i) = acc_i - \frac{loss_i}{100} * (inputs_i - limit)$$

Ο λόγος που επιλέχθηκε η μεγιστοποίηση του ποσοστού ταξινόμησης έχει να κάνει με το γεγονός ότι κατά την δημιουργία και την εκπαίδευση

του μοντέλου στην προηγούμενη εργαστηριακή άσκηση σε πάρα πολύ μικρές τιμές του loss η ταξινόμηση δεν ήταν πάντα σωστή και δεν συμβάδιζε με το accuracy του μοντέλου. Ακόμα ο λόγος που εφαρμόζουμε ποινή σε άτομα που έχουν περισσότερες από 392 εισόδους είναι γιατί θα έχουμε μία μείωση 50% αλλά για λιγότερες εισόδους από αυτές θα χάνουμε χαρακτηριστικά που χρειαζόμαστε αφού το σχήμα των ψηφιών στις εικόνες συνεχώς μεταβάλλεται, με αποτέλεσμα να μην έχουμε καλή ταξινόμηση.

4. Γενετικοί Τελεστές:

(α') Επιλογή:

- Ρουλέτα βάσει κόστους: Με αυτή την μέθοδο από N άτομα επιλέγουμε με βάση μία πιθανότητα p_i τα άτομα τα οποία θα γίνει η διασταύρωση. Η πιθανότητα υπολογίζεται ως εξής:

$$p_i = \frac{f_i}{\sum_{i=1}^N f_i}$$

όπου,

f_i : το fitness του i -οστού ατόμου,

N : το μέγεθος του πληθυσμού.

Κατα αυτό τον τρόπο τα άτομα με το μεγαλύτερο fitness έχουν μεγαλύτερη πιθανότητα να επιλεγούν, δηλαδή βαρύνουν περισσότερο.

- Ρουλέτα βάσει κατάταξης: Με την μέθοδο αυτή τα άτομα κατατάσσονται με σύμφωνα με την καταλληλότητα σε αύξουσα σειρά. Στο άτομα με τη μικρότερη καταλληλότητα ανατίθεται κατάταξη 1, στο αμέσως επόμενο 2 και η διαδικασία συνεχίζεται μέχρι το καταλληλότερο άτομα να έχει κατάταξη N , όπου N το μέγεθος του πληθυσμού. Το μέγεθος που πιάνει το άτομα πάνω στη ρουλέτα σε τοις εκατό προκύπτει από:

$$\frac{r_i}{\sum_{i=1}^N r_i} \times 100$$

όπου,

r_i : η κατάταξη του i -οστού ατόμου,

N : το μέγεθος του πληθυσμού.

Σύμφωνα με το παραπάνω τα άτομα βαρύνουν με τον ίδιο τρόπο ανεξαρτήτως της μεγάλης απόκλισης που μπορεί να έχει η συνάρτηση καταλληλότητας τους.

- Τουρνουά: Με την μέθοδο αυτή δημιουργούμε ένα τουρνουά μεγέθους K . Σε κάθε τουρνουά διαλέγουμε τυχαία K άτομα, από αυτά επιλέγεται αυτό με το καλύτερο fitness.

Από αυτές αυτή που επιλέχθηκε για τον αλγόριθμο μας σύμφωνα με το πρόβλημα που έχουμε να επιλύσουμε είναι η μέθοδος του Τουρνουά. Ο λόγος που επιλέχθηκε βασίζεται στο γεγονός ότι τα άτομα που θα επιλεγούν για να διασταυρωθούν θα έχουν μεγαλύτερες διαφοροποιήσεις μεταξύ τους, διότι επιλέγονται K άτομα τυχαία. Στις υπόλοιπες μεθόδους οι επιλογή γίνεται με βάση κάποια πιθανότητα από τον αρχικό πληθυσμό με αποτέλεσμα τα άτομα με καλύτερη καταλληλότητα να είναι πιο πιθανό να επιλεγούν. Άρα με την χρήση του τουρνουά και την ποικιλομορφία που προσφέρει στη διασταύρωση μπορεί να προκύψει κάποιο άτομο με πολύ καλύτερο fitness από το καλύτερο άτομο της προηγούμενης γενιάς με αποτέλεσμα ο αλγόριθμος να συγκλίνει πιο γρήγορα.

(β') Διασταύρωση:

- Διασταύρωση μονού σημείου: Κατά την διασταύρωση μονού σημείου δημιουργούνται ζευγάρια ατόμων από αυτά που επιλέχθηκαν και επιλέγεται τυχαία ένας ακέραιος αριθμός N . Τα άτομα είναι μήκους M και ισχύει ότι $0 \leq N \leq M - 1$. Το N είναι το σημείο που θα γίνει η διασταύρωση των ατόμων, δηλαδή θα γίνει ανταλλαγή των δυαδικών ψηφίων N μέχρι $M - 1$.
- Διασταύρωση διπλού σημείου: Η μέθοδος είναι ίδια με την παραπάνω με την διαφορά ότι επιλέγονται δύο ακεραίοι αντί ενός και η διασταύρωση θα γίνει μεταξύ των δυαδικών ψηφίων που περιέχονται εντός των ορίων που θέτουν οι ακεραίοι που επιλέχθηκαν.
- Ομοιόμορφη διασταύρωση: Για αυτού του τύπου την διασταύρωση δημιουργείται μια φόρμα από δυαδικά ψηφία μήκους ίδιου με του ατόμου. Στα σημεία που η φόρμα έχει τον ψηφίο 1 το πρώτο παιδί παίρνει το αντίστοιχο ψηφίο του δεύτερου γονέα, ενώ το δεύτερο παιδί του πρώτου γονέα. Όταν το ψηφίο της φόρμας είναι 0 το πρώτο παιδί παίρνει το αντίστοιχο ψηφίο το πρώτου γονέα ενώ το δεύτερο παιδί του δεύτερου γονέα.

Στον αλγόριθμο μας επιλέξαμε διασταύρωση διπλού σημείο. Εστιάζοντας στο πρόβλημα που έχουμε, πρέπει να αφαιρέσουμε pixels

από μία εικόνα αλλά ταυτοχρόνως να μην αφαιρέσουμε σημαντική πληροφορία για τον νευρωνικό μας. Αν χρησιμοποιούσαμε την διασταύρωση μονού σημείου ή την ομοιόμορφη οι απόγονοι κατά το πέρας των γενεών θα περιορίζονταν, δηλαδή θα αφαιρούσαν pixels από κοντινές περιοχές και αυτό θα μπορούσε να οδηγήσει σε πολύ χαμηλής ποιότητας λύση αφού αν παρατηρήσουμε τις εικόνες που έχουμε στο σύνολο δεδομένων μας θα δούμε ότι μας ενδιαφέρει κυρίως είναι να αφαιρέσουμε το περίγραμμα και να κρατήσουμε τα κεντρικά pixels διότι αυτά είναι που απεικονίζουν τον αριθμό που θέλουμε να ταξινομηθεί. Αν χρησιμοποιούσαμε ομοιόμορφη διασταύρωση θα είχαμ πολλαπλά σημεία κοπής στους γονείς με αποτέλεσμα να ανακατεύεται το γενετικό υλικό και να χανουμε χρήσιμη πληροφορία. Αν χρησιμοποιούσαμε διασταύρωση μονού σημείου μετά από κάποιες γενεές οι απόγονοι θα μοιάζουν πολύ με τους γονείς με αποτέλεσμα η καλύτερη λύση να είναι κάποιο τοπικό ακρότατο.

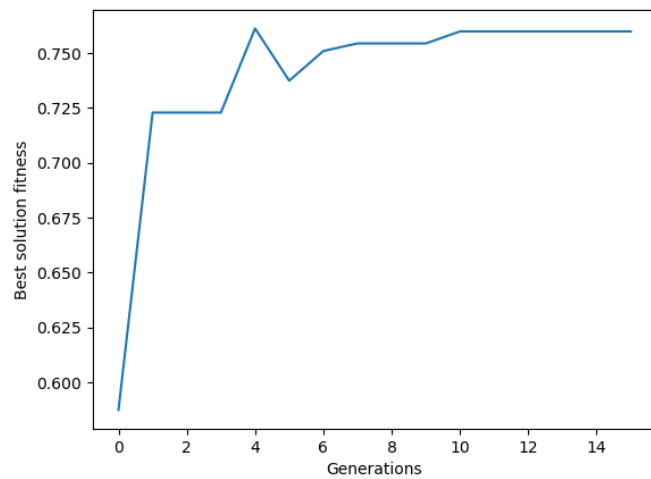
(γ') Μετάλλαξη: Η μετάλλαξη βοηθάει τον αλγόριθμό να εξερευνήσει λύσεις σε όλο το επίπεδο λύσεων. Ο ελιτισμός ουσιαστικά περνάει τα άτομα με το καλύτερο fitness αμετάλαχτα στην επόμενη γενιά. Επιλέξαμε να μην χρησιμοποιήσουμε τον ελιτισμό διότι η μετάλλαξη όπως είπαμε βοηθάει να γίνει εξερεύνηση περισσότερων λύσεων και επίσης σε περίπτωση που έχουμε μικρό πληθύσμο θα εποφεληθεί πολύ από αυτό. Σύμφωνα με το δικό μας πρόβλημα επειδή χρησιμοποιούμε 10.000 εικόνες από τις 60.000, ο ελιτισμός μπορεί να οδηγήσει σε υπερπροσαρμογή την βέλτιστη λύση πάνω σε αυτές τις 10.000 εικόνες και να μειώσει την γενικευτική ικανότητα του δικτύου.

B3. Αξιολόγηση και επίδραση παραμέτρων

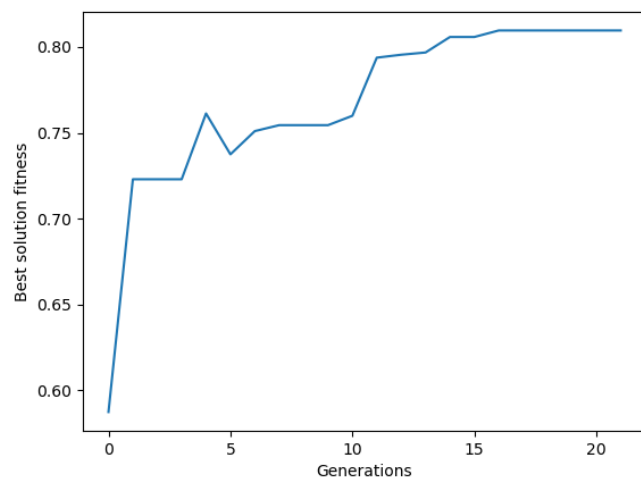
1. Πίνακς αποτελεσμάτων

| A/A | ΜΕΓΕΘΟΣ ΠΛΗΘΥΣΜΟΥ | ΠΙΘΑΝΟΤΗΤΑ ΔΙΑΣΤΑΥΡΩ- ΣΗΣ | ΠΙΘΑΝΟΤΗΤΑ ΜΕΤΑΛΛΑΞΗΣ | ΜΕΣΗ ΤΙΜΗ ΒΕΛΤΙΣΤΟΥ | ΜΕΣΟΣ ΑΡΙΘΜΟΣ ΓΕΝΕΩΝ |
|------------|------------------------------|--|----------------------------------|--------------------------------|-------------------------------------|
| 1 | 20 | 0.6 | 0.00 | 0.739 | 8.5 |
| 2 | 20 | 0.6 | 0.01 | 0.767 | 11.5 |
| 3 | 20 | 0.6 | 0.10 | 0.798 | 12.5 |
| 4 | 20 | 0.9 | 0.01 | 0.808 | 11.5 |
| 5 | 20 | 0.1 | 0.01 | 0.688 | 5 |
| 6 | 200 | 0.6 | 0.00 | 0.886 | 18 |
| 7 | 200 | 0.6 | 0.01 | 0.905 | 27 |
| 8 | 200 | 0.6 | 0.10 | 0.913 | 25.5 |
| 9 | 200 | 0.9 | 0.01 | 0.914 | 24.5 |
| 10 | 200 | 0.1 | 0.01 | 0.841 | 29.5 |

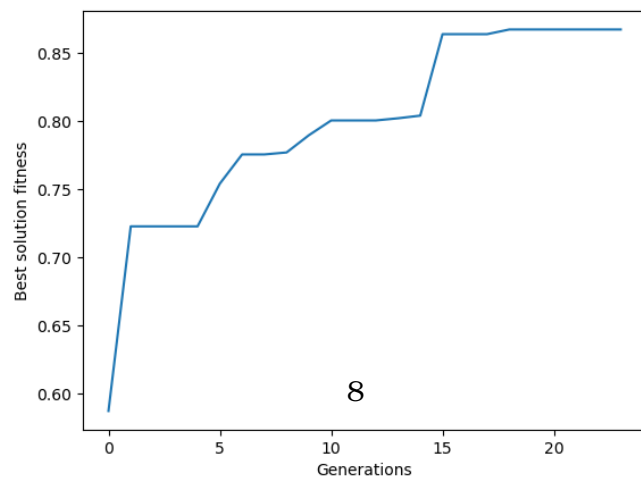
2. Καμπύλες εξέλιξης:



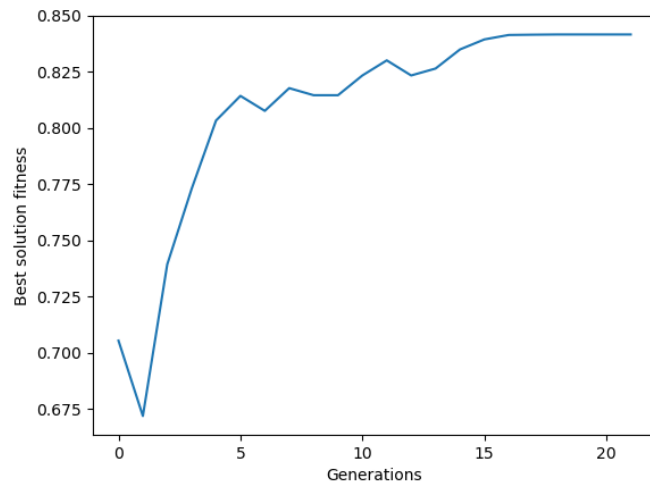
Περίπτωση: 20, 0.6 ,0.00



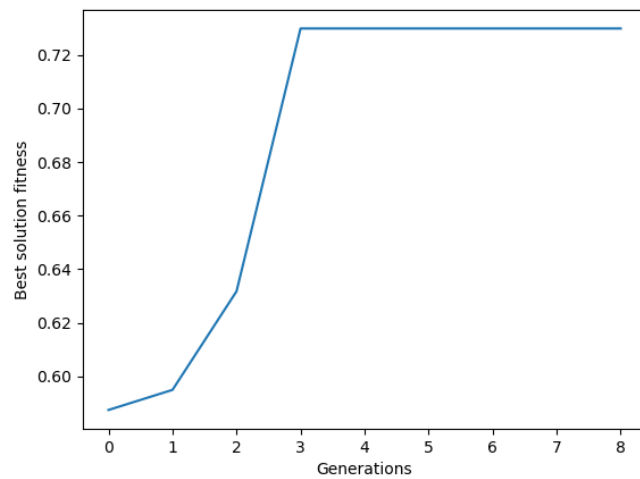
Περίπτωση: 20, 0.6 ,0.01



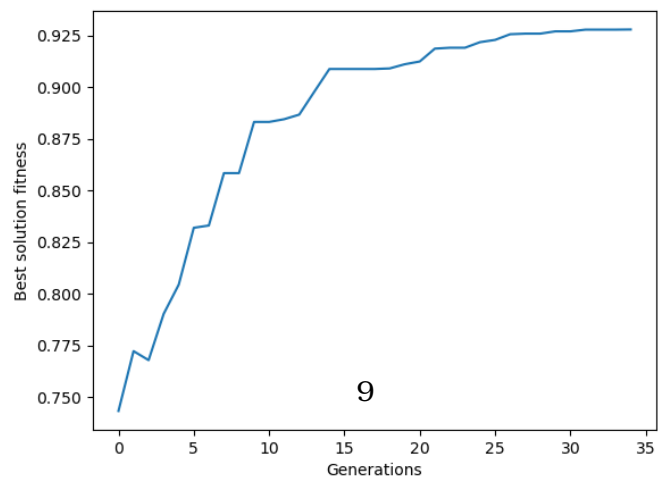
Περίπτωση: 20, 0.6 ,0.10



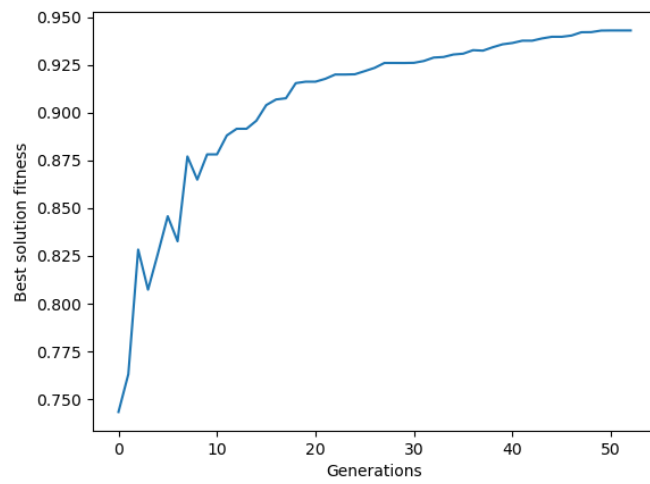
Περίπτωση: 20, 0.9 ,0.01



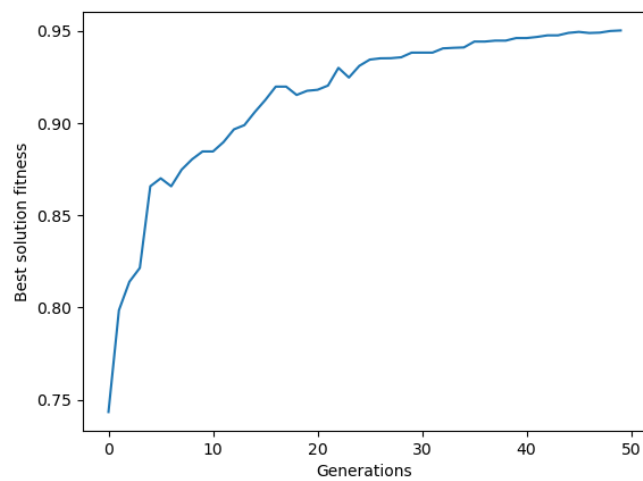
Περίπτωση: 20, 0.1 ,0.01



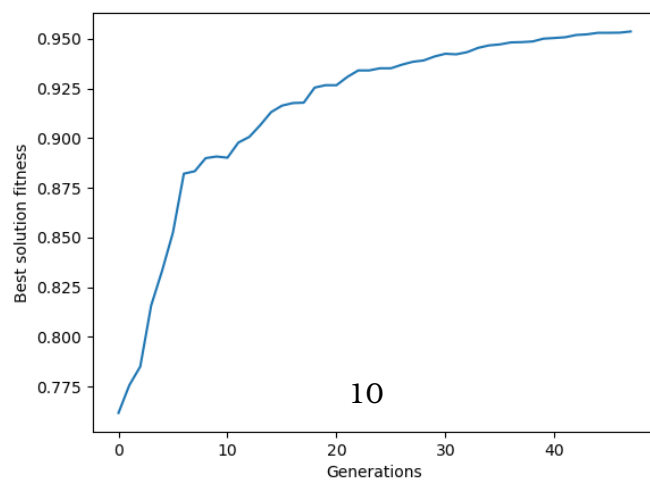
Περίπτωση: 200, 0.6 ,0.00



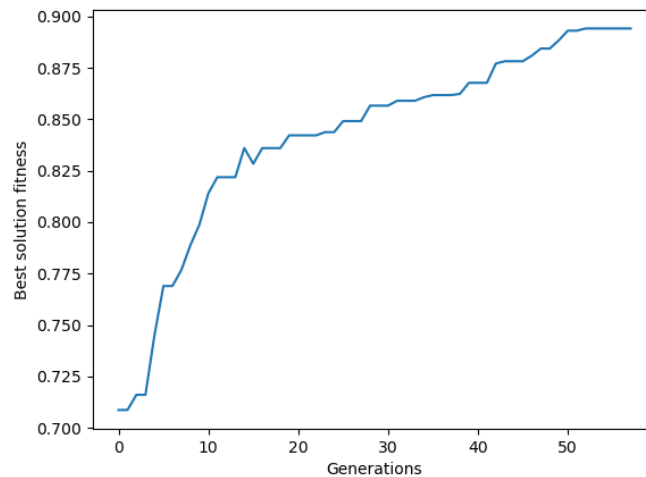
Περίπτωση: 200, 0.6 ,0.01



Περίπτωση: 200, 0.6 ,0.10



Περίπτωση: 200, 0.9 ,0.01



Περίπτωση: 200, 0.1 ,0.01

3. Συμπεράσματα: Σύμφωνα με τα αποτελέσματα που φαίνονται στον πίνακα και τις γραφικές παραστάσεις, παρατηρούμε αρχικά στις περιπτώσεις που ο αρχικός πληθυσμός είναι 20 έχουμε μικρή καταλληλότητα για την καλύτερη λύση σε σχέση με πληθυσμό 200 αλλά και ότι η εξέλιξη σταμάτησε σε μικρό αριθμό γενεών πράγμα που σημαίνει ότι η καλύτερη λύση σταμάτησε να βελτιώνεται. Το παραπάνω είναι λογικά αφού λόγω του μειωμένου πληθυσμού έχουμε λιγότερες πιθανότητες να γίνει η διασταύρωση οπότε δεν θα εξεταστεί μεγάλο πλήθος λύσεων. Σε περίπτωση που δεν είχαμε τα κριτήρια σταματήματος για την εξέλιξη με πληθυσμό 20 ίσως βρισκάμε κάποια καλύτερη καταλληλότητα μετά από έναν μεγάλο αριθμό γενεών αλλά υπάρχει μεγάλη πιθανότητα αυτή η λύση να μην είναι η βέλτιστη αλλά κάποιο τοπικό βέλτιστο. Στο πληθυσμό 200 βλέπουμε ότι τα αποτελέσματα είναι καλύτερα διότι υπάρχει μεγαλύτερη ποικιλομορφία στα άτομα και περισσότερες πιθανότητες διασταύρωσης, οπότε η εξέλιξη συνεχίζει για αρκετές γενεές. Παρά όμως το γεγονός ότι η καλύτερη λύση έχει πολύ υψηλή καταλληλότητα, ο χρόνος εκτέλεσης ήταν αρκετά υψηλός.

Στο πίνακα παρατηρούμε ότι η πιθανότητα διασταύρωσης όσο μεγαλύτερη είναι τόσο καλύτερη είναι η τιμή του fitness της καλύτερης λύσης αλλά και η εξέλιξη ολοκληρώνεται γρηγορότερα. Άρα, εξαιτίας της μεγαλύτερης πιθανότητας διασταύρωσης τα άτομα διασταυρώνονται πιο συχνά οπότε η απόγονοι είναι διαφορετικοί από τους γονείς που ισοδύναμι σε μία καινούργια λύση. Όταν η πιθανότητα διασταύρωσης είναι μικρή βλέπουμε ότι η καλύτερη λύση έχει πολύ μικρή καταλληλότητα

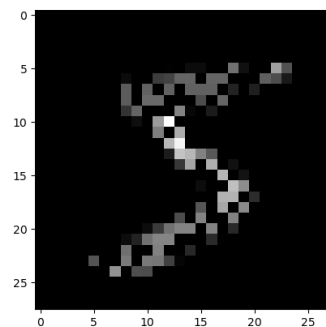
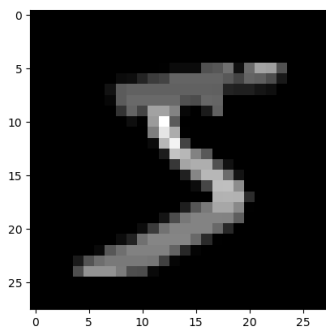
διότι οι απόγονοι μοιάζουν πολύ με τους γονείς τους. Τέλος βλέπουμε ότι στον πίνακα ότι η μέση τιμή του βέλτιστου στις περιπτώσεις που η πιθανότητα μετάλλαξης δεν είναι μηδενική ο μέσος αριθμός γενεών αυξάνεται και στην περίπτωση που η τιμή της είναι 0.1 έχουμε την καλύτερη μέση τιμή του βέλτιστου. Αυτό συμβαίνει διότι με την μετάλλαξη δίνεται η δυνατότητα στον πληθυσμό να εξερευνήσει μεγαλύτερο πλήθος λύσεων και αποτρέπει τα χρωμοσώματα να γίνουν ίδια μετά από κάποιες γενιές εξέλιξης και να κολλήσουν σε κάποιο τοπικό ακρότατο.

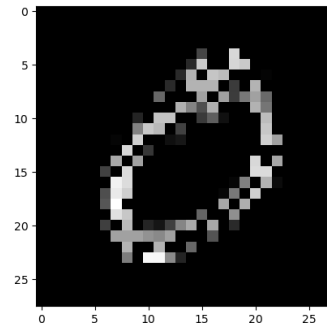
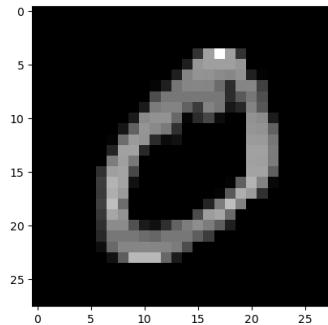
B4. Αξιολόγηση ΤΝΔ

1.)

(α') Η καλύτερη λύση του γενετικού αλγορίθμου μας έδωσε accuracy στα άγνωστα δεδομένα 0.95 ενώ το μοντέλο από την εργασία Α 0.98. Αυτό σημαίνει ότι η καλύτερη λύση του γενετικού μας αλγορίθμου δεν έχει καλύτερη γενικευτική ικανότητα αν και η απόκλιση είναι πολύ μικρή. Ο λόγος που συμβαίνει αυτό είναι που ο γενετικός μας δεν έχει εκπαιδευτεί για όλο το σύνολο δεδομένων του train αλλά σε ένα μέρος αυτού(10.000). Η περίπτωση που θα έχει εκπαιδευτεί σε όλο θα εξεταστεί αργότερα.

(β') Παρακάτω παρουσιάζονται δύο εικόνες πριν και μετά την εφαρμογή της επιλογής χαρακτηριστικών σύμφωνα με την καλύτερη λύση:



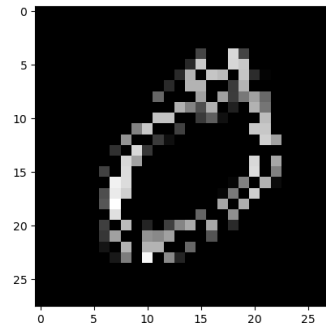
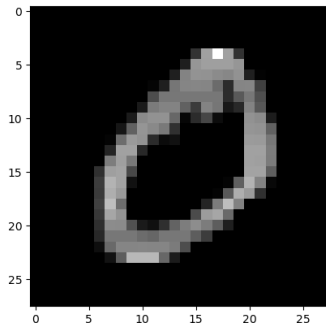
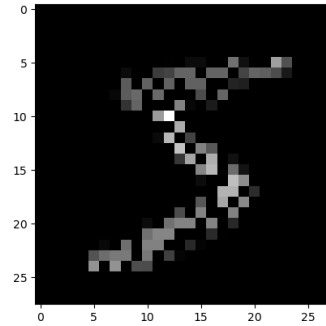
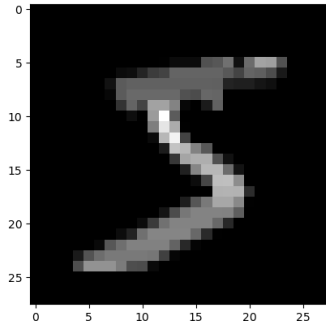


Παρατηρούμε ότι υπάρχουν κάποιες αφαιρέσεις εικονοστοιχίων από το περίγραμμα του αριθμού το οποίο είναι ανμενόμενο. Υπάρχουν όμως και σημεία μέσα πάνω στον αριθμό που έχουν αφαιρεθεί pixels, διότι αρχικά οι γενετικοί είναι καθαρά μια στοχαστική διαδικασία οπότε δεν μπορούμε να διασφαλίσουμε τη πιο βέλτιστη λύση το μόνο που μπορούμε είναι να βρούμε κάποια πολύ κοντά σε αυτή.

- (γ') Το γεγονός που έχουν αφαιρεθεί εικονοστοιχία πανω στον αριθμό δείχνει κάποια υπερπροσαρμογή στα δεδομένα που εκπαιδεύτηκε ο γενετικός μας αλγόριθμος. Διότι μπορεί να υπήρχαν εικόνες που στα σημεία αυτά να μην περιείχαν τίποτα ή η είσοδος που είχαν να μην επηρέαζε το τελικό αποτέλεσμα οπότε αφαιρέθηκε.

2.)

- (α') Μετά την εξέλιξη του αλγορίθμου σε όλο το σύνολο των δεδομένων και τον έλεγχο στο άγνωστο σύνολο(mnist_test) τα αποτελέσματα ήταν ίδια με το προηγούμενο σε ότι αφορά την γενικευτική ικανότητα του δικτύου.
- (β') Παρακάτω παρουσιάζονται οι ίδιες εικόνες πριν και μετά την εφαρμογή της επιλογής χαρακτηριστικών :



παρατηρείται ότι οι εικόνες και στις δύο διαφορετικές περιπτώσεις είναι σχεδόν ίδιες, Αυτό μας οδηγεί στο συμπέρασμα ότι η βέλτιστη λύση και στις δύο περιπτώσεις είναι ίδια ανεξαρτήτως του αριθμού των εικόνων. Οι μικρές διαφορές που υπάρχουν οφείλονται στη στοχαστική φύση του αλγορίθμου αλλά σύμφωνα με το γεγονός ότι δεν αλλάζει η ακρίβεια του νευρωνικού σημαίνει ότι οι διαφορές που υπάρχουν δεν ήταν σε pixels που παρείχαν πολύτιμη πληροφορία.

- (γ') Εφόσον εξεταστικές και αυτή η περίπτωση και τα αποτελέσματα είναι ίδια, η υπερπροσαρμογή είναι αρκετά απίθανη και αυτό είναι λογικό αν αναλογιστούμε τα δεδομένα που έχουμε. Το dataset είναι αρκετά optimized για τέτοιου είδους προβλήματα. Ουσιαστικά οι εικόνες κατανέμονται ομοιόμορφα μέσα στο dataset αρά στην πρώτη περίπτωση που ο αλγόριθμος έτρεξε για μικρότερο πλήθος δεδομένων αρκούσε ώστε να δει αρκετές εικόνες και να καταλήξει σε αυτή τη λύση.