



ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Playing with MNIST

Συγγραφέας : Δέκας Δημήτριος
Αριθμός Ειδικού Μητρώου : 3063

23 Δεκεμβρίου 2020

Καθηγητής : Τέφας Αναστάσιος

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

© 2020 Δέκας Δημήτριος

Created with L^AT_EX

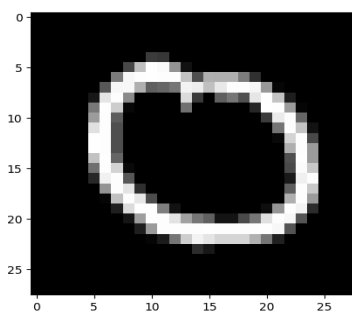
0 1 2 3 4 5 6 7 8 9 10

1. Περίληψη

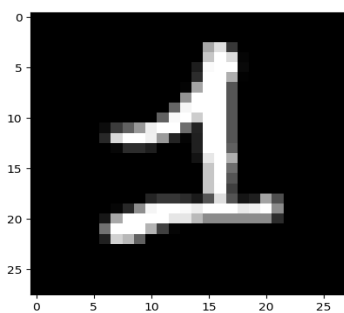
Η παρούσα εργασία εκπονήθηκε στο πλαίσιο του προπτυχιακού μαθήματος «Νευρωνικά Δίκτυα και Βαθιά Μάθηση», το οποίο πραγματοποιήθηκε κατά το χειμερινό εξάμηνο του διδακτικού έτους 2020-2021 στο Τμήμα Πληροφορικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, και αποτελεί μια εκτενή ενασχόληση με τις μεθόδους οι οποίες συζητήθηκαν στα πλαίσια του μαθήματος καθώς και εφαρμογή αυτών στην δημόσια βάση χειρόγραφων ψηφίων MNIST. Για το προγραμματιστικό μέρος της εργασίας χρησιμοποιήθηκε η γλώσσα προγραμματισμού Python 3.8.

2. Βάση Δεδομένων

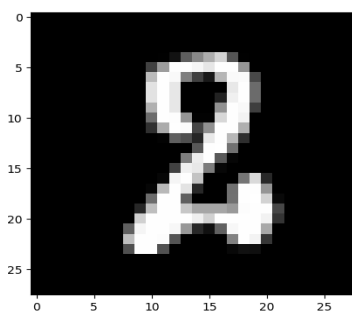
Για τους σκοπούς της εργασίας επιλέχθηκε η δημόσια βάση χειρόγραφων ψηφίων MNIST η οποία μπορεί να βρεθεί στην ηλεκτρονική διεύθυνση <http://yann.lecun.com/exdb/mnist/>. Ένα χαρακτηριστικό δείγμα των δεδομένων της εν λόγω βάσης παρουσιάζεται στις παρακάτω εικόνες.



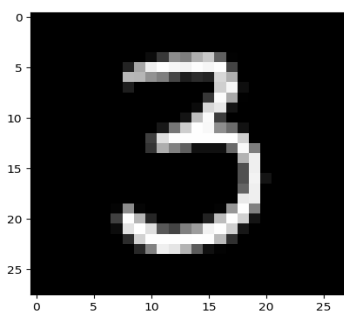
Το ψηφίο 0



Το ψηφίο 1



Το ψηφίο 2



Το ψηφίο 3

Σχήμα 1: Χαρακτηριστικές εικόνες από την βάση MNIST

Οι εικόνες αποτελούνται από 784 pixels (ή αλλιώς 28x28) και είναι ασπρόμαυρες. Η βάση περιέχει 60.000 εικόνες ψηφίων με σκοπό την χρήση τους στην εκπαίδευση της εκάστοτε μεθόδου και 10.000 για τον έλεγχο της ποιότητας της εκπαίδευσης. Επίσης, παρέχονται και οι αντίστοιχες ετικέτες, οι οποίες υποδεικνύουν το ψηφίο το οποίο απεικονίζεται στην εκάστοτε εικόνα. Σκοπός της εργασίας είναι μια εκτενής ενασχόληση με την βάση, η ανάπτυξη, εκπαίδευση και αξιολόγηση των βασικών μεθόδων που διδάχθηκαν στα πλαίσια του μαθήματος και όχι η επίτευξη ανταγωνιστικών επιδόσεων. Ως εκ τούτου, καθόλη την εργασία θα χρησιμοποιηθούν μόλις τα πρώτα τέσσερα ψηφία της βάσης (0-3). Ο σκοπός αυτής της σύμβασης είναι η διευκόλυνση παραγωγής αποτελεσμάτων σε εύλογο χρονικό διάστημα χωρίς την απαίτηση μεγάλων υπολογιστικών δυνατοτήτων, έτσι ώστε να διευκολυνθεί η εκπαιδευτική διαδικασία. Έτσι το πλήθος των δεδομένων μειώνεται σημαντικά και απομένουν 24754 εικόνες εκπαίδευσης και 4157 εικόνες για τον έλεγχο της εκπαίδευσης. Οι κλάσεις υπό εξέταση είναι πλέον μόλις τέσσερις.

3. Εξαγωγή Χαρακτηριστικών

Κατά την εξέταση των διαφορετικών μεθόδων που πρόκειται να αναλυθούν στην παρούσα εργασία είναι αναγκαστική η διαχείριση των χαρακτηριστικών τα οποία συνθέτουν κάθε δείγμα της βάσης όπως αυτή έχει ήδη εισαχθεί στην προηγούμενη ενότητα. Είναι εύκολο να αντιληφθεί κανείς πως, κάθε εικόνα εκπαίδευσης από τις 24754 που έχουν απομείνει μετά την υιοθέτηση της σύμβασης, και κάθε εικόνα ελέγχου, χαρακτηρίζεται πλήρως από τα 784 pixels που την απαρτίζουν.

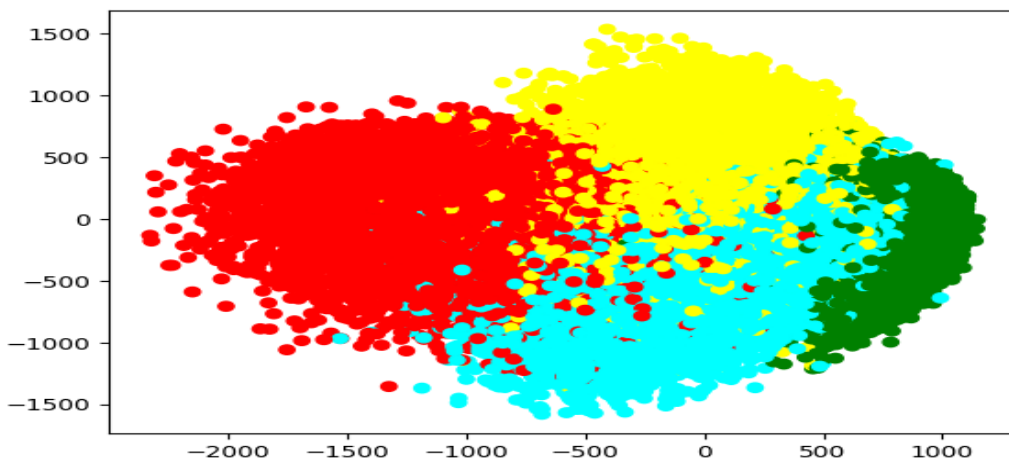
Είναι εύλογο να θελήσει κάποιος να μειώσει τον αριθμό των χαρακτηριστικών αυτών, δίχως να χάνεται πληροφορία για τα δείγματα της βάσης, η οποία θα μπορούσε να φανεί πολύτιμη για την επίλυση προβλημάτων από τις μεθόδους που θα αναπτυχθούν. Έχοντας αυτή την συλλογιστική κατα νού, προτάσσεται η χρήση της μεθόδου ανάλυσης κυρίων συνιστωσών, γνωστής και ως PCA, ενώ παράλληλα διατηρείται τουλάχιστον το 90% της πληροφορίας που λαμβάνεται από το σύνολο των χαρακτηριστικών.

3.1. Μέθοδος Ανάλυσης Κυρίων Συνιστωσών

Κάνοντας χρήση της μεθόδου PCA τα χαρακτηριστικά ενός δείγματος μπορούν να μειωθούν από 784 σε οποιοδήποτε αριθμό $V < 784$. Για τον σκοπό αυτό γίνονται οι εξής υπολογισμοί:

1. Έστω είσοδος μία εικόνα x με 784 pixels, σχήματος 28×28 .
2. Τα 784 pixels της εικόνας αυτής συμβολίζονται ως το διάνυσμα \vec{x} .
3. Γίνεται κανονικοποίηση προς τον μέσο για το \vec{x} , έστω \vec{x}_{norm} .
4. Υπολογίζεται ο πίνακας αυτοσυσχέτισης για το \vec{x}_{norm} , έστω R_x .
5. Εύρεση των ιδιοτιμών και των ιδιοδιανυσμάτων του πίνακα R_x .
6. Ταξινόμηση των ιδιοτιμών και των ιδιοδιανυσμάτων του προηγούμενου βήματος.
7. Διατηρούνται τα ιδιοδιανύσματα των V μεγαλύτερων ιδιοτιμών στον πίνακα $U_{reduced}$.
8. Δημιουργία του νέου διανύσματος χαρακτηριστικών, $x_{trans} = x_{norm} * U_{reduced}$.

Έτσι κάθε δείγμα έχει πλέον 2 χαρακτηριστικά και μπορεί να οπτικοποιηθεί βάση αυτού του νέου διανύσματος και της ετικέτας που προσδιορίζει το ψηφίο το οποίο εμφανίζεται στην εικόνα. Σε κάθε κλάση ανατίθεται ένα χρώμα. Στην προκειμένη περίπτωση, ισχύουν οι εξής αναθέσεις: κόκκινο για το ψηφίο 0, πράσινο για το ψηφίο 1, κυανό για το ψηφίο 2 και κίτρινο για το ψηφίο 3. Για να διατηρηθεί τουλάχιστον το 90% της πληροφορίας που προσφέρεται από το σύνολο των χαρακτηριστικών κάθε δείγματος, απαιτείται η χρήση V μεγαλύτερου ή ίσου με 75 για τα δείγματα εκπαίδευσης και 71 για τα δείγματα δοκιμής. Επομένως γίνεται χρήση του $V = 75$.



Σχήμα 2: Οπτικοποίηση της βάσης με χρήση της μεθόδου της ανάλυσης κυρίων συνιστωσών για $V = 2$

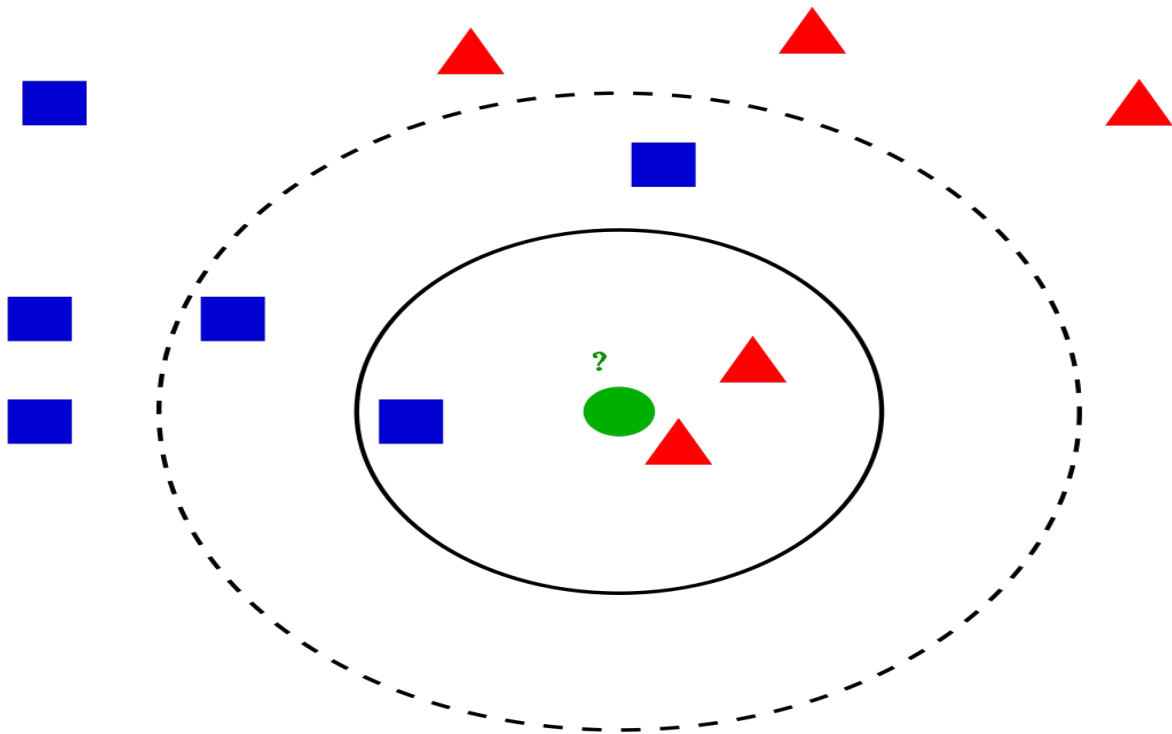
4. Μετρικές Απόστασης

Οι μετρικές που χρησιμοποιήθηκαν στην παραπάνω πειραματική διαδικασία για κάθε μέθοδο είναι οι εξής:

- Ευκλείδειες αποστάσεις με χρήση της `np.linalg.norm(point1 - point2)`.
- Αποστάσεις Minkowski με χρήση της `scipy.spatial.distance.minkowski` για όρισμα $p = 3$.
- Αποστάσεις Manhattan με χρήση της `scipy.spatial.distance.cityblock`.
- Αποστάσεις Chebyshev με χρήση της `scipy.spatial.distance.chebyshev`.

5. Κατηγοριοποιητής Κ Πλησιέστερων Γειτόνων

Ο κατηγοριοποιητής πλησιέστερων γειτόνων είναι μία ιδιαίτερα διαδεδομένη επιβλεπόμενη μέθοδος. Ο αλγόριθμος δέχεται ως είσοδο τα σημεία τα οποία πρέπει να χρησιμοποιήσει για να λάβει μια μελλοντική απόφαση κατηγοριοποίησης, τις ετικέτες αυτών των σημείων, την παράμετρο k , καθώς και το/τα σημείο/σημεία που καλείται να κατηγοριοποιήσει. Η παράμετρος k που αναφέρθηκε προηγουμένος καθορίζει το πλήθος των σημείων που εν τέλει θα επηρεάσουν την απόφαση που θα ληφθεί για ένα καινούργιο σημείο προς κατηγοριοποίηση. Η γενική ιδέα του κατηγοριοποιητή αυτού είναι η προσπάθεια να εκμεταλλευτεί την τοπικότητα η οποία χαρακτηρίζει τα δείγματα που ανήκουν στην ίδια κλάση για το εκάστοτε πρόβλημα. Σχηματικά μπορεί να περιγραφεί ως εξής:



Σχήμα 3: Source: <https://commons.wikimedia.org/wiki/File:KnnClassification.svg>

Στην παραπάνω εικόνα αποτυπώνεται ξεκάθαρα το πλαίσιο μέσα στο οποίο δίνεται η απάντηση κατηγοριοποίησης από τον εν λόγω αλγόριθμο. Το πράσινο κυκλικό δείγμα απεικονίζει το νέο δείγμα τα οποίο καλείται η μέθοδος να εισάγει σε μία από τις δύο άλλες κλάσεις (κόκκινα τρίγωνα ή μπλέ τετράγωνα). Η απόφαση αυτή λαμβάνεται με την εξής συλλογιστική πορεία:

- Για $k = 1$, η μέθοδος εκφυλίζεται στον κατηγοριοποιητή πλησιέστερου γείτονα, και έτσι εξετάζεται μόνο το κοντινότερο δείγμα το οποίο στην προκειμένη περίπτωση είναι το κόκκινο τρίγωνο. Έτσι το νέο πράσινο κυκλικό δείγμα θα κατηγοριοποιηθεί στην κόκκινη τριγωνική κλάση.

- Για $k = 3$, γίνεται χρήση όλων των δειγμάτων μέσα στον συνεχή κύκλο και αποφασίζεται ποι-ά κλάση υπερτερεί αριθμητικά. Τελικά, το πράσινο κυκλικό δείγμα θα κατηγοριοποιηθεί στην κόκκινη τριγωνική κλάση.
- Για $k = 5$, γίνεται χρήση όλων των δειγμάτων μέσα στον διακεκομμενη κύκλο και αποφασίζεται και πάλι ποιά κλάση υπερτερεί αριθμητικά. Τελικά, το πράσινο κυκλικό δείγμα θα κατηγοριο-ποιηθεί στην μπλέ τετραγωνική κλάση.

Τα βήματα του αλγορίθμου τα οποία θα χρησιμοποιηθούν παρακάτω είναι απλά και περιγράφονται στις επόμενες κουκίδες στις οποίες ως απόσταση πειράγεται ένα σύνολο διαφορετικών μετρικών που μπορούν να χρησιμοποιηθούν (π.χ Ευκλείδεια, Μανχάταν, Μινκόβσκι) και ως σημεία αναφοράς, τα σημεία τα οποία χρησιμοποιούνται για ληφθεί η απόφαση κατηγοριοποίησης.

- Λάβε ως είσοδο τα σημεία τα οποία πρέπει να χρησιμοποιήσεις για να λάβεις μια μελλοντική απόφαση κατηγοριοποίησης, τις ετικέτες αυτών των σημείων, την παράμετρο k , καθώς και το/τα σημείο/σημεία που καλείσαι να κατηγοριοποιήσεις.
- Για κάθε δείγμα που πρέπει να κατηγοριοποιηθεί και για κάθε δείγμα αναφοράς.
- Υπολόγισε την απόσταση ανάμεσα στο σημείο προς κατηγοριοποίηση και τα σημεία αναφοράς.
- Επίλεξε τα K κοντινότερα δείγματα βάση της απόστασης που υπολόγισες.
- Λάβε την απόφαση κατηγοριοποίησης βάση της συχνότερης ετικέτας στα K επιλεγμένα δείγματα.

Ο κατηγοριοποιητής K Πλησιέστερων Γειτόνων εμφανίζει τα εξής πειραματικά αποτελέσματα για διάφορες μετρικές απόστασης, 75 χαρακτηριστικά εισόδου και $K = 1$. Ο χρόνος καταγράφεται σε δευτερόλεπτα:

Distance Metric Used	Training Time	Training Accuracy	Testing Accuracy
Euclidean	0.1	100%	99.45%
Minkowski	0.1	100%	99.42%
Manhattan	0.1	100%	99.21%
Chebyshev	0.1	100%	99.21%

Ο χρόνος εκπαίδευσης είναι μηδαμινός αφού το μόνο που χρειάζεται η διαδικασία εκπαίδευσης της μεθόδου είναι η λήψη των δεδομένων αναφοράς, η οποία γίνεται σε $O(1)$ χρονική πολυπλοκότητα. Για την κατηγοριοποίηση των δειγμάτων δοκιμής απαιτείται ο υπολογισμός των αποστάσεων για κάθε ζεύγος σημείων ανάμεσα στα διανύσματα εκπαίδευσης και σε αυτά που καλείται η μέθοδος να κατηγοριοποιήσει και άρα η μέθοδος εμφανίζει χρονική πολυπλοκότητα $O(nm)$. Η πολυπλοκότητα αυτής της τάξης γίνεται απαγορευτική για μεγάλα n και m . Επίσης ο υπολογισμός της απόστασης Minkowski είναι κατά πολύ πολυπλοκότερος υπολογιστικά των άλλων, ενώ για την απόσταση Manhattan είναι ελάχιστα ευκολότερος. Για $K = 1$ η μέθοδος απολαμβάνει 100% επιτυχίας στα δείγματα εκπαίδευσης για οποιαδήποτε μετρική αφού ουσιαστικά το μόνο δείγμα που χρησιμοποιείται για την κατηγοριοποίηση τους είναι ο ίδιος τους ο ευατός και η απόσταση τους από το εν λόγω δείγμα είναι μηδενική.

Η χρονική επίδοση του αλγορίθμου είναι πολύ κακή κάτι το αναμενόμενο άλλωστε. Για κάθε δείγμα προς κατηγοριοποίηση (4157) θα πρέπει να υπολογισθεί η απόσταση από κάθε δείγμα αναφοράς (24754). Για κάθε τέτοιο υπολογισμό απαιτούνται βασικές πράξεις ίσες με τα χαρακτηριστικά κάθε δείγματος. Έτσι στην συγκεκριμένη περίπτωση ο αλγόριθμος χαρακτηρίζεται από πολυπλοκότητα $O(4157 * 24754 * \text{number of characteristics in each sample})$, η οποία για 75 χαρακτηριστικά οδηγεί σε 7.717.678.350 βασικές πράξεις. Ο πραγματικός χρόνος που δαπανάται από το πρόγραμμα δεν αποτελεί το καλύτερο μέτρο σύγκρισης αφού εξαρτάται κατά πολύ από το εκάστοτε υπολογιστικό σύστημα, της δυνατότητες του σε πόρους καθώς και τον φόρτο εργασίας του συστήματος από άλλους παράγοντες πέραν του αλγορίθμου.

Ο κατηγοριοποιητής K Πλησιέστερων Γειτόνων εμφανίζει τα εξής πειραματικά αποτελέσματα για διάφορες μετρικές απόστασης, 75 χαρακτηριστικά εισόδου και $K = 3$. Ο χρόνος καταγράφεται σε δευτερόλεπτα:

Distance Metric Used	Training Time	Training Accuracy	Testing Accuracy
Euclidean	0.1	99.71%	99.42%
Minkowski	0.1	99.69%	99.49%
Manhattan	0.1	99.60%	99.16%
Chebyshev	0.1	99.53%	99.25%

Η επίδοση της μεθόδου μειώνεται αφού όταν συμβάλλουν τρεις γείτονες στην απόφαση κατηγοριοποίησης προκύπτουν λάθη σε οριακά δείγματα των κλάσεων. Η χρονική επίδοση του κατηγοριοποιητή δεν επηρεάζεται από την αύξηση του αριθμού των γειτόνων αφού δεν επηρεάζεται η βασική πράξη της μεθόδου.

6. Κατηγοριοποιητής Πλησιέστερου Κέντρου

Ο κατηγοριοποιητής πλησιέστερου κέντρου είναι μια μέθοδος που μοιάζει πολύ σε λογική με αυτή των K-μέσων. Ο αλγόριθμος δέχεται ως είσοδο τα σημεία τα οποία πρέπει να χρησιμοποιήσει για να λάβει μια μελλοντική απόφαση κατηγοριοποίησης, τις ετικέτες αυτών των σημείων, καθώς και το/τα σημείο/σημεία που καλείται να κατηγοριοποιήσει. Η γενική ιδέα του κατηγοριοποιητή αυτού είναι η δημιουργία ενός κύκλου γύρω από κάθε κλάση ο οποίος έχει ως κέντρο του έχει το διάνυσμα με τις μέσες τιμές για κάθε χαρακτηριστικό του συνόλου των σημείων που συμμετέχουν στην κλάση αυτή.

Τα βήματα του αλγορίθμου τα οποία θα χρησιμοποιηθούν παρακάτω είναι απλά και περιγράφονται στις επόμενες κουκίδες στις οποίες ως απόσταση πειγράφεται ένα σύνολο διαφορετικών μετρικών που μπορούν να χρησιμοποιηθούν (π.χ Ευκλείδεια, Μανχάταν, Μινκόβσκι) και ως σημεία αναφοράς, τα σημεία τα οποία χρησιμοποιούνται για ληφθεί η απόφαση κατηγοριοποίησης.

- Λάβε ως είσοδο τα σημεία τα οποία πρέπει να χρησιμοποιήσεις για να λάβεις μια μελλοντική απόφαση κατηγοριοποίησης, τις ετικέτες αυτών των σημείων, την παράμετρο k , καθώς και το/τα σημείο/σημεία που καλείσαι να κατηγοριοποιήσεις.
- Για κάθε κλάση, υπολόγισε το άθροισμα κάθε χαρακτηριστικού για κάθε δείγμα που ανήκει στην κλάση.
- Υπολόγισε το κέντρο κάθε κλάσης διαιρώντας το διάνυσμα του αθροίσματος που βρήκες στο προηγούμενο βήμα με τον αριθμό των δειγμάτων που ανήκουν στην αντίστοιχη κλάση.
- Υπολόγισε την απόσταση του δείγματος προς κατηγοριοποίηση με το κέντρο κάθε κλάσης.
- Λάβε την απόφαση κατηγοριοποίησης του δείγματος βάση της μικρότερης απόστασης από το κέντρο μίας κλάσης. Ανέθεσε την κλάση αυτή στο δείγμα αυτό.

Ο κατηγοριοποιητής Πλησιέστερου Κέντρου εμφανίζει τα εξής πειραματικά αποτελέσματα για διάφορες μετρικές απόστασης και 75 χαρακτηριστικά εισόδου. Ο χρόνος καταγράφεται σε δευτερόλεπτα:

Distance Metric Used	Training Time	Training Accuracy	Testing Accuracy
Euclidean	1.5	91.26%	91.75%
Minkowski	1.5	90.52%	91.20%
Manhattan	1.5	91.65%	91.85%
Chebyshev	1.5	84.98%	85.60 %

Η επίδοση της μεθόδου σε σύγκριση με τον κατηγοριοποιητή K πλησιέστερων γειτόνων είναι σαφώς χειρότερη. Αυτό είναι προφανώς αναμενόμενο αφού πλέον για την κατηγοριοποίηση γίνεται χρήση του κέντρου και όχι των γειτόνων κάθε δείγματος. Επομένως τα δείγματα τα οποία κατηγοριοποιούνται με λάθος τρόπο επειδή ανήκουν σε οριακές περιπτώσεις ανάμεσα σε δύο κλάσεις αυξάνουν κατά πολύ.

Απο άποψη χρονικής επίδοσης οι δύο μέθοδοι είναι η μέρα με την νύχτα. Δεν υπάρχει καμία σύγκριση για το ποιά είναι ανώτερη σε αυτό το κομμάτι. Αυτή η παρατήρηση είναι προφανώς αναμενόμενη, και αυτό διότι, για κάθε νέο δείγμα η κατηγοριοποίηση του δεν απαιτεί τον υπολογισμό της απόστασης του από κάθε προυπάρχων δείγμα αλλά απο το κέντρο των δύο κλάσεων. Έτσι η βασική πράξη πλέον δεν είναι ο υπολογισμός των αποστάσεων ανάμεσα στα σημεία αλλά ο καθορισμός του κέντρου για κάθε κλάση. Για κάθε δείγμα εκπαίδευσης (24754) θα πρέπει να υπολογισθεί η συμβολή του στο κέντρο της κλάσης στην οποία ανήκει. Για κάθε τέτοιο υπολογισμό απαιτούνται βασικές πράξεις ίσες με τα χαρακτηριστικά κάθε δείγματος. Έτσι στην συγκεκριμένη περίπτωση ο αλγόριθμος χαρακτηρίζεται από πολυπλοκότητα $O(24754 * \text{number of characteristics in each sample})$ η οποία για 75 χαρακτηριστικά οδηγεί σε 1.856.550 βασικές πράξεις. Ο πραγματικός χρόνος που δαπανάται απο το πρόγραμμα δεν αποτελεί το καλύτερο μέτρο σύγκρισης όπως έχει ήδη εξηγηθεί.

Οι παρακάτω γενικές παρατηρήσεις αξίζει να αναφερθούν σχετικά με τις δοκιμές που έλαβαν χώρα με τις προαναφερθείσες μετρικές για κάθε μέθοδο:

- Η χρήση της απόστασης Manhattan βελτιώνει την επίδοση του κατηγοριοποιητή πλησιέστερου κέντρου ενώ βελτιώνει και τον χρόνο.
- Η χρονική απόδοση για την απόσταση Minkowski για τον κατηγοριοποιητή K πλησιέστερων γειτόνων, είναι απαγορευτική, χωρίς να βελτιώνει την επίδοση την μεθόδου. Ακόμη και αν η θεωρητική πολυπλοκότητα του αλγορίθμου παραμεινεί η ίδια, η πρακτική πολυπλοκότητα η οποία εξαρτάται από το πόσο γρήγορα υπολογίζεται η απόσταση ανάμεσα στα σημεία καθιστά την μέθοδο των K πλησιέστερων γειτόνων υπερβολικά αργή για την συγκεκριμένη μετρική απόστασης.
- Η μέθοδος του πλησιέστερου κέντρου φαίνεται να ευνοείται από απλότερες μετρικές, όπως η Manhattan και η Ευκλείδεια σε σχέση με τις πιο περίπλοκες Minkowski και Chebyshev.

7. Νευρωνικό Δίκτυο Συνάρτησης Ακτινικής Βάσης

Ένα βασικό Νευρωνικό Δίκτυο Συνάρτησης Ακτινικής Βάσης αποτελεί μία συγκεκριμένη αρχιτεκτονική τριών επιπέδων νευρώνων. Στο πρώτο επίπεδο βρίσκονται οι νευρώνες εισόδου οι οποίοι συνδέονται πλήρως με το δεύτερο στρώμα το οποίο αποτελείται από νευρώνες με συνάρτηση ενεργοποίησης κάποια συνάρτηση ακτινικής βάσης, εξού και το όνομα του δικτύου. Τέλος οι εξόδοι του μεσαίου στρώματος προωθούνται στο τελευταίο στρώμα, αυτό της εξόδου, το οποίο για προβλήματα κατηγοριοποίησης αποτελείται από K νευρώνες, όπου K ο αριθμός των πιθανών κλάσεων για ένα δείγμα.

Ο βασικός στόχος του δικτύου είναι η δημιουργία συστάδων γύρω από κάποια κέντρα οι οποίες θα περιέχουν το σύνολο των δειγμάτων που θεωρεί το δίκτυο πως ανήκουν στην ίδια κλάση. Έτσι όταν καλείται να κατηγοριοποιήσει κάποιο νέο δείγμα υπολογίζει την απόσταση του από το κέντρο κάθε συστάδας και παράγει το ανάλογο αποτέλεσμα.

Για το μεσαίο στρώμα, το οποίο είναι και υπεύθυνο για την δημιουργία συστάδων και τον υπολογισμό των αποστάσεων κάθε δείγματος απο τα κέντρα τους, απαιτείται το κέντρο καθώς και η διασπορά για κάθε νευρώνα του.

Για να βρεθούν τα κέντρα μπορεί να γίνει τυχαία επιλογή από τα δείγματα εκπαίδευσης ή χρήση κάποιας μεθόδου ομαδοποίησης, όπως αυτή των K -Μέσων. Για την διασπορά κάθε συστάδας γίνεται χρήση της μέγιστης απόστασης ανάμεσα σε δύο διαφορετικά κέντρα από αυτά που έχουν επιλεγεί στο προηγούμενο βήμα.

Η εκπαίδευση γίνεται σε δύο στάδια:

- Στο πρώτο στάδιο ρυθμίζονται τα κέντρα καθώς και η διασπορές κάθε νευρώνα του μεσαίου επιπέδου στο δίκτυο.
- Στο δεύτερο στάδιο πρέπει να βρεθούν τα βάρη των συνδέσεων ανάμεσα στο μεσαίο και το τελευταίο επίπεδο του δικτύου, τα οποία ορίζουν κάποια γραμμική σχέση ανάμεσα στις εξόδους του επιπέδου της συνάρτησης ακτινικής βάσης και την εξόδο.

Στην παρούσα εργασία γίνεται χρήση της μεθόδου Γραμμικής Παλινδρόμησης Ελαχίστων Τετραγώνων με χρήση του ψευδοαντίστροφου πίνακα για την εύρεση των βαρών του τελευταίου επιπέδου του δικτύου.

Από τα παρακάτω αποτελέσματα φαίνεται πώς:

- Η χρήση της τυχαίας αρχικοποίησης κέντρων φαίνεται να έχει καλύτερη χρονική απόδοση από αυτή των K-Μέσων όπως ήταν αναμενόμενο.
- Η χρήση των K-Μέσων και για μικρά K βλεπώνει αισθητά την επίδοση των κατηγοριοποιήσεων της μεθόδου. Ενώ το K μεγαλώνει, υπάρχει σύγκλιση την επίδοση των δυο αρχικοποιήσεων, αφού τα βάρη προσεγγίζουν όλο και περισσότερο στην ελαχιστοποίηση της συνάρτησης κόστους.
- Ο χρόνος εκπαίδευσης αυξάνεται κατά πολύ σε αναλογία με τον αριθμό των κέντρων που χρησιμοποιούνται.
- Φαίνεται πώς η μέθοδος φτάνει στις καλύτερες περιπτώσεις και για τα πειράματα που πραγματοποιήθηκαν το 94% επιτυχίας.
- Για μεγαλύτερα από 300 κέντρα η επίδοση συνεχίζει να αυξάνεται έως ότου το μοντέλο υπερεκπαιδεύεται, όμως οι χρονικές απαιτήσεις της μεθόδου αυξάνουν κατά πολύ λόγω της πολυπλοκότητας της. Η εν λόγω αδυναμία πιθανότατα θα μπορούσε να καταπολεμηθεί με κάποια παραλλαγή της αναπτυχθείσας υλοποίησης και συνδυασμό των Νευρωνικών Δικτύων Συνάρτησης Ακτινικής Βάσης με κάποιον κανόνα οπισθοδρομικής ενημέρωσης βαρών ο οποίος χρησιμοποιείται στις υλοποιήσεις των νευρωνικών δικτύων. Έτσι θα μπορούσε, αφού γίνει χρήση κάποιου αριθμού κέντρου (π.χ. 100) και καταφέρει το μοντέλο να φτάσει μία ορισμένη επίδοση, ο κανόνας οπισθοδρομικής θα επιφέρει τις απαραίτητες αλλαγές με ελάχιστες χρονικές απαιτήσεις στα βάρη του τελευταίου στρώματος αυξάνοντας ακόμη περισσότερο τις επιδόσεις του μοντέλου και συγκλίνοντας γρήγορα στο ελάχιστο για τα δεδομένα του προβλήματος υπό εξέταση.

7.1. Αποτελέσματα Τυχαίων Κέντρων

Centers Used	Training Time	Training Accuracy	Testing Accuracy
4	1	59.20%	60.14%
10	2	83.93%	85.11%
75	12	91.93%	92.57%
150	36	90.31%	90.93%
300	120	91.35%	91.51%

7.2. Αποτελέσματα K-Μέσων

Centers Used	Training Time	Training Accuracy	Testing Accuracy
4	5	87.81%	87.92%
10	11	91.50%	91.41%
75	110	91.14%	92.06%
150	184	89.32%	89.94%
300	226	91.69%	92.08%

8. Σχολιασμός Αποτελεσμάτων και Σύγκριση Μεθόδων

Η πρώτη μέθοδος που εξετάστηκε είναι ο Κατηγοριοποιητής K Πλησιέστερων Γειτόνων για $K = 1$ και $K = 3$. Το μεγάλο πλεονέκτημα της μεθόδου είναι ο χρόνος εκπαίδευσης ο οποίος είναι μηδενικός αφού η εκπαίδευση που απαιτείται από την μέθοδο είναι ένα απλό πέρασμα των δειγμάτων εκπαίδευσης στην συνάρτηση που την υλοποιεί. Από την άλλη πλευρά, το μεγάλο μειονέκτημα της εν λόγω μεθόδου ανεξαρτήτως του K , είναι ο χρόνος που απαιτείται για να παράξει τις κατηγοριοποιήσεις νέων δειγμάτων. Εφόσον κάθε νέο δείγμα προς κατηγοριοποίηση απαιτεί τον υπολογισμό της απόστασης του από κάθε δείγμα εκπαίδευσης, γίνεται εύκολα κατανοητό ότι η πολυπλοκότητα της μεθόδου αυξάνει ραγδαία για μεγάλο αριθμό δειγμάτων αναφοράς και δοκιμής.

Για $K = 1$, η ακρίβεια στα δείγματα εκπαίδευσης είναι 100%. Αυτή η τέλεια ακρίβεια επιτυγχάνεται αφού κάθε δείγμα εκπαίδευσης έχει ως μοναδικό δείγμα αναφοράς κατά την κατηγοριοποίηση του τον εαυτό του και έτσι λαμβάνει εγγυημένα την σωστή ετικέτα κλάσης. Για $K = 3$ αυτή η τέλεια ακρίβεια χάνεται και παρατηρείται μία λίγο χειρότερη ακρίβεια (99.7%). Αυτή η μείωση οφείλεται σε οριακές περιπτώσεις λάθος κατηγοριοποίησης στις οποίες, αν και ο ένας εκ των τριών γειτόνων που χρησιμοποιούνται για την κατηγοριοποίηση του δείγματος εκπαίδευσης είναι το ίδιο το δείγμα με σωστή ετικέτα, οι άλλοι δύο γείτονες ανήκουν σε γειτονική κλάση, και έτσι η τελική απόφαση είναι λανθασμένη.

Για τα δοκιμαστικά δείγματα προς κατηγοριοποίηση, παρατηρείται, για $K = 1$ ακρίβεια της τάξης του 99.45%, και για $K = 3$ ακρίβεια της τάξης του 99.42% για Ευκλείδειες αποστάσεις, οι οποίες παρουσιάζουν την καλύτερη σχέση χρονικής επίδοσης και ακρίβειας. Φαίνεται πως η ακρίβεια του Κατηγοριοποιητή K Πλησιέστερων Γειτόνων τείνει να μειωθεί για νέα δείγματα προς κατηγοριοποίηση, καθώς αυξάνει το K , χωρίς να επηρεάζεται η χρονική επίδοση της μεθόδου αφού η επιλογή των K κοντινότερων γειτόνων ενός δείγματος δεν είναι η βασική πράξη της μεθόδου. Αυτή η συμπεριφορά ενδεχομένως να οφείλεται σε μεγάλα ποσοστά υπερκαληψης των δύο κλάσεων για οριακές περιπτώσεις κατηγοριοποίησης ανάμεσα σε δύο κλάσεις. Αντιθέτως, με χρήση της απόστασης Minkowski παρατηρείται η ακριβώς αντίστροφη συμπεριφορά, δηλαδή η ακρίβεια για νέα δείγματα προς κατηγοριοποίηση αυξάνει μαζί με το K . Μία απιθανή εξήγηση για την παρατήρηση αυτή θα μπορούσε να είναι ο πολυπλοκότερος υπολογισμός που επιτελείται κατά τον υπολογισμό των αποστάσεων Minkowski, η οποία ενδέχεται να ευνοεί το πρόβλημα κατηγοριοποίησης προς εξέταση και τις οριακές περιπτώσεις που εμφανίζει.

Στην συνέχεια δοκιμάστηκε η μέθοδος του Κατηγοριοποιητή Πλησιέστερου Κέντρου. Το μεγάλο πλεονέκτημα της μεθόδου είναι ο σχεδόν μηδενικός χρόνος εκπαίδευσης σε συνδυασμό με τον πολύ μικρό χρόνο που απαιτεί για την κατηγοριοποίηση νέων δειγμάτων, δηλαδή μερικά δευτερόλεπτα. Αρχικά η μέθοδος χρειάζεται τα δείγματα εκπαίδευσης και πολύ λίγο χρόνο, προκειμένου να βρεί τα κέντρα των κλάσεων που καλείται να διαχειριστεί, και έτσι προκύπτει ο χρόνος εκπαίδευσης. Έπειτα, και αφού έχει υπολογίσει τα κέντρα στο προηγούμενο βήμα, κατηγοριοποιεί τα νέα δείγματα βάση της απόστασης τους από τα παραπάνω κέντρα και άρα ο χρόνος που δαπανάται στην κατηγοριοποίηση είναι ελάχιστος σε σχέση με την μέθοδο των K Πλησιέστερων Γειτόνων η οποία απαιτούσαι τον υπολογισμό την απόστασης του νέου δείγματος από κάθε δείγμα εκπαίδευσης. Έτσι ο χρόνος που χρειάζεται η μέθοδος για να κάνει νέες προβλέψεις μειώνεται δραματικά όταν την συγκρίνει κανείς με την πρώτη μέθοδο που εξετάστηκε.

Το μεγάλο μειονέκτημα του Κατηγοριοποιητή Πλησιέστερου Κέντρου είναι η ακρίβεια που παρέχει στις προβλέψεις του για τις ετικέτες των δειγμάτων. Η μέθοδος παρουσιάζει κατα πολύ την χειρότερη επίδοση σε σχέση με όλες τις υπόλοιπες μεθόδους που εξετάστηκαν με μόλις 91.65% ακρίβεια για τα δείγματα εκπαίδευσης και 91.85% ακρίβεια για τα δείγματα δοκιμής. Η παρατήρηση αυτή είναι αναμενόμενη αφού ο εν λόγω κατηγοριοποιητής δεν εκμεταλλεύεται σε ικανοποιητικό βαθμό την τοπικότητα που παρουσιάζουν τα δείγματα της εκάστοτε κλάσης. Έτσι, πολλά δείγματα καταλήγουν με λάθος ετικέτα κλάσης μετά την κατηγοριοποίησή τους. Εάν σχηματίσουμε έναν κύκλο γύρω από το κέντρο κάθε κλάσης, με κατάλληλη ακτίνα, έτσι ώστε να καλύπτει τα δείγματα που δεν μπορούν να κατηγοριοποιηθούν σε κάποια άλλη κλάση βάση απόστασης, πολλά δείγματα θα παραμείνουν εκτός κάποιου κύκλου. Αυτά ακριβώς τα δείγματα παρουσιάζουν πιθανότητα λανθασμένης κατηγοριοποίησης, και ένα ποσοστό από αυτά θα κατηγοριοποιηθεί λανθασμένα στο τέλος της μεθόδου.

Τέλος, εξετάσθηκε το Νευρωνικό Δίκτυο Συνάρτησης Ακτινικής Βάσης. Το μεγάλο πλεονέκτημα της μεθόδου είναι πως συνδυάζει τα καλά επίπεδα επίδοσης και την χρονική απόδοση. Καμία από τις κλασσικές μεθόδους που εξετάσθηκαν στα πλαίσια της εργασίας δεν προσφέρουν αυτόν τον συνδυασμό και αυτό είναι που ξεχωρίζει την μέθοδο. Εάν μάλιστα αναπτυχθεί η παραλλαγή που προτίνεται στην ενότητα που περιγράφει την μέθοδο οι επιδόσεις θα βελτιωθούν κατά πολύ και μάλιστα σε εύλογο χρονικό διάστημα.