



ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

Playing with MNIST

Συγγραφέας : Δέκας Δημήτριος
Αριθμός Ειδικού Μητρώου : 3063

6 Δεκεμβρίου 2020

Καθηγητής : Τέφας Αναστάσιος

Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης

© 2020 Δέκας Δημήτριος

Created with L^AT_EX

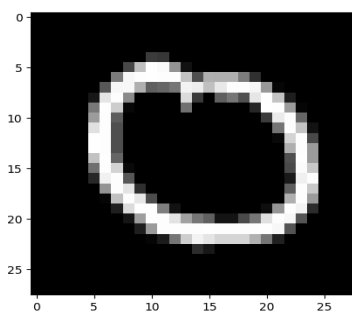
0 1 2 3 4 5 6 7 8 9 10

1. Περίληψη

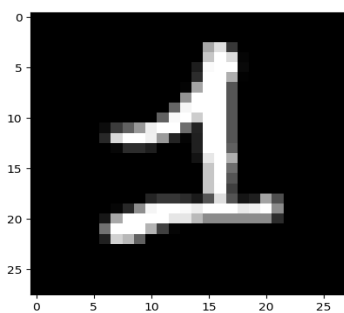
Η παρούσα εργασία εκπονήθηκε στο πλαίσιο του προπτυχιακού μαθήματος «Νευρωνικά Δίκτυα και Βαθιά Μάθηση», το οποίο πραγματοποιήθηκε κατά το χειμερινό εξάμηνο του διδακτικού έτους 2020-2021 στο Τμήμα Πληροφορικής του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης, και αποτελεί μια εκτενή ενασχόληση με τις μεθόδους 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

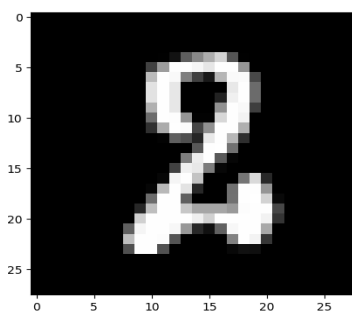
Για τους σκοπούς της εργασίας επιλέχθηκε η δημόσια βάση χειρόγραφων ψηφίων MNIST η οποία μπορεί να βρεθεί στην ηλεκτρονική διεύθυνση <http://yann.lecun.com/exdb/mnist/>. Ένα χαρακτηριστικό δείγμα των δεδομένων της εν λόγω βάσης παρουσιάζεται στις παρακάτω εικόνες.



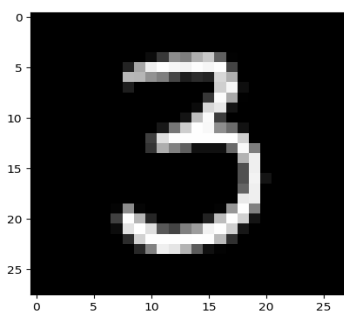
Το ψηφίο 0



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 πώς, 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.

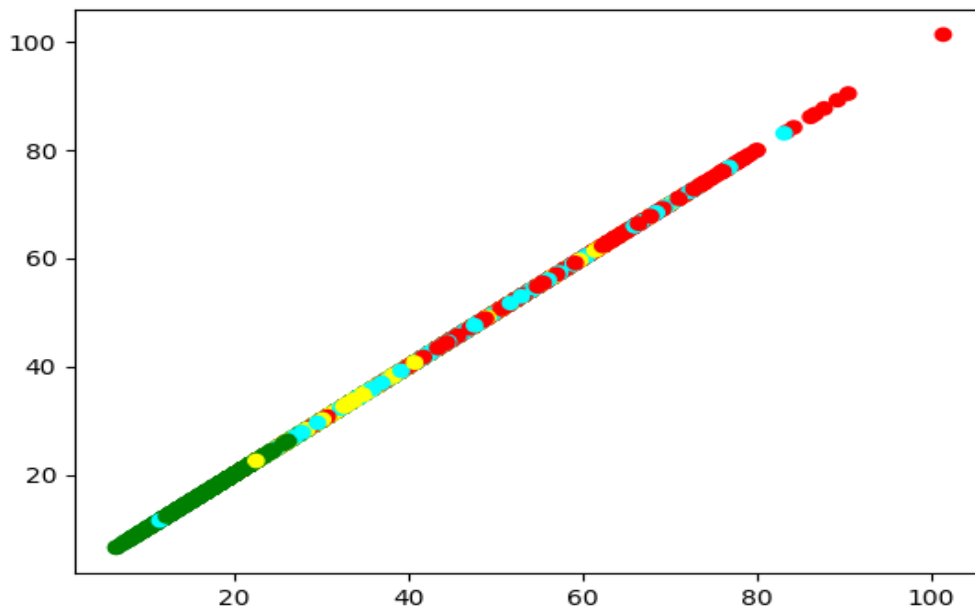
Είναι εύλογο να θελήσει κάποιος να μειώσει τον αριθμό των χαρακτηριστικών αυτών, δίχως να χάνεται πληροφορία για τα δείγματα της βάσης, η οποία θα μπορούσε να φανεί πολύτιμη για την επίλυση προβλημάτων από τις μεθόδους που θα αναπτυχθούν. Έχοντας αυτή την συλλογιστική κατα νού, προτάσσονται δύο πιθανές λύσεις για να αντιμετωπίσουν τον παραπάνω προβληματισμό, και οι οποίες παρουσιάζονται παρακάτω.

3.1. Μέθοδος Μέσης Φωτεινότητας

Η πρώτη μέθοδος προτείνει την μείωση των χαρακτηριστικών από 784 σε μόλις 2 για κάθε δείγμα της βάσης, κάνοντας χρήση της μέσης φωτεινότητας που παρουσιάζεται στις γραμμές και τις στήλες την εκάστοτε εικόνας. Για να επιτευχθεί 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. Υπολογίζεται η μέση φωτεινότητα γραμμών για την εικόνα x , έστω x_{arb} .
4. Υπολογίζεται η μέση φωτεινότητα στηλών για την εικόνα x , έστω x_{acb} .
5. Δημιουργία του νέου διανύσματος χαρακτηριστικών, $x_{trans} = (x_{arb}, x_{acb})$.

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.



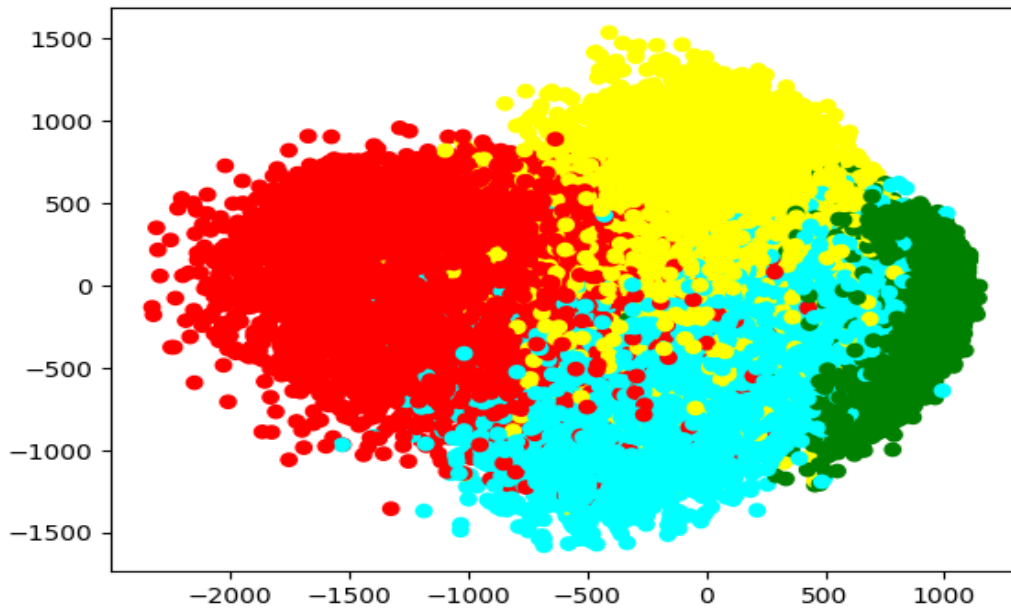
Σχήμα 2: Οπτικοποίηση της βάσης με χρήση της μεθόδου μέσης φωτεινότητας

3.2. Μέθοδος Ανάλυσης Κυρίων Συνιστωσών

Η δεύτερη μέθοδος είναι η πολύ δημοφιλής Ανάλυση Κυρίων Συνιστωσών, γνωστή και ως PCA. Κάνοντας χρήση της συγκεκριμένης μεθόδου 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 οι V μεγαλύτερες ιδιοτιμές και τα αντίστοιχα ιδιοδιανύσματα στον πίνακα $U_{reduced}$.
8. Δημιουργία του νέου διανύσματος χαρακτηριστικών, $x_{trans} = x_{norm} * U_{reduced}$.

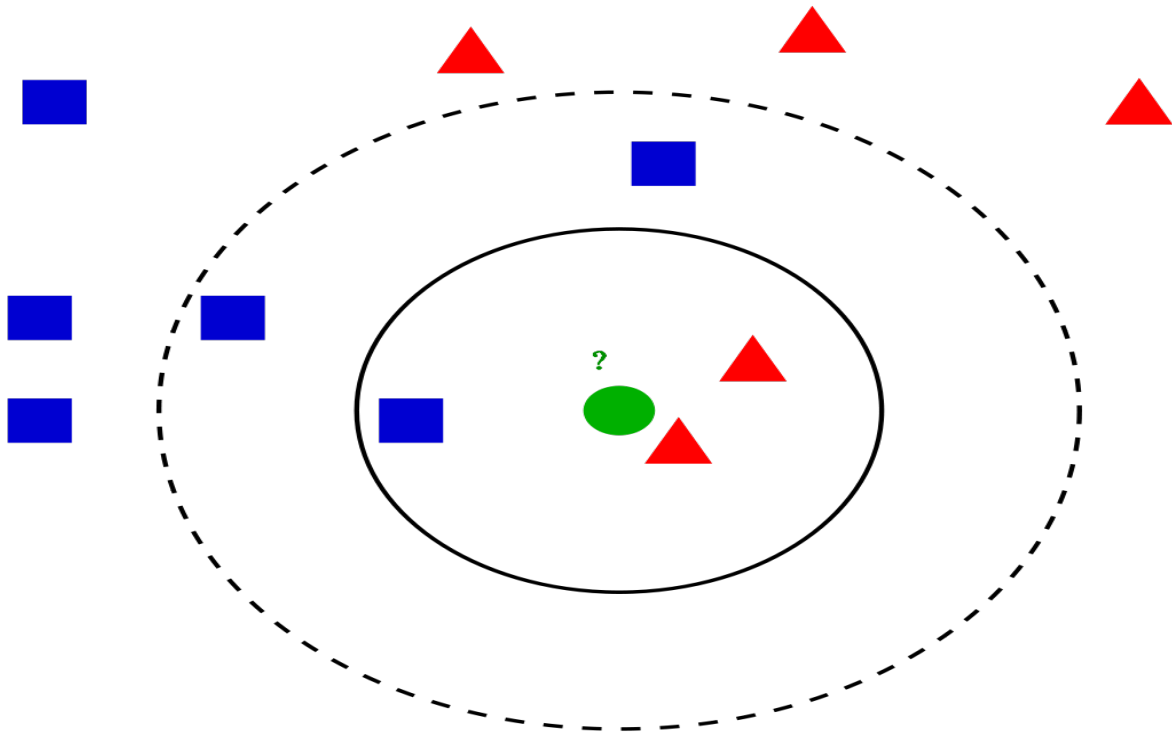
Έτσι κάθε δείγμα έχει πλέον 2 χαρακτηριστικά και μπορεί να οπτικοποιηθεί βάση αυτού του νέου διανύσματος και της ετικέτας που προσδιορίζει το ψηφίο το οποίο εμφανίζεται στην εικόνα. Σε κάθε κλάση ανατίθεται ένα χρώμα. Στην προκειμένη περίπτωση, ισχύουν οι εξής αναθέσεις: κόκκινο για το ψηφίο 0, πράσινο για το ψηφίο 1, κυανό για το ψηφίο 2 και κίτρινο για το ψηφίο 3.



Σχήμα 3: Οπτικοποίηση της βάσης με χρήση της μεθόδου της ανάλυσης κυρίων συνιστωσών για $V = 2$

4. 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 προηγουμένως 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 4: Source: <https://commons.wikimedia.org/wiki/File:KnnClassification.svg>

Εδώ 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 ποιά 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 κατηγοριοποιηθεί στην μπλέ τετραγωνική κλάση.

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 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 διάφορα χαρακτηριστικά εισόδου, $K = 1$ και ευκλείδειες αποστάσεις:

# of Characteristics	Accuracy Percentage	Time Elapsed in Seconds	Basic Operations
Whole Input (784)	99.27832571566033	842.987550497055	80.675.464.352
500 via PCA	45.15275439018523	825.3182067871094	51.451.189.000
300 via PCA	44.96030791436132	680.122180223465	30.870.713.400
100 via PCA	43.78157324993986	655.605224609375	10.290.237.800
2 via PCA	51.81621361558817	556.8936283588409	205.804.756
2 via Average Brightnesses	44.86408467644936	491.9704728126526	205.804.756

Η επίδοση της μεθόδου φαίνεται να είναι εξαιρετική όταν γίνεται χρήση ολόκληρης της εισόδου. Η περίπτωση για την οποία η μέθοδος δίνει πολύ χαμηλή επίδοση, είναι μετά τον μετασχηματισμό των χαρακτηριστικών της εισόδου με την μέθοδο της μέσης φωτεινότητας ή της ανάλυσης κυρίων συνιστωσών. Μια εύλογη εξήγηση για το φαινόμενο αυτό θα μπορούσε να είναι η τοπικότητα η οποία χαρακτηρίζει τα δεδομένα της βάσης. Όταν όμως γίνεται χρήση μεθόδου μείωσης των διαστάσεων των διανυσμάτων που περιέχουν τα χαρακτηριστικά, αυτή η τοπικότητα χάνεται, όπως μπορεί κανείς να διακρίνει στα σχήματα 2 και 3, και οι ομάδες κάθε κλάσεις υπερκαλύπτουν μεγάλο μέρος άλλων κλάσεων δημιουργώντας μεγάλες περιοχές για τις οποίες ο αλγόριθμος σφάλει στην κατηγοριοποίηση. Τέλος φαίνεται πως με χρήση της μεθόδου της ανάλυσης των κυρίων συνιστωσών, και για 2 συνιστώσες, παρατηρείται μια μικρή αύξηση στην επίδοση του αλγορίθμου. Αυτό οφείλεται κυρίως στο γεγονός ότι οι συνιστώσες που προκύπτουν από την εν λόγω μέθοδο αποτελούν ένα άθροισμα των 784 χαρακτηριστικών που εκφράζουν πλήρως το κάθε δείγμα της βάσης και δίνουν ένα μέτρο για την συσχέτιση μεταξύ των δειγμάτων αυτών. Έτσι υπάρχει η περίπτωση, η οποία πιθανότατα λαμβάνει χώρα και στην παρούσα εργασία, τα χαρακτηριστικά της εισόδου να εκφράζονται καλύτερα με μόλις δύο συνιστώσες σε σχέση με την εκφραστικότητα που προσφέρουν περισσότερες συνιστώσες (π.χ. 100,300,500).

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 784 χαρακτηριστικά οδηγεί σε 80.675.464.352 ενώ για 2 σε 205.804.756 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 δυνατότητας 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 διάφορα χαρακτηριστικά εισόδου, $K = 3$ και ευκλείδειες αποστάσεις:

# of Characteristics	Accuracy Percentage	Time Elapsed in Seconds	Basic Operations
Whole Input (784)	99.15804666827039	843.8720397949219	80.675.464.352
500 via PCA	45.297089247053165	783.790962934494	51.451.189.000
300 via PCA	45.12869858070724	698.9823195934296	30.870.713.400
100 via PCA	43.05989896560019	670.4753487110138	10.290.237.800
2 via PCA	52.97089247053164	629.3050184249878	205.804.756
2 via Average Brightnesses	46.54799133990859	570.3843824863434	205.804.756

Η επίδοση της μεθόδου μειώνεται αφού όταν συμβάλλουν τρεις γείτονες στην απόφαση κατηγοριοποίησης προκύπτουν λάθη σε οριακά δείγματα των κλάσεων. Αντιθέτως, για είσοδο που αποτελείται από δείγματα με 2 χαρακτηριστικά η μέθοδος με τρεις γείτονες βοηθάει στην καταπολέμηση του προβλήματος της λάθος πρόβλεψης σε οριακά δείγματα που βρίσκονται ανάμεσα σε δύο κλάσεις τα οποία είναι πολλά όπως αποτυπώνεται και στα σχήματα 2 και 3. Η χρονική επίδοση του κατηγοριοποιητή δεν επηρεάζεται από την αύξηση του αριθμού των γειτόνων αφού δεν επηρεάζεται η βασική πράξη της μεθόδου.

5. 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 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 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 διάφορα χαρακτηριστικά εισόδου και για ευκλείδειες αποστάσεις:

# of Characteristics	Accuracy Percentage	Time Elapsed in Seconds	Basic Operations
Whole Input (784)	91.72480153957181	16.601562976837158	19.407.136
500 via PCA	40.004811161895596	8.968602180480957	12.377.000
300 via PCA	40.004811161895596	5.4434897899627686	7.426.200
100 via PCA	40.052922780851574	2.23103666305542	2.475.400
2 via PCA	53.11522732739957	0.25918030738830566	49.508
2 via Average Brightnesses	50.084195333172964	0.1655721664428711	49.508

Η επίδοση της μεθόδου σε σύγκριση με τον κατηγοριοποιητή K πλησιέστερων γειτόνων είναι σαφώς χειρότερη. Οι περιπτώσεις των οποίων η επίδοση είναι ανώτερη, είναι αυτές στις οποίες γίνεται χρήση μόνο 2 χαρακτηριστικών για κάθε δείγμα. Αυτό ίσως οφείλεται στο γεγονός ότι με την χρήση του κέντρου σε αντίθεση με τους K γείτονες μπορεί να καταπολεμηθεί η έλλειψη τοπικότητας των δειγμάτων και να επιτευχθεί καλύτερη κατηγοριοποίηση.

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 προϋπάρχον δείγμα αλλά από το κέντρο των τεσσάρων 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 784 χαρακτηριστικά οδηγεί σε 19.407.136 ενώ για 2 σε 49.508 βασικές πράξεις. Ο πραγματικός χρόνος που δαπανάται από το πρόγραμμα δεν αποτελεί το καλύτερο μέτρο σύγκρισης όπως έχει ήδη εξηγηθεί.

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`.

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 Minkowski βελτιώνει την επίδοση του κατηγοριοποιητή πλησιέστερου κέντρου ενώ δεν επηρεάζει τον χρόνο δραματικά. Επίσης όσο μεγαλώνει το όρισμα της (p), τόσο χειροτερεύει η επίδοσή της, αλλά μόνο για την κατηγοριοποίηση με όλα τα χαρακτηριστικά των δειγμάτων (784). Αυτό μπορεί να οφείλεται στην χρήση της μεθόδου PCA και στο γεγονός ότι αυτή ευνοεί την χρήση αποστάσεων υψηλότερων διαστάσεων.
- Η απόδοση του αλγορίθμου πλησιέστερου κέντρου για τα 2 χαρακτηριστικά της μεθόδου μέσης φωτεινότητας παραμένει σταθερή ανεξαρτήτως της απόστασης που επιλέγεται αφού μιλάμε για σημεία επάνω στην ίδια ευθεία, την $x = y$.
- Η απόσταση Manhattan μειώνει κατά πολύ το ποσοστό επιτυχίας της κατηγοριοποίησης των μεθόδων αφού απλοποιεί κατά πολύ τον υπολογισμό των αποστάσεων.
- Η χρονική απόδοση για την απόσταση Minkowski για τον κατηγοριοποιητή K πλησιέστερων γειτόνων, είναι απαγορευτική, χωρίς να βελτιώνει δραματικά την επίδοση την μεθόδου. Ακόμη και αν η θεωρητική πολυπλοκότητα του αλγορίθμου παραμένει 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.
- Υπάρχει δραματική μείωση της απόδοσης της μεθόδου K πλησιέστερων γειτόνων, για την μετρική Minkowski, και για δείγματα με 500,300 και 100 χαρακτηριστικά τα οποία προήλθαν από την χρήση της PCA.

6. Νευρωνικό Δίκτυο

Το νευρωνικό δίκτυο το οποίο απαιτείται στα πλαίσια της εργασίας προορίζεται για την κατηγοριοποίηση των ψηφίων της βάσης σε μία από τις πιθανές κλάσεις και θα αναπτυχθεί με χρήση του πακέτου λογισμικού `torch` το οποίο είναι διαθέσιμο για την έκδοση 3.8 της γλώσσας προγραμματισμού Python. Είναι εύκολο να κατανοήσει κανείς ότι στο επίπεδο εισόδου του δικτύου ο αριθμός των νευρώνων θα ισούται με 784 ενώ σε αυτό της εξόδου με 4. Στα κρυφά επίπεδα μπορούν να γίνουν διάφορες επιλογές για τον αριθμό των νευρώνων. Για το συγκεκριμένο πρόβλημα θα γίνει χρήση ενός πλήρως συνδεδεμένου νευρωνικού δικτύου. Επίσης για την συνάρτηση ενεργοποίησης κάθε νευρώνα του δικτύου γίνεται χρήση της συνάρτησης `ReLU` λόγω της δημοφιλίας της σε σχέση με τις υπόλοιπες επιλογές (`Linear`, `Sigmoid`, `Tanh`). Επίσης γίνεται χρήση της συνάρτησης `LogSoftMax()` στην έξοδο του δικτύου με σκοπό την αποτύπωση των αποτελεσμάτων σε διάστημα πιθανότητας από 0 έως 1. Η εν λόγω συνάρτηση στην έξοδο του δικτύου, δίνει καλύτερη επίδοση από άποψη υπολογιστικού κόστους και ταχύτητας εκπαίδευσης. Τέλος εφόσον γίνεται χρήση της `LogSoftMax()`, θα γίνει χρήση της `nll_loss()` συνάρτησης κόστους που προσφέρεται από το πακέτο, και η οποία λειτουργεί λαμβάνοντας εισόδους πιθανότητες στο διάστημα $[0, 1]$.

Τα βήματα τα οποία απαιτούνται για την εκπαίδευση κάθε νευρωνικού δικτύου, και κατ' επέκταση και στο συγκεκριμένο δίκτυο, είναι τα εξής:

- Για κάθε εποχή εκπαίδευσης και για κάθε δέσμη δεδομένων ακολούθησε την παρακάτω διαδικασία.
- Υπολόγισε την έξοδο την οποία δίνει το δίκτυο όταν δεχεται σειριακά ως είσοδο τα δείγματα της δέσμης.
- Υπολόγισε την συνάρτηση κόστους που έχει επιλεχθεί για το δίκτυο, με είσοδο τα αποτελέσματα κατηγοριοποίησης του δικτύου για τα δείγματα της δέσμης και τις πραγματικές ετικέτες αυτών των δειγμάτων.
- Κάνε χρήση του κανόνα Back-Propagation για να ενημερώσεις κατάλληλα τα βάρη ανάμεσα στους νευρώνους του δικτύου.
- Εάν το κόστος που παρατηρήθηκε στην εποχή είναι μικρότερο από μία προκαθορισμένη παράμετρο, τερμάτισε την εκπαίδευση.

Κατά την διάρκεια της εκπαίδευσης πολλές παράμετροι μπορούν να αλλάξουν δίνοντας διαφορετικά πλεονεκτήματα/μειονεκτήματα στην τελική απόδοση του δικτύου. Για την παράμετρο εκπαίδευσης β , ξεκινώντας από μία τιμή ίση με 0.01, γίνεται έλεγχος για τα αποτελέσματα που έχει η μείωση αυτής, κατά μία τάξη μεγέθους κάθε φορά, έως ότου φτάσει την τιμή 0.000001. Παρατηρείται πως αρχικά η τιμή της παραμέτρου είναι υπερβολικά υψηλή για το δίκτυο και έτσι καταλήγει να ταλαντώνει ανάμεσα σε απόμακρα σημεία της συνάρτησης που προσπαθεί να προσεγγίσει. Με την σταδιακή μείωση του β αυτή η ταλάντωση παύει να υφίσταται και το δίκτυο παρουσιάζει σύγκλιση προς το ολικό ελάχιστο, μειώνοντας διαρκώς την αριθμητική τιμή της συνάρτησης κόστους. Όταν πλέον η τιμή της β είναι υπερβολικά μικρή, η παράμετρος δεν έχει το επιθυμητό αποτέλεσμα αφού στην ουσία 'παγιδεύει' το δίκτυο υπο εξέταση σε κάποιο τοπικό ελάχιστο χωρίς αυτό να έχει την περαιτέρω δυνατότητα απεγκλωβισμού από αυτό το ελάχιστο. Μία λύση για το πρόβλημα αυτό είναι η χρήση της λεγόμενης ορμής, η οποία δίνει την επιπλέον 'ώθηση' που χρειάζεται το δίκτυο για να ανταπεξέλθει στα τοπικά ελάχιστα.

Μία άλλη σημαντική παρατήρηση για τις παραμέτρους που επηρεάζουν την διαδικασία εκπαίδευσης είναι η σχέση που έχει το μέγεθος της δέσμης με τις επιδόσεις που επιτυγχάνονται από το δίκτυο και τον χρόνο που δαπανάται στην κάθε εποχή. Όσο το μέγεθος της εκάστοτε δέσμης αυξάνεται τόσο γρηγορότερα ολοκληρώνεται η εκπαίδευση του μοντέλου. Αντιθέτως, με την αύξηση αυτή, οι επιδόσεις του δικτύου μειώνονται δραστικά, και έτσι για ίδιο αριθμό εποχών με μεγαλύτερο μέγεθος δέσμης, η συνάρτηση κόστους λαμβάνει πολύ μεγαλύτερες τιμές. Αυτό το αποτέλεσμα λαμβάνει χώρα επειδή το δίκτυο δεν καταφέρνει να 'παγιδεύσει' αρκετή πληροφορία στα βάρη του μέσω των αλλαγών που προκαλεί ο αλγόριθμος οπισθοδρόμησης για μεγάλα μεγέθη δέσμης, αφού σε κάθε πέρασμα του αλγορίθμου εκπαίδευσης, χρησιμοποιείται μεγάλος αριθμός δειγμάτων. Γι' αυτό και οι συσχετίσεις ανάμεσα στα δεδομένα εκπαίδευσης δεν καταγράφονται ικανοποιητικά από το μοντέλο.

Μία πολύ καλή επίδοση δίνεται για δίκτυο τεσσάρων στρωμάτων (με δύο κρυφά στρώματα εξηνητα-τεσσάρων νευρώνων το καθένα), μέγεθος δέσμης = 64, ρυθμός εκπαίδευσης $\beta = 0.0001$ και 50 εποχές. Εφόσον γίνεται τυχαία ανάμιξη των δεδομένων πριν την έναρξη του αλγορίθμου εκπαίδευσης, και αφού η σειρά εισόδου των δεδομένων παίζει ρόλο στην εκπαίδευση του δικτύου, η ακρίβεια της κατηγοριοποίησης μπορεί να αυξομειώνεται για τις συγκεκριμένες παραμέτρους και διαφορετικές εκτελέσεις της υλοποίησης.

Προφανώς και το νευρωνικό δίκτυο με τις κατάλληλες παραμέτρους εκπαίδευσης έχει επιδόσεις που δεν μπορούν να προσεγγιστούν με τις μεθόδους των Κ Πλησιέστερων Γειτόνων και του Πλησιέστερου Κέντρου. Επίσης μπορεί να υποστηριχθεί ότι η εκπαίδευση του δικτύου είναι γρηγορότερη από την μέθοδο των πλησιέστερων γειτόνων. Παρ' όλ' αυτά η εύρεση των καλύτερων δυνατών παραμέτρων δεν είναι εύκολη και είναι αρκετά χρονοβόρα. Έτσι μπορεί να εξαχθεί σαν γενικό συμπέρασμα ότι για μεγάλο σύνολο δεδομένων τα νευρωνικά δίκτυα είναι ανώτερα των κλασικών μεθόδων κατηγοριοποίησης. Όμως για απλότερα προβλήματα όπως αυτό που αναλύθηκε στην παρούσα εργασία, η επίδοσης των κλασικών αλγορίθμων μπορεί να είναι αρκετά ικανοποιητική, και έτσι η επιλογή του

καταλληλότερου κατηγοριοποιητή εναποφύεται στην κρίση του χρήστη και το πρόβλημα κατηγοριοποίησης που αυτός καλείται να αντιμετωπίσει.

Τα κύρια προβλήματα που αντιμετώπισα κατά την ενασχόληση μου με το νευρωνικό δίκτυο και το πακέτο λογισμικού torch ήταν τέσσερα, και σας τα παραθέτω παρακάτω:

- Πως ακριβώς καθορίζεται ο καταλληλότερος αριθμός επιπέδων και νευρώνων ανα επίπεδο για το δίκτυο μας πέραν της μεθόδου της δοκιμής και πλάνης.
- Δεν κατανοήθηκε πλήρως η ανάγκη της χρήσης των συναρτήσεων `network.train()/network.eval()`.
- Δεν κατανοήθηκε το πλήρες φάσμα στο οποίο μπορούν να χρησιμοποιηθούν διαφορετικές συναρτήσεις κόστους αναλογα με την συνάρτηση στο τελευταίο επίπεδο του κάθε δικτύου και ποίοι είναι οι βέλτιστοι συνδυασμοί.
- Δεν κατανοήθηκαν πλήρως οι λόγοι οι οποίοι θα ωθούσαν κάποιον στο να χρησιμοποιήσει κάποια μέθοδος μείωσης χαρακτηριστικών για τα δείγματα στην είσοδο του νευρωνικού δικτύου