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

Created with LaTeX

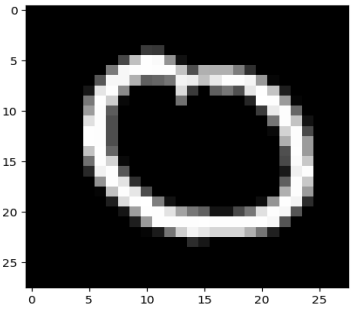
0 1 2 3 4 5 6 7 8 9 10

1. Περίληψη

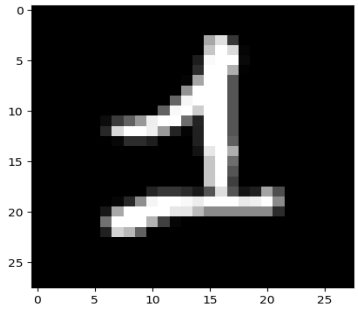
Η παρούσα εργασία εκπονήθηκε στο πλαίσιο του προπτυχιακού μαθήματος «Νευρωνικά Δίκτυα και Βαθιά Μάθηση», το οποίο πραγματοποιήθηκε κατά το χειμερινό εξάμηνο του διδακτικού έτους 2020-2021 στο Τμήμα Πληροφορικής του Αριστοτελείου Πανεπιτημίου Θεσσαλονίκης, και αποτελεί μια εκτενή ενασχόληση με τις μεθόδους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

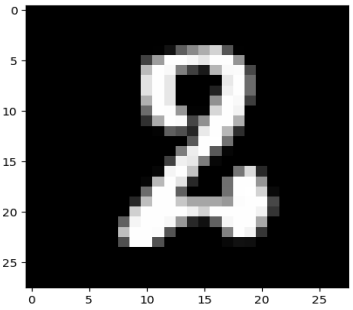
Για τους σκοπούς της εργασίας επιλέχθηκε η δημόσια βάση χειρόγραφων ψηφιών MNIST η οποία μπορεί να βρεθέι στην ηλεκτρονική διεύθυνση `http://yann.lecun.com/exdb/mnist/`. Ένα χαρακτηριστικό δείγμα των δεδομένων της εν λόγω βάσης παρουσιάζεται στις παρακάτω εικόνες.



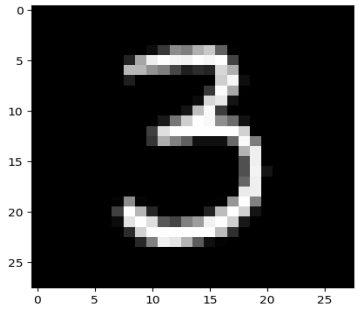
Το ψηφίο 0



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 τέσσερεις. Εάν επιπλέον κάνουμε και διάκριση των κλάσεων σε μονά και ζυγά ψηφία, μένουμε με μόλις δύο κλάσεις.

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.

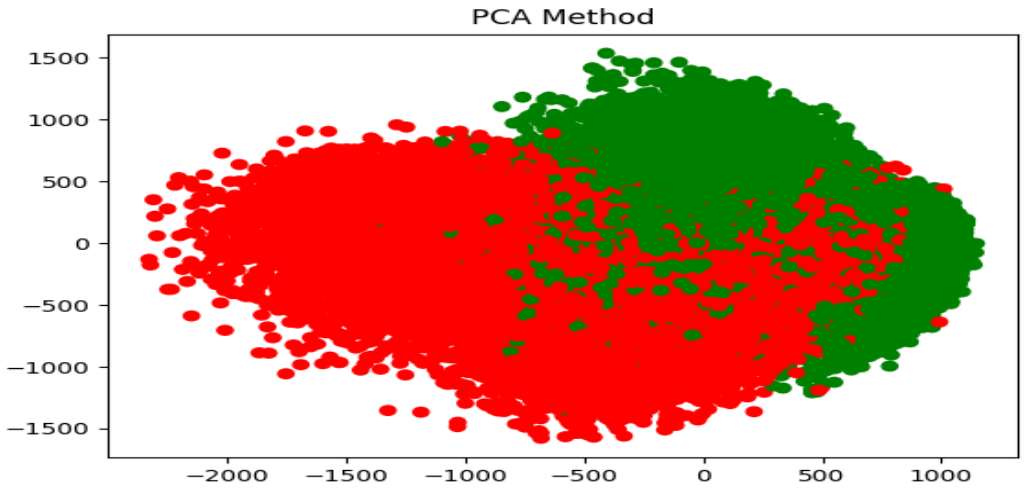
Είναι εύλογο να θελήσει κάποιος να μειώσει τον αριθμό των χαρακτηριστικών αυτών, δίχως να χάνεται πληροφορία για τα δείγματα της βάσης, η οποία θα μπορούσε να φανεί πολύτιμη για την επίλυση προβλημάτων από τις μεθόδους που θα αναπτυχθούν. Έχοντας αυτή την συλλογιστική κατα νού, προτάσσεται η χρήση της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

3.2. 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 28x28.
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 εάν το ψηφίο το οποίο εμφανίζεται στην εικόνα αποτελεί άρτιο η περιττό αριθμό. Σε κάθε κλάση ανατίθεται ένα χρώμα. Στην προκειμένη περίπτωση, ισχύουν οι εξής αναθέσεις: κόκκινο για τα άρτια ψηφία και πράσινο για τα περιττά ψηφία. 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 $V = 2$

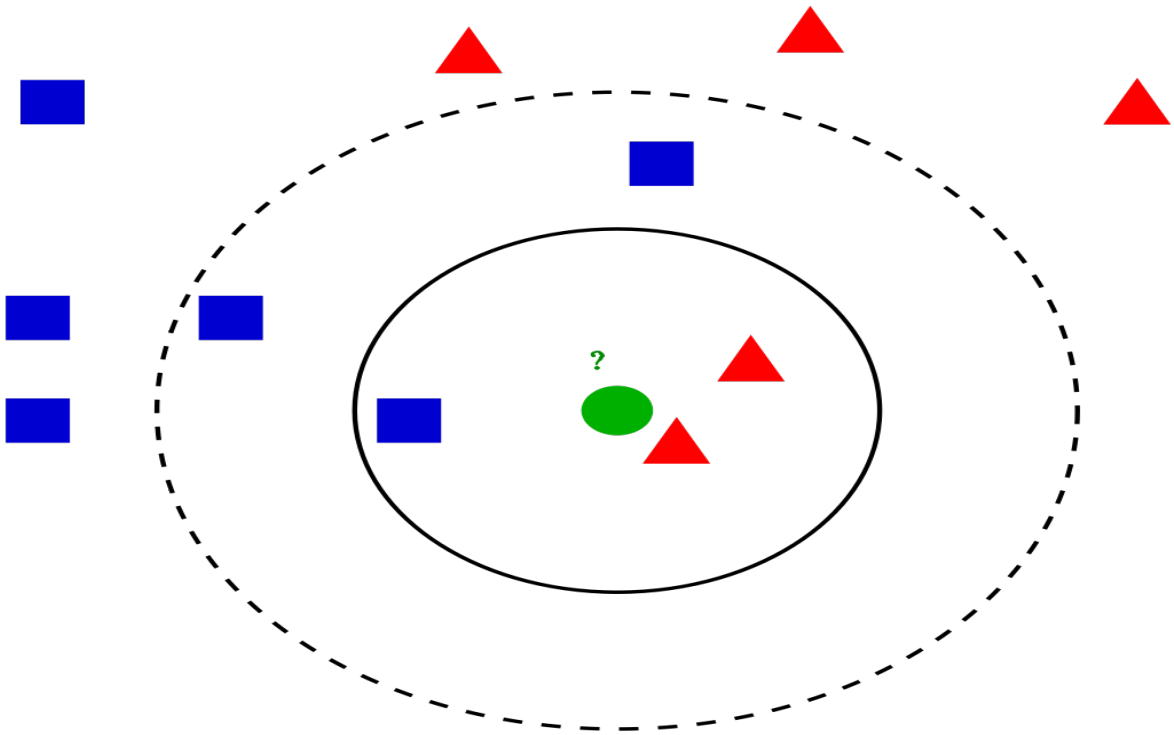
4. Μετρικές Απόστασης

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

- Ευκλείδειες αποστάσεις με χρήση της $np.linalg.norm(\vec{point}_1 - \vec{point}_2)$.
- Αποστάσεις Minkowski με χρήση της scipy.spatial.distance.minkowski για όρισμα $p = 3$.
- Αποστάσεις Manhattan με χρήση της scipy.spatial.distance.cityblock.
- Αποστασεις Chebyshev με χρήση της scipy.spatial.distance.chebyshev.

5. Κατηγοριοποιητής Κ Πλησιέστερων Γειτόνων

Ο κητηγοριοποιητής 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 κ, καθώς και το/τα σημείο/σημεία που καλείται να κατηγοριοποιήσει. Η παράμετρος κ 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 απο 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, γινεται 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 κ, 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 Κ κοντινότερα δείγματα βάση της απόστασης που υπολόγισες.

- Λάβε την απόφαση κατηγοριοποίησης βάση της συχνότερης ετικέτας στα Κ επιλεγμένα δείγματα.

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

Distance Metric Used	Training Time	Training Accuracy	Testing Accuracy
Euclidean	0.1	100%	99.59%
Minkowski	0.1	100%	99.61%
Manhattan	0.1	100%	99.47%
Chebyshev	0.1	100%	99.45%

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 Κ = 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 τέτοιον υπολογισμό απαιτούνται βασικές πράξεις ίσες με τα χαρακτηριστικά κάθε δείγματος. Έτσι στην συγκεκριμένη περίπτωση ο αλγόριθμος χαρακτηρίζεται από πολυπλοκότητα $O(4157 * 24754 * number\ of\ characteristics\ in\ each\ sample)$, η οποία για 75 χαρακτηριστικά οδηγεί σε 7.717.678.350 βασικές πράξεις. Ο πραγματικός χρόνος που δαπανάται από το πρόγραμμα δεν αποτελέι το καλύτερο μέτρο σύγκρισης αφού εξαρτάται κατά πολύ από το εκάστοτε υπολογσιτικό σύστημα, της δυνατότητες του σε πόρους καθώς και τον φόρτο εργασίας του συστήματος απο άλλους παράγοντες πέραν του αλγορίθμου.

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

Distance Metric Used	Training Time	Training Accuracy	Testing Accuracy
Euclidean	0.1	99.798%	99.64%
Minkowski	0.1	99.802%	99.69%
Manhattan	0.1	99.733%	99.40%
Chebyshev	0.1	99.729%	99.54%

Η επίδοση της μεθόδου μειώνεται αφού όταν συμβάλλουν τρείς 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 Κ-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 εχει 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 κ, 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 διαιρόντας 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	88.00%	87.59%
Minkowski	1.5	87.66%	86.96%
Manhattan	1.5	88.53%	87.85%
Chebyshev	1.5	84.69%	84.39%

Η επίδοση της μεθόδου σε σύγκριση με τον κατηγοριοποιητή Κ πλησιέστερων γειτόνων είναι σαφώς χειρότερη. Αυτό είναι προφανώς αναμενόμενο αφού πλεόν 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.

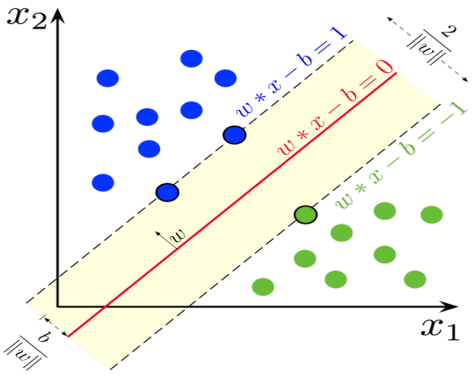
Απο άποψη χρονικής επίδοσης οι δύο μέθοδοι είναι η μέρα με την νύχτα. Δεν υπάρχει καμία σύγκριση για το ποιά είναι ανώτερη σε αυτό το κομμάτι. Αυτή η παρατήρηση είναι προφανώς αναμενόμενη, και αυτό διότι, για κάθε νέο δείγμα η κατηγοριοποίηση του δεν απαιτεί τον υπολογισμό της απόστασης του από κάθε 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 τέτοιον υπολογισμό απαιτούνται βασικές πράξεις ίσες με τα χαρακτηριστικά κάθε δείγματος. Έτσι στην συγκεκριμένη περίπτωση ο αλγόριθμος χαρακτηρίζεται από πολυπλοκότητα $O(24754 * number\ of\ characteristics\ in\ each\ sample)$ η οποία για 75 χαρακτηριστικά οδηγεί σε 1.856.550 βασικές πράξεις. Ο πραγματικό χρόνος που δαπανάται απο το πρόγραμμα δεν αποτελέι το καλύτερο μέτρο σύγκρισης όπως έχει ήδη εξηγηθεί.

Οι παρακάτω γενικές παρατηρήσεις αξίζει να αναφερθούν σχετικά με τις δοκιμές που έλαβαν χώρα με τις προαναφερθήσες 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.
- Η απόσταση Manhattan μειώνει κατα πολύ το ποσοστό επιτυχιάς του κατηγοριοποιητή Κ-πλησιέστερων γειτόνων αφού απλοποιεί κατά πολύ τον υπολογισμό των αποστάσεων δίνοντας περισσότερα λάθη σε οριακές κατηγοριοποιήσεις.
- Η χρονική απόδοση για την απόσταση Minkowski για τον κατηγοριοποιητή Κ πλησιέστερων γειτόνων, είναι απαγορευτική, χωρίς να βελτιώνει δραματικά την επίδοση την μεθόδου. Ακόμη και άν η θεωρητική πολυπλοκότητα του αλγορίθμου παραμενει η ίδια, η πρακτική πολυπλοκότητα η οποία εξαρτάται από το πόσο γρήγορα υπολογίζεται η απόσταση ανάμεσα στα σημεία καθιστά την μέθοδο των Κ 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 Ευκλείδια σε σχέση με τις πίο περίπλοκες Minkowski και Chebyshev.

7. Μηχανή Διανυσμάτων Υποστήριξης

Η Μηχανές Διανυσάτων Υποστήριξης (SVM) βασίζονται στην ιδέα που αποτυπώνεται κατάλληλα στο Σχήμα 4.

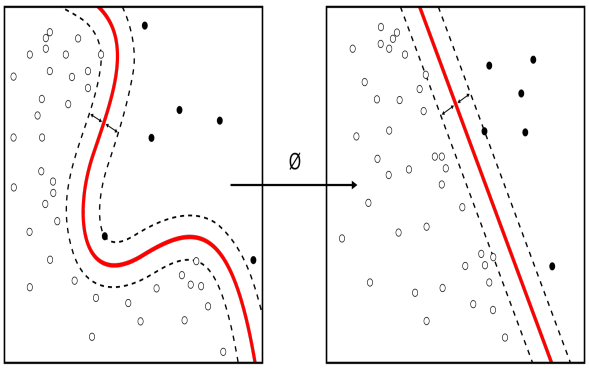


Σχήμα 4: Source: https://en.wikipedia.org/wiki/File:SVM_margin.png

Στην εν λόγω εικόνα παρουσιάζονται τα δείγματα δύο διαφορετικών κλάσεων τις οποίες επιθυμούμε να διαχωρίσουμε κατάλληλα. Ανάγοντας την εικόνα στο πρόβλημα το οποίο εξετάζεται στην παρούσα εργασία, έστω ότι τα μπλέ δείγματα ανταποκρίνονται σε μονούς αριθμούς, ενώ τα πράσινα δείγματα σε ζυγούς. Ένας εύκολος τρόπος διαχωρισμού των δειγμάτων θα ήταν η εύρεση κάποιας διαχωριστικής υπερεπιφάνειας η οποία να επιτυγχάνει την διχοτόμηση των δειγμάτων, έτσι ώστε κάθε κλάση να έχει δείγματα της μόνο στην μία από τις δύο πλευρές της υπερεπιφάνειας αυτής. Επιπλεόν μία ακόμη επιθυμητή ιδιότητα θα ήταν η μεγιστοποίηση της απόστασης των οριακών δειγμάτων των δύο κλάσεων από την επιλεγμένη υπερεπιφάνεια. Τα οριακά δείγματα είναι και τα μόνα που εν τέλει παίζουν ρόλο στην επιλογή της υπερεπιφάνειας, αφού με διαγραφή των υπολοίπων δειγμάτων, δεν επηρεάζεται η επιλογή της, και αποκαλούνται διανύσματα υποστήριξης. Με την ιδιότητα αυτή επιτυγχάνεται η ελαχιστοποίηση της πιθανότητας λανθασμένης κατηγοριοποίησης για οριακά δείγματα που ίσως προκύψουν μετά την εκπαίδευση. Αυτού του τύπου η μηχανή αποκαλείται Γραμμική Μηχανή Διανυσμάτων Υποστήριξης (LSVM) και αποτελεί την απλούστερη μορφή SVM.

Τι συμβαίνει όμως όταν το πρόβλημα που καλείται η μηχανή να αντιμετωπίσει δεν είναι γραμμικά διαχωρίσιμο;

Η μηχανή πλέον καλείται να αντιμετωπίση ένα πρόβλημα το οποίο δεν μπορεί να διαχωριστεί με μία ευθεία και επομένως η απλή Γραμμική Μηχανή Διανυσμάτων Υποστήριξης δεν μπορεί να βοηθήσει. Έτσι για την αντιμετώπιση του προβλήματος πρέπει να γίνει χρήση κάποιου μεταχηματισμού, με σκοπό να μετατραπεί ο χώρος των δειγμάτων όπως φαίνεται στο Σχήμα 5.



Σχήμα 5: Source: https://commons.wikimedia.org/wiki/File:Kernel_Machine.svg

Ο μετασχηματισμός αυτός γίνεται με χρήση των Συναρτήσεων Πυρήνα (Kernel Machines) οι οποίες επιτυγχάνουν την μεταφορά των δειγμάτων σε έναν χώρο στον οποίο αυτά είναι πλεόν γραμμικά διαχωρίσιμα. Η πιο συχνές Συναρτήσης Πυρήνα είναι οι παρακάτω:

- Γραμμικός Πυρήνας: $k(x, y) = x^T * y$ ο οποίος ουσιαστικά χρησιμοποιείται από τα LSVM.
- Πολυωνυμικός Πυρήνας: $k(x, y) = (g * x^T * y + c)^d$ όπου d ο βαθμός του πολυωνύμου, c κάποια σταθερά και g κάποιος παράγοντας κανονικοποίησης.
- Γκαουσιανός Πυρήνας: $k(x, y) = exp\{-g * ||x - y||^2\}$ όπου g κάποιος παράγοντας κανονικοποίησης.
- Σιγμοειδής Πυρήνας: $k(x, y) = tanh(g * x^T * y + c)$ όπου c κάποια σταθερά και g κάποιος παράγοντας κανονικοποίησης.

Μία ακόμη σημαντική παράμετρος των SVMs είναι η λεγόμενη ανοχή σε λάθη κατηγοριοποίησης, η οποία συμβολίζεται με C. Υπάρχουν και SVMs που κάνουν χρήση άλλων παραμέτρων και όχι της C, όπως π.χ. τα nuSVM τα οποία, όπως δηλώνει και το όνομα τους, κάνουν χρήση της παραμέτρου nu, η οποία αποτελεί ένα άνω όριο για το ποσοστό των οριακών λαθών και ένα κάτω όριο για το ποσοστό των διανυσμάτων υποστήριξης σε σχέση με το σύνολο των δειγμάτων εκπαίδευσης. Για παράδειγμα, για nu = 0.05 η μηχανή μπορεί να εγγυηθεί ότι θα προκύψουν το πολύ 5% λάθος κατηγοριοποιήσεις και ο λιγότερος αριθμός διανυσμάτων υποστήριξης σε σχέση με τα δείγματα εκπαίδευσης θα είναι 5%. Στην παρούσα εργασία θα εξετασθούν μόνο τα C-SVMs, με χρήση των συναρτήσεων πυρήνα που αναφέρθηκαν παραπάνω.

7.1. Γραμμική Μηχανή Διανυσμάτων Υποστήριξης

Για την υλοποίηση της Γραμμικής Μηχανής Διανυσμάτων Υποστήριξης (LSVM) χρησιμοποιείται το πακέτο λογισμικού sklearn.svm και πιο σγκεκριμένα η μέθοδος SVC που παρέχεται από αυτό, με χρήση της γραμμικής συνάρτησης πυρήνα ('linear'). Για την εύρεση της καλύτερης παραμέτρου εκπαίδευσης C, έγινε χρήση της μεθόδου 10-Fold Cross Validation, μέσω της μεθόδου που επίσης παρέχεται από το εν λόγω πακέτο λογισμικού, την GridSearchCV. Οι παράμετροι που δοκιμάστηκαν κατά την διαδικάσια αυτή είναι οι εξής: C = [0.001, 0.01, 0.1, 1, 10, 100]. Η δοκιμαστική αυτή διαδικασία καταλήγει στο συμπέρασμα ότι η καλύτερη, από τις πιθανές τιμές της παραμέτρου C, είναι η 0.1 με τελική απόδοση ίση με 0.976.

7.2. Μή-Γραμμική Μηχανή Διανυσμάτων Υποστήριξης

7.2.1. Πολυωνυμικός Πυρήνας

Για την υλοποίηση της Μή-Γραμμικής Μηχανής Διανυσμάτων Υποστήριξης (SVM) με Πολυωνυμικό Πυρήνα (polynomial) χρησιμοποιείται το πακέτο λογισμικού sklearn.svm και πιο σγκεκριμένα η μέθοδος SVC που παρέχεται από αυτό, με χρήση της γκαουσιανής συνάρτησης πυρήνα ('poly'). Για την εύρεση των καλύτερων παραμέτρων εκπαίδευσης C, gamma, coef0 και degree, έγινε χρήση της μεθόδου 10-Fold Cross Validation, μέσω της μεθόδου που επίσης παρέχεται από το εν λόγω πακέτο λογισμικού, την GridSearchCV. Οι παράμετροι που δοκιμάστηκαν κατά την διαδικάσια αυτή είναι οι εξής: C = [0.001, 0.01, 0.1, 1, 10, 100], gamma = [0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000], coef0 = [0, 1] και degree = [2, 3, 4, 5, 6]. Η δοκιμαστική αυτή διαδικασία καταλήγει στο συμπέρασμα ότι ο καλύτερος, από τους 480 πιθανούς συνδυασμούς για τις παραμέτρους C, gamma, coef0 και degree, είναι ο C = 1, gamma = 0.1, coef0 = 0 και degree = 3 με τελική απόδοση ίση με 0.995. Μέτα την δοκιμή της τιμής που δίνεται αυτόματα από την μέθοδο στην παράμετρο gamma, και για C = 10 παρατηρείται καλύτερη απόδοση και έτσι προτιμούνται οι συγκεκριμένες παράμετροι εκπαίδευσης.

7.2.2. Γκαουσιανός Πυρήνας

Για την υλοποίηση της Μή-Γραμμικής Μηχανής Διανυσμάτων Υποστήριξης (SVM) με Γκαουσιανό Πυρήνα (RBF) χρησιμοποιείται το πακέτο λογισμικού sklearn.svm και πιο σγκεκριμένα η μέθοδος SVC που παρέχεται από αυτό, με χρήση της γκαουσιανής συνάρτησης πυρήνα ('rbf'). Για την εύρεση των καλύτερων παραμέτρων εκπαίδευσης C και gamma, έγινε χρήση της μεθόδου 10-Fold Cross Validation, μέσω της μεθόδου που επίσης παρέχεται από το εν λόγω πακέτο λογισμικού, την GridSearchCV. Οι παράμετροι που δοκιμάστηκαν κατά την διαδικάσια αυτή είναι οι εξής: C = [0.001, 0.01, 0.1, 1, 10, 100], gamma = [0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000]. Η δοκιμαστική αυτή διαδικασία καταλήγει στο συμπέρασμα ότι ο καλύτερος, από τους 48 πιθανούς συνδυασμούς για τις παραμέτρους C και gamma, είναι ο C = 10, gamma = 0.01, με τελική απόδοση ίση με 0.995. Μέτα την δοκιμή της τιμής που δίνεται αυτόματα από την μέθοδο στην παράμετρο gamma, και για C = 10 παρατηρείται καλύτερη απόδοση και έτσι προτιμούνται οι συγκεκριμένες παράμετροι εκπαίδευσης.

7.2.3. Σιγμοειδής Πυρήνας

Για την υλοποίηση της Μή-Γραμμικής Μηχανής Διανυσμάτων Υποστήριξης (SVM) με Σιγμοειδή Πυρήνα (sigmoid) χρησιμοποιείται το πακέτο λογισμικού sklearn.svm και πιο σγκεκριμένα η μέθοδος SVC που παρέχεται από αυτό, με χρήση της γκαουσιανής συνάρτησης πυρήνα ('sigmoid'). Για την εύρεση των καλύτερων παραμέτρων εκπαίδευσης C, gamma και coef0, έγινε χρήση της μεθόδου 10-Fold Cross Validation, μέσω της μεθόδου που επίσης παρέχεται από το εν λόγω πακέτο λογισμικού, την GridSearchCV. Οι παράμετροι που δοκιμάστηκαν κατά την διαδικάσια αυτή είναι οι εξής: C = [0.001, 0.01, 0.1, 1, 10, 100], gamma = [0.001, 0.01, 0.1, 1, 10, 100] και coef0 = [0, 1]. Η δοκιμαστική αυτή διαδικασία καταλήγει στο συμπέρασμα ότι ο καλύτερος, από τους 7 2 πιθανούς συνδυασμούς για τις παραμέτρους C, gamma και coef0, είναι ο C = 100, gamma = 0.001, coef0 = 0 με τελική απόδοση ίση με 0.975.

7.3. Αποτελέσματα

Distance Metric Used	Training Time	Training Accuracy	Testing Accuracy
Linear	5.5	97.74%	98.12%
Polynomial	6.2	99.92%	99.66%
Gaussian (RBF)	2.8	99.96%	99.76%
Sigmoid	6.9	97.69%	98.03%

8. Σχολιασμός Αποτελεσμάτων και Σύγκριση Μεθόδων

Η πρώτη μέθοδος που εξετάστηκε είναι ο Κατηγοριοποιητής Κ Πλησιέστερων Γειτόνων για Κ = 1 και Κ = 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 Κ, 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 Κ = 1, η ακρίβεια στα δείγματα εκπαίδευσης είναι 100%. Αυτή η τέλεια ακρίβεια επιτυγχάνεται αφού κάθε δείγμα εκπαίδευσης έχει ως μοναδικό δείγμα αναφοράς κατά την κατηγοριοποίση του τον ευατό του και έτσι λαμβάνει εγγυημένα την σωστή ετικέτα κλάσης. Για Κ = 3 αυτή η τέλεια ακρίβεια χάνεται και παρατηρείται μία λίγο χειρότερη ακρίβεια (99.8%). 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 τελικη απόφαση είναι λανθασμένη.

Για τα δοκιμαστικά δείγματα προς κατηγοριοποίηση, παρατηρείται, για K = 1 ακρίβεια της τάξης του 99.59%, και για K = 3 ακρίβεια της τάξης του 99.64% για Ευκλείδειες αποστάσεις, οι οποίες παρουσιάζουν την καλύτερη σχέση χρονικής επίδοσης και ακρίβειας. Φαίνεται πως η ακρίβεια του Κατηγοριοποιητή Κ Πλησιεστερων Γειτόνων τείνει να αυξάνει για νέα δείγματα προς κατηγοριοποίηση, καθώς αυξάνει το Κ, χωρίς να επηρεάζεται η χρονική επίδοση της μεθόδου αφού η επιλογή των Κ κοντινότερων γειτόνων ενός δείγματος δεν είναι η βασική πράξη της μεθόδου. Αυτή η συμπεριφορά είναι απολύτως αναμενόμενη αφού τα δείγματα της ίδιας κλάσης χαρακτηρίζονται από τοπικότητα και καθώς αυξάνεται η τιμή του Κ, όλο και περισσότερα δείγματα της σωστής κλάσης θα συσχετίζονται με το προς κατηγοριοποίηση δείγμα και θα λαμβάνονται υπόψην από την μέθοδο πρίν την απόφαση κατηγοριοποίησης του νέου δείγματος δοκιμής. Παρ΄όλάυτά το Κ δεν θα πρέπει να ξεπερνά κάποια τιμή η οποία εγγυάται την τοπικότητα που αναφέρθηκε προηγουμένως αφού θα εμφανίζονται όλο και περισσότερα σφάλματα για οριακές περιπτώσεις ανάμεσα σε κλάσεις. Συνήθως το Κ λαμβάνει τις τιμές 3, 5, 7, ανάλογα με το πρόβλημα.

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 κατηγοριοποίση είναι ελάχιστος σε σχέση με την μέθοδο των Κ 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 88% ακρίβεια για τα δείγματα εκπαίδευσης και 87.6% 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 κατηγοριοποίηση τους. Εάν σχηματίσουμε έναν κύκλο γύρο 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, εξετάσθηκαν οι Μηχανές Διανυσμάτων Υποστήριξης. Πρίν δοκιμασθεί οποιαδήποτε μηχανή θα πρέπει να βρεθεί το κατάλληλο σύνολο παραμέτρων για την συγκεκριμένη μηχανή και το συγκεκριμένο πρόβλημα που καλείται να αντιμετωπίσει. Αυτό γίνεται μέσω της διαδικασίας 10-Fold Cross Validation για όλες τις μηχανές που δοκιμάστηκαν στην παρούσα εργασία. Η εν λόγω διαδικασία μπορεί να είναι εξαιρετικά αργή για μεγάλα σύνολα παραμέτρων και ενδεχομένως να διαρκέσει πολλές ώρες. Αυτό είναι και το μόνο μειονέκτημα που παρουσιάζει η συγκεκριμένη μέθοδος σε σχέση με τις προηγούμενες που χρησιμοποιήθηκαν για να λύσουν το πρόβλημα κατηγοριοποίησης.

Όσον αφορά την χρονική επίδοση των Μηχανών Διανυσμάτων Υποστήριξης, αυτή είναι πολύ καλή. Κάποια δευτερόλεπτα απαιτούνται για την ολοκλήρωση της εκπαίδευσης τους και συγκεκριμένα για τον Γκαουσιανό πυρήνα η εκπαίδευση διαρκεί μόλις 3 δευτερόλεπτα. Ακόμη και οι προβλέψεις για τις ετικέτες των δειγμάτων δοκιμών λαμβάνονται πολύ γρήγορα από την μηχανή, μόλις σε λίγα δευτερόλεπτα. Έτσι η χρονική επίδοση των μηχανών αυτών είναι με διαφορά καλύτερη από αυτή των κλασσικών μεθόδων Κ Πλησιέστερων Κέντρων και Πλησιέστερου Κέντρου.

Για να επιτύγχουν οι μηχανές αποτελέσματα με υψηλλή ακρίβεια απαιτείται η χρήση της σωστής συνάρτησης πυρήνα. Έτσι για την γραμμική και την σιγμοειδή συνάρτηση η ακρίβεια είναι χειρότερη από αυτή του Κατηγοριοποιητή Κ Πλησιέστερων γειτόνων και για τα δείγματα δοκιμής αλλά και για τα δείγματα εκπαίδευσης. Όταν γίνεται χρήση ομως τον σωστών συναρτήσεων πυρήνα (πολυωνυμική, γκαουσιανή) η ακρίβεια των αποτελεσμάτων είναι με διαφορά η καλύτερη σε σχέση με οποιαδήποτε άλλη μέθοδο. Συγκεκριμένα για τον πυρήνα RBF, ο οποίος είναι και αυτός με τα καλύτερα αποτελέσματα, παρατηρείται ποσοστό επιτυχίας στα δείγματα εκπαίδευσης ίσο με 99.96% και ίσο με 99.76% για τα δείγματα δοκιμής. Αυτό καταδεικνύει το πολύ καλό επίπεδο εκπαίδευσης αλλά και γενίκευσης που παρέχουν οι Μηχανές Διανυσμάτων Υποστήριξης με χρήση του Γκαουσιανού Πυρήνα, οι οποίες και υπερτερούν σε σχέση με όλους τους υπόλοιπους κατηγοριοποιητές που δοκιμάσθηκαν.