Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ

ΜΗ ΓΡΑΜΜΙΚΗ ΜΕΙΩΣΗ ΔΙΑΣΤΑΣΕΩΝ ΣΕ ΕΦΑΡΜΟΓΕΣ ΑΝΑΓΝΩΡΙΣΗΣ ΠΡΟΤΥΠΩΝ

Εκπόνηση:
Πέτρος Κατσιλέρος

Επίβλεψη:
Νικόλαος Πιτσιάνης
Νίκος Σισμάνης

Με την εκπόνηση της εν λόγω διπλωματικής εργασίας ολοκληρώνεται ο κύκλος των προπτυχιακών μου σπουδών αποκτώντας δίπλωμα Ηλεκτρολόγου Μηχανικού και Μηχανικού Η/Υ απο το Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης.

Περίληψη

Η τεχνητή νοημοσύνη μέσω της μηχανικής μάθησης είναι αναμφισβήτητα ένας επιστημονικός κλάδος ο οποίος επικεντρώνει το ενδιαφέρον ολοένα και περισσότερων μηχανικών-ερευνητών. Το γεγονός αυτό οφείλεται στην επιτυχία τέτοιου είδους εφαρμογών σε διάφορους κλάδους της καθημερινότητάς μας όπως αυτός της ρομποτικής, της υγείας, της εξόρυξης γνώσης κλπ. Επίσης οι σημερινοί υπολογιστές λόγω της ραγδαίας ανάπτυξης της τεχνολογίας παρέχουν τους απαραίτητους πόρους ώστε να μπορέσουν να αναπτυχθούν και να διερευνηθούν τέτοιου είδους προβλήματα. Παρ' όλα αυτά, όσους πόρους και αν διαθέσουμε δεν μπορούμε σε καμιά περίπτωση να δημιουργήσουμε κάτι αντίστοιχο με τον ανθρώπινο εγκέφαλο.

Γνωρίζουμε οτι ο ανθρώπινος εγκέφαλος είναι ένα τρομερά περίπλοκο σύστημα εκατομμυρίων νευρώνων συνδεδεμένων μεταξύ τους οι οποίοι είναι σε θέση να εκτελούνε σε κλάσματα του δευτερολέπτου έναν τεράστιο αριθμό λογικών πράξεων. Το μοντέλο αυτό είναι αδύνατον να προσομοιωθεί με οποιοδήποτε υπολογιστικό σύστημα διαθέτει ο άνθρωπος σήμερα. Στην προσπάθεια των Μηχανικών να μοντελοποιήσουν τις λειτουργίες του λαμβάνοντας φυσικά υπόψιν ευρήματα και αποτελέσματα των επιστημόνων της Ιατρικής σημαντικές λύσεις και βελτιστοποιήσεις έρχονται να δώσουν αλγόριθμοι οι οποίοι έχουν ως στόχο να μειώσουν τις παραμέτρους τις οποίες πρέπει να εκτιμήσει κάποιο υπολογιστικό σύστημα ώστε τελικά να μπορέσει να εξάγει συμπεράσματα, αντίστοιχα ενός ανθρώπου.

Χαρακτηριστικά παραδείγματα τέτοιων εφαρμογών με τα οποία καταπιάνεται και η εργασία αυτή είναι η μείωση των παραμέτρων σε μοτίβα εικόνων ή άλλων δεδομένων με στόχο την εξαγωγή συμπεράσματος για την ταξινόμηση των δεδομένων σε κλάσεις. Συγκεκριμένα γίνεται εφαρμογή του αλγορίθμου Locally Linear Embedding σε τρία σετ δεδομένων. Τα δύο περιέχουν εικόνες με ψηφία αριθμών και στόχος είναι να γίνει ταξινόμηση των ψηφίων αυτών, και το τελευταίο σετ περιέχει ιατρικά δεδομένα ασθενών με σκοπό την πρόβλεψη εμφάνισης ή όχι κάποιας μορφής καρκίνου. Αφού γίνει μείωση των διαστάσεων οι οποίες και λαμβάνονται τελικά υπόψιν για την εξαγωγή του τελικού συμπεράσματος, εφαρμόζεται ο ταξινομητής κοντινότερων γειτόνων (K-Nearest Neighbors (KNN)) ο οποίος εξάγει και το τελικό συμπέρασμα για την ταξινόμηση των δεδομένων στις κατάλληλες κλάσεις.

Ευχαριστίες

Με την ολοκλήρωση αυτής της διπλωματικής εργασίας θα ήθελα καταρχήν να ευχαριστήσω τον επίκουρο καθηγητή του τμήματός Ηλεκτρολόγων Μηχανικών και Μηχανικών Η/Υ του Αριστοτελείου Πανεπιστημίου Θεσσαλονίκης κ.Νικόλαο Πιτσιάνη ο οποίος μου έδωσε το ερέθισμα, απαραίτητες χρήσιμες συμβουλές αλλά και πόρους ώστε να μπορέσω να ολοκληρώσω την έρευνα για το συγκεκριμένο θέμα. Επίσης ένα μεγάλο ευχαριστώ στον υποψήφιο διδάκτορα του τμήματος Νίκο Σισμάνη για την καθοδήγηση του καθ΄όλη την διάρκεια εκπλήρωσης της εργασίας μου αυτής.

Τέλος, ένα πολύ θερμό και μεγάλο ευχαριστώ στους γονείς μου οι οποίοι με στήριξαν τόσο οικονομικά όσο και ψυχολογικά όλα αυτά τα χρόνια ώστε να μπορέσω να αποκτήσω το δίπλωμά μου. Στο σημείο αυτό δεν θα μπορούσα να παραλείψω τον σκύλο μου,την κοπέλα και τους φίλους μου διότι ο καθένας ξεχωριστά με τον τρόπο του βοήθησαν στην αντιμετώπιση των δυσκολιών που συνάντησα καθόλη την διάρκεια των σπουδών μου.

Κατσιλέρος Πέτρος

Θεσσαλονίκη, Μάρτιος 2016

Αφιέρωση

Αφιερώνω την διπλωματική αυτή εργασία πρωτίστως στον εαυτό μου για τον κόπο μου όλα αυτά τα χρόνια ώστε να μπορέσω να αποκτήσω το δίπλωμα Ηλεκτρολόγου Μηχανικού και Μηχανικού Η/Υ και κατά δεύτερον στους γονείς μου οι οποίοι με στήριξαν ανελλιπώς και με κάθε τρόπο σε όλη αυτή την πορεία.

Κατσιλέρος Πέτρος

Θεσσαλονίκη, Μάρτιος 2016

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κεφάλαιο 1

Εισαγωγή

1.1 Αναγνώριση προτύπων και μηχανική μάθηση

Αναγνώριση προτύπων καλείται η επιστημονική περιοχή που έχει στόχο την ταξινόμηση αντικειμένων σε κατηγορίες ή κλάσεις. Ανάλογα με την κάθε εφαρμογή τα δεδομένα μπορεί να είναι είτε εικόνες, είτε σήματα είτε οποιοδήποτε άλλο σετ δεδομένων χρειάζεται για κάποιο λόγο να ταξινομηθεί. Στις μέρες μας η ανάγκη διαχείρισης αλλά και ανάκτησης πληροφοριών μέσω ηλεκτρονικών υπολογιστών αποκτά τεράστια σπουδαιότητα. Αυτό διότι ο όγκος των πληροφοριών αυξάνεται ραγδαία με ρυθμό αδύνατο να διαχειριστεί ο άνθρωπος. Επίσης η ανάπτυξη της τεχνολογίας μας παρέχει πολύ ισχυρά υπολογιστικά συστήματα με τη χρήση των οποίων μπορούμε να δημιουργήσουμε πολύπλοκα μοντέλα εξόρυξης γνώσης.

Επιστημονικοί κλάδοι στους οποίους έχει τεράστια σημασία η αναγνώριση προτύπων είναι οι αυτός της Ιατρικής, της Βιολογίας, ο χώρος των αγορών και των επιχειρήσεων και τέλος η διαχείριση και η εξόρυξη γνώσης απο τον τεράστιο όγκο της πληροφορίας που είναι διαθέσιμος στο διαδίκτυο. Φυσικά η αναγνώριση προτύπων είναι ένα πολύ σημαντικό μέρος του κλάδου της μηχανικής μάθησης σε ρομποτικά/υπολογιστικά συστήματα.

Η υπολογιστική όραση για παράδειγμα είναι αντικείμενο ιδιαίτερα χρήσιμο τόσο στον χώρο της ρομποτικής όσο σε αυτόν της ιατρικής αλλά και προφανώς της βιομηχανίας. Τέτοιου είδους εφαρμογές έχουν εισέλθει πολύ δυναμικά στην καθημερινότητά μας τα τελευταία χρόνια. Συγκεκριμένα στον χώρο της βιομηχανίας υπάρχουν συστήματα τα οποία επιβλέπουν μέσω μια κάμερας την γραμμή παραγωγής καθώς και ρομπότ τα οποία μεταφέρουν και συναρμολογούν αντικείμενα. Επίσης υπάρχουν εφαρμογές οι οποίες αναγνωρίζουν για παράδειγμα πρόσωπα τραβώντας μια εικόνα με το κινητό μας τηλέφωνο. Τέλος στον χώρο της αυτοκινητοβιομηχανίας δεν είναι λίγες αντίστοιχες εφαρμογές οι οποίες έχουν συμβάλει δυναμικά στην αυτόνομη οδήγηση αλλά και στην προειδοποίηση για εμπόδια κλπ.

Ιδιαίτερη έμφαση αξίζει να δωθεί στην εξόρυξη γνώσης σε κλάδους όπως στη Βιολογία αλλά και στην Ιατρική. Για παράδειγμα η πρόβλεψη εμφάνισης ασθενειών όπως ο καρκίνος μέσω αναγνώρισης συγκεκριμένων μοτίβων σε εικόνες απο μαγνητικό τομογράφο, η μελέτη της αλύσίδας του γεννετικού υλικού αλλά και ο χώρος των εγχειρίσεων υψηλής ακρίβειας με τη χρήση της ρομποτικής.

1.2 Ερεθίσματα απο τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου

Απο μελέτες που έχουν γίνει για την λειτουργία του ανθρώπινου εγκεφάλου γνωρίζουμε ότι για οποιοδήποτε σύνολο μετρήσεων προέρχεται για παράδειγμα είτε απο την όραση μας είτε απο την ακοή μας ο εγκέφαλός μας μετασχηματίζει το σύνολο των δεδομένων αυτών σε ένα νέο σύνολο χαρακτηριστικών. Με τον τρόπο αυτό, επιλέγοντας προφανώς κάθε φορά τα κατάλληλα χαρακτηριστικά, επιτυγχάνεται τεράστια συμπίεση του όγκου της πληροφορίας σε σύγκριση με τα αρχικά δεδομένα εισόδου. Αυτο σημαίνει λοιπόν ότι το μεγαλύτερο μέρος της πληροφορίας για παράδειγμα μια σκηνής που βλέπουμε και στην οποία θέλουμε να αναγνωρίσουμε τα αντικείμενα που περιέχονται, συμπιέζεται σε έναν πολύ μικρό αριθμό χαρακτηριστικών. Η παραπάνω διαδικασία χαρακτηρίζεται ως τεχνική μείωσης διάστασης γνωστή στην βιβλιογραφία με τον όρο Dimensionality Reduction.

Ας πάρουμε για παράγειγμα τον κλάδο της υπολογιστικής όρασης ο οποίος αποτελεί και αντικείμενο μελέτης της εν λόγω εργασίας και ας αναρωτηθούμε το εξής: Πόσο δύσκολο είναι για κάποιον απο εμάς να αγνωρίσει κάποιο νούμερο αποτυπωμένο σε μια εικόνα· Η προφανής απάντηση είναι καθόλου. Και αυτή είναι μια πολύ σωστή απάντηση, διότι για τον ανθρώπινο εγκέφαλο το να καταλάβει οτι το ψηφίο το οποίο βρίσκεται στην εικόνα είναι για παράδειγμα το 1 και όχι το 9 είναι ένα πολύ απλό πρόβλημα.

Πιο συγκεκριμένα βλέποντας μια οποιαδήποτε σκηνή ο ανθρώπινος εγκέφαλος προσπαθεί να εντοπίσει σημεία ενδιαφέροντος τα οποία αποτελούν χαρακτηριστικά σημεία της. Τέτοια μπορεί να είναι πολύ έντονες αλλαγές στην φωτεινότητα όπως για παράδειγμα γωνίες, κενά ή τρύπες. Στην συνέχεια εντοπίζει πιο σύνθετες γεωμετρίες όπως ευθείες ή καμπύλες γραμμές και τέλος προσδιορίζει πιο ολοκληρωμένες δομές τρισδιάστατων αντικειμένων. Το ίδιο ακριβώς γίνεται και στην παραπάνω περίπτωση με το ψηφίο. Εντοπίζουμε αρχικά οτι το μοτίβο του ψηφίου 1 είναι πολύ κοντά σε αυτά των ψηφίων εφτά και τέσσερα αλλά σε καμιά περίπτωση δεν θα λέγαμε οτι έχει τρομερές ομοιότητες με αυτό του δύο ή του οχτώ για παράδειγμα.

Το παραπάνω παράδειγμα είναι ένα πολύ απλό δείγμα του τρόπου με τον οποίο ο ανθρώπινος εγκέφαλος προσπαθεί με κάθε τρόπο να ελαχιστοποιήσει τις παραμέτρους που πρέπει να εκτιμήσει. Φυσικά αν αναλογιστούμε ένα ρεαλιστικό περίπλοκο πρόβλημα της καθημερινότητάς μας θα δούμε οτι απαιτούνται πολύ πιο σύνθετοι υπολογισμοί και θα πρέπει να συνδιάσουμε ένα πλήθος απο παραμέτρους ώστε τελικά να καταλήξουμε στο τελικό συμπέρασμα για κάποια απόφαση. Σε κάθε περίπτωση όμως γίνεται τεράστια συμπίεση της αρχικής πληροφορίας μέσω τεχνικών μείωσης διαστάσεων ώστε να ελαχιστοποιηθούν οι παράμετροι που πρέπει να υπολογιστούν και προφανώς να επιταχυνθεί η διαδικασία εξαγωγής της τελικής μας απόφασης.

Το γεγονός αυτό και δεδομένου οτι το όραμα της επιστημονικής κοινότητας των μηχανικών που ασχολούνται με την μηχανική μάθηση και την εξόρυξη γνώσης είναι να δημιουργηθεί ένα μοντέλο αντίστοιχο με αυτό του ανθρώπινου εγκεφάλου δεν θα μπορούσε να τους αφήσει αδιάφορους ώστε να μελετήσουν και να αναπτύξουν αντίστοιχους αλγορίθμους.

1.2.1 Μάθηση με επίβλεψη - χωρίς επίβλεψη - με ημιεπίβλεψη

Ένα πολύ εύλογο ερώτημα το οποίο προκύπτει απο την παραπάνω ανάλυση είναι πως ο ανθρώπινος εγκέφαλος έχει μάθει και τελικώς έχει αποθηκεύσει το σύνολο αυτών των μοντέλων για τον κάθε αριθμό ή για οποιοδήποτε άλλο αντικείμενο ή μοτίβο μπορεί να αναγνωρίσει με τόσο μεγάλη ταχύτητα και ευκολία. Η απάντηση είναι προφανώς η συνεχής εκπαίδευση και η διαρκής υπενθύμιση των συγκεκριμένων προτύπων.

Πιο συγκεκριμένα ο άνθρωπος απο την μέρα που αρχίζει να αλληλεπιδρά με το περιβάλλον παίρνει διάφορα ερεθίσματα τα οποία καιρό με τον καιρό μαθαίνει να τα ταξινομεί κατάλληλα και να τα χρησιμοποιεί σε περίπτωση που εμφανιστούν μπροστά του. Τα ερεθίσματα αυτά είναι είτε εικόνες, είτε ήχοι είτε ερεθίσματα τα οποία μπορεί να προέρχονται απο τις υπόλοιπες αισθήσεις του.

Ο τρόπος με τον οποίο καταφέρνουμε να συγκρατούμε και να μπορούμε να διαχειριστούμε ανα πάσα στιγμή τον τεράστιο όγκο πληροφοριών που βρίσκονται καταχωρημένες στον εγκέφαλό μας είναι ένας συνδιασμός τεχνικών μάθησης και συνεχούς εκπαίδευσης. Οι τεχνικές αυτές στον χώρο της τεχνητής νοημοσύνης αναφέρονται ως τεχνικές μάθησης με επίβλεψη, χωρίς επίβλεψη και με ημιεπίβλεψη. Θα μπορούσε κάποιος αρχικά να υποστηρίξει ότι ο ανθρώπινος εκγέφαλος χρησιμοποιεί κατεξοχήν τεχνικές μάθησης χωρίς επίβλεψη διότι μπορεί να μαθαίνει μόνος του νέα πράγματα. Είναι όμως πραγματικά αυτό το οποίο συμβαίνει· Η απάντηση είναι μάλλον όχι, και αυτό διότι απο την πολύ νεαρή του υλικία ο καθένας μας έχει γύρω του ανθρώπους οι οποίοι προσπαθούν συνεχώς να μας μεταφέρουν γνώση και να μας μάθουν τι βρίσκεται γύρω μας και πως να αλληλεπιδρούμε μαζί του. Παρ΄όλα αυτά μετά απο κάποιο σημείο ο ανθρώπινος εγκέφαλος αποκτά δυνατότητες με τις οποίες μπορεί να αξιολογεί και να μαθαίνει μόνος του πολύ σύνθετα προβλήματα. Αυτό το επιτυγχάνει αναλύοντάς τα σε απλούστερα τα οποία γνωρίζει ήδη πως να τα διαχειριστεί. Επίσης είναι στην φύση του ανθρώπου να εξερευνεί συνεχώς άγνωστα μονοπάτια και να αναζητεί απαντήσεις σε άγνωστα προβλήματα επιτυγχάνοντας αξιοθαύμαστα αποτελέσματα.

Απο τα παραπάνω καταλήγουμε στο συμπέρασμα ότι ο άνθρωπος χρησιμοποιεί τεχνικές ημιεπίβλεψης για την εκπαίδευση του εγκεφάλου του γεγονός το οποίο του δίνει την δυνατότητα να μπορεί να διαχειριστεί αλλά και να μάθει πολύ σύνθετα μοντέλα. Μέσα απο αυτή την διαδικασία είναι σε θέση με το πέρασμα του χρόνου να δημιουργήσει ένα τεράστιο και πανίσχυρο δίκτυο πληροφοριών, ταξινομημένο με τρόπο τον οποίο δεν μπορούμε ακόμα να εξηγήσουμε και να κατανοήσουμε. Με αυτό το μοντέλο είναι σε θέση ταχύτατα να αποφασίζει που βρίσκεται ο ευρύτερος χώρος της πληροφορίας που θέλει να αντλήσει και στην συνέχεια να αποφασίζει με τεράστια ακρίβεια και ταχύτητα την τελική του απόφαση.

Το μοντέλο αυτό με το οποίο λειτουργεί ο ανθρώπινος εγκέφαλος είναι αν μη τι άλλο αξιοθαύμαστο και ανεξήγητο. Παρ΄όλα είναι πολύ δύσκολο να εφαρμοστεί στον τομέα της τεχνητής νοημοσύνης και αυτό διότι ακόμα δεν είμαστε σε θέση να δώσουμε εξηγήσεις για τον ακριβή τρόπο λειτουργίας του. Το συνηθέστερο και πιο αποτελεσματικό μέχρι στιγμής μοντέλο το οποίο χρησιμοποιείται στην εξόρυξη γνώσης μέσω ηλεκτρονικών υπολογιστών είναι αυτό της μάθησης με επίβλεψη. Σύμφωνα με το μοντέλο αυτό θα πρέπει αν συλλέξουμε ένα μεγάλο συνήθως όγκο δεδομένων τον οποίο να τροφοδοτήσουμε στην συνέχεια ως είσοδο στο σύστημά μας και με την κατάλληλη μεθοδολογία να το καθοδηγήσουμε ώστε τελικά να μάθει συγκεκριμένα μοντέλα τα οποία να μπορεί να χρησιμοποιήσει στην συνέχεια με σκοπό την εξαγωγή κάποιου συμπεράσματος.

1.3 Μείωση της διάστασης των δεδομένων

Στην παραπάνω διαδικασία δεδομένου ότι στις περισσότερες περιπτώσεις έχουμε να αντιμετωπίσουμε πολύ σύνθετα υπολογιστικά προβλήματα ο αριθμός των παραμέτρων που πρέπει να υπολογιστούν είναι σε συγκεκριμένες εφαρμογές απαγορευτικά μεγάλος. Σε κάποιες εφαρμογές το πρόβλημα είναι θέμα χρόνου όπου πρέπει να γίνει μείωση των παραμέτρων ώστε να ελαχιστοποιηθεί ο χρόνος εξαγωγής του συμπεράσματος. Σε άλλες είναι θέμα χώρου διότι ένας μεγάλος αριθμός πολυδιάστατων δεδομένων μπορεί να αποτελεί πρόβλημα σε συγκεκριμένες εφαρμογές. Τέλος υπάρχουν περιπτώσεις στις οποίες χρειαζόμαστε την μείωση των διαστάσεων ώστε να διώξουμε εντελώς παραμέτρους οι οποίες επιδρούν σαν θόρυβος και επηρεάζουν αρνητικά την εξαγωγή ορθού συμπεράσματος ταξινόμησης. Προφανώς σε πολλές πρακτικές εφαρμογές επικρατεί ένας συνδυασμός των παραπάνω αναγκών.

Αντικείμενο λοιπόν της εν λόγω διπλωματικής εργασίας είναι η διερεύνηση και η χρήση του αλγορίθμου Locally Linear Embeddings για την μείωση των διαστάσεων σε πρακτικά προβλήματα όπως η αναγνώριση ψηφίων αλλά και η ταξινόμηση ασθενών με βάση το αν πρόκειται να εμφανίσουν κάποιας μορφής καρκίνου ή όχι. Τα αποτελέσματα των πειραμάτων είναι ιδιαίτερα ενθαρυντικά και δείχνουν σε όλες τις περιπτώσεις ότι η μείωση των διαστάσεων επιδρά δραματικά στην μείωση του κόστους των υπολογισμών αλλά και στην αύξηση της σωστής πρόβλεψης λόγω απομάκρυνσης του θορύβου. Επίσης παρουσιάζονται δύο πρακτικές και ρεαλιστικές μέθοδοι εφαρμογής του αλγορίθμου σε πραγματικά προβλήματα απο τις οποίες η πρώτη έρχεται να αντιμετωπίσει το πρόβλημα της πολύ μεγάλης μνήμης που απαιτεί η εκτέλεση του αλγορίθμου και η δεύτερη παρέχει την δυνατότητα για την ταξινόμηση των αποτελεσμάτων και την εξαγωγή συμπεράσματος σε πραγματικό χρόνο.

Κεφάλαιο 2

Μαθηματικό και θεωρητικό υπόβαθρο

2.1 Διανύσματα βάσης

Έστω ότι έχουμε ένα σύνολο δειγμάτων εισόδου με αντίστοιχο διάνυσμα $\mathbf{x}$ διάστασης $N \times 1$,

$$\mathbf{x}^T = \Big[x(0), \ldots, x(N-1) \Big]$$

Έστω επίσης ορθοκανονικό μητρώο A, τάξης $N \times N$. Τότε ορίζουμε το μετασχηματισμένο διάνυσμα $\mathbf{y}$ του $\mathbf{x}$ ώς

$$\mathbf{y} = \mathbf{A}^H \mathbf{x} \equiv \begin{bmatrix} \mathbf{a}_0^H \\ \vdots \\ \mathbf{a}_{N-1}^H \end{bmatrix} \mathbf{x} \tag{2.1.1}$$

Το H δηλώνει τον Hermitian τελεστή, δηλαδή τον μιγαδικό συζηγή του ανάστροφου. Απο τον ορισμό των ορθοκανονικών μητρώων έχουμε

$$\mathbf{x} = \mathbf{A}\mathbf{y} = \sum_{i=0}^{N-1} y(i) \mathbf{a}_i \tag{2.1.2}$$

Οι στήλες του $\mathbf{A}$, $\mathbf{a}_i = 0, 1, \ldots, N-1$ καλούνται *διανύσματα βάσης* του μετασχηματισμού. Τα στοιχεία $y(i)$ του $\mathbf{y}$ είναι οι προβολές του διανύσματος $\mathbf{x}$ σε αυτά τα διανύσματα βάσης. Λαμβάνοντας υπόψιν την ιδιότητα της ορθοκανονικότητας μπορούμε να επαληθεύσουμε την παραπάνω διατύπωση υπολογίζοντας το εσωτερικό γινόμενο του $\mathbf{x}$ με το $\mathbf{a}_j$. Έχουμε:

$$\langle \mathbf{a}_j, \mathbf{x} \rangle \equiv \mathbf{a}_j^H \mathbf{x} = \sum_{i=0}^{N-1} y(i) \langle \mathbf{a}_j, \mathbf{a}_i \rangle = \sum_{i=0}^{N-1} y(i) \delta_{ij} = y(j) \tag{2.1.3}$$

2.1.1 Διάνυσμα εικόνας

Άν πάρουμε για παράδειγμα μια εικόνα, το σύνολο των δειγμάτων εισόδου είναι μια δυδιάστατη ακολουθία $X(i,j), i, j = 0, 1, \ldots, N-1$, η οποία ορίζει ένα μητρώο Q, τάξεως $N \times N$. Σε αυτή την περίπτωση μπορούμε να μετατρέψουμε την είσοδο αυτή σε ένα διάνυσμα $\mathbf{x}$ διάστασης N^2 διατάσσοντας για παράδειγμα τις γραμμές του μητρώου την μία μετά την άλλη έχοντας τελικά

$$\mathbf{x}^T = \Big[X(0,0), \ldots, X(0, N-1), \ldots, X(N-1, 0), \ldots, X(N-1, N-1) \Big] \tag{2.1.4}$$

Με αυτό τον μετασχηματισμό όμως ο αριθμός των πράξεων που απαιτούνται για τον πολλαπλασιασμό ενός τετραγωνικού μητρώου τάξεως $N \times N$ με ένα διάνυσμα $\mathbf{x}$ διαστάσεων $N^2 \times 1$, είναι της τάξης $\mathcal{O}(N^4)$ μέγεθος απαγορευτικό για τις περισσότερες ρεαλιστικές εφαρμογές.

2.1.2 Ορθοκανονικά ιδιοδιανύσματα

Το παραπάνω εμπόδιο μπορεί να ξεπεραστεί αν μετασχηματίσουμε το μητρώο Q μέσω ενός συνόλου *μητρώων βάσης*. Έστω λοιπόν U και V ορθοκανονικά μητρώα διάστασης $N \times N$. Ορίζουμε τότε το μετασχηματισμένο μητρώο Y του X ως

$$Y = U^H X V \tag{2.1.5}$$

ή

$$X = U Y V^H \tag{2.1.6}$$

Μέσω αυτού του μετασχηματισμού ο αριθμός των πράξεων μειώνεται σε $\mathcal{O}(N^3)$. Πιο αναλυτικά η παραπάνω εξίσωση θα μπορούσε να γραφεί ως

$$Q = \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} Y(i,j) \mathbf{u}_i \mathbf{v}_j^H \tag{2.1.7}$$

όπου $\mathbf{u}_i$ είναι τα διανύσματα στήλης του U και $\mathbf{v}_j$ τα διανύσματα στήλης του V. Η παραπάνω εξίσωση είναι ένα ανάπτυγμα του μητρώου X ως προς τις $N \times 2$ εικόνες βάσης. Τέλος κάθε ένα απο τα γινόμενα $\mathbf{u}_i \mathbf{v}_j$ είναι ένα μητρώο $N \times N$

$$\mathbf{u}_i \mathbf{v}_j = \begin{bmatrix} u_{i0} v_{j0}^* & \dots & u_{i0} v_{jN-1}^* \\ \vdots & \vdots & \vdots \\ u_{iN-1} v_{j0}^* & \dots & u_{iN-1} v_{jN-1}^* \end{bmatrix} \tag{2.1.8}$$

Στην περίπτωση κατα την οποία το Y είναι διαγώνιο τότε έχουμε

$$Q = \sum_{i=0}^{N-1} Y(i,i)\mathbf{u}_i\mathbf{v}_i^H \tag{2.1.9}$$

με αποτέλεσμα το πλήθος των μητρώων-εικόνων βάσης να μειώνεται σε N. Τέλος έπειτα απο μερικές πράξεις και τροποποιήσεις μπορούμε να ορίσουμε κάθε στοιχείο (i, j) του μετασχηματισμένου μητρώου ως τον πολλαπλασιασμό κάθε στοιχείου του X με τον συζυγή του αντίστοιχου στοιχείου του A_{ij} και αθροίζοντας όλα τα γινόμενα. Δηλαδή

$$\langle A, B\rangle = \sum_{m=0}^{N-1}\sum_{n=0}^{N-1} A(m,n)^* B(m,n) \tag{2.1.10}$$

και τελικά

$$Y(i,j) = \langle A_{i,j}, X\rangle \tag{2.1.11}$$

2.2 Ο μετασχηματισμός Karhunen-Loeve - PCA

Ο μετασχηματισμός Karhunen-Loeve αξιοποιεί την στατιστική πληροφορία που περιγράφει τα δεδομένα και ο υπολογισμός του μητρώου γίνεται χωρίς επίβλεψη. Ας υποθέσουμε και πάλι ένα διάνυσμα $\mathbf{x}$ το οποίο αποτελείται απο τα δείγματα μια εικόνας τα οποία έχουν διαταχθεί λεξικογραφικά όπως περιγράφτηκε παραπάνω. Πρέπει να επισυμανθεί στο σημείο αυτό η επιθυμητή ιδιότητα των εξαχθέντων χαρακτηριστικών να είναι αμοιβαίως ασυσχέτιστα και αυτό για την αποφυγή πλεονάζουσας πληροφορίας. Η πιο συνηθισμένη συνθήκη για την γέννηση τέτοιου είδους χαρακτηριστικών είναι η μέση τιμή των δεδομένων να έχει μηδενική τιμή. Δηλαδή θέλουμε την ιδιότητα

$$E[y(i)y(j)] = 0, i \neq j \tag{2.2.1}$$

Έστω

$$\mathbf{y} = A^T \mathbf{x} \tag{2.2.2}$$

Εφόσον έχουμε υποθέσει ότι $E[x] = 0$ αμέσως βλέπουμε ότι $E[y] = 0$ και

$$R_y = E[\mathbf{y}\mathbf{y}^T] = E[A^T \mathbf{x}\mathbf{x}^T A] = A^T R_x A \tag{2.2.3}$$

Πρακτικά το R_x αντιπροσωπεύει μια μέση τιμή πάνω στο δοθέν σύνολο διανυσμάτων εκπαίδευσης. Επίσης είναι συμμετρικό μητρώο και επομένως τα ιδιοδιανύσματά του είναι αμοιβαίως ορθογώνια. Άρα έστω ότι επιλέγεται ένα μητρώο A με στήλες τα ορθοκανονικά ιδιοδιανύσματα $\mathbf{a}_i, i = 0, 1, \ldots, N-1$ του R_x τότε το R_y είναι διαγώνιο.

$$R_y = A^T R_x A = \Lambda \tag{2.2.4}$$

Το Λ είναι διαγώνιο μητρώο με διαγώνια στοιχεία τις αντίστοιχες ιδιοτιμές $\lambda_i, i = 0, 1, \ldots, N-1$ του R_x. Αποτέλεσμα της παραπάνω διαδικασίας είναι ένας μετασχηματισμός, ο μετασχηματισμός Karhunen-Loeve ο οποίος επιτυγχάνει τον αρχικό μας στόχο, δηλαδή την δημιουργία χαρακτηριστικών τα οποία είναι στατιστικώς ανεξάρτητα.

2.2.1 Προσέγγιση μέσου τετραγωνικού σφάλματος - MSE

Σε αυτή την υποενότητα θα αναλυθεί η διαδικασία με την οποία μπορούμε να οδηγηθούμε στην επιλογή κάποιων, έστω m το πλήθος, κυρίαρχων χαρακτηριστικών μέσω της προσέγγισης μέσου τετραγνικού σφάλματος. Ας πάρουμε ξανά τις εξισώσεις (2.1.1) και (2.1.2) τότε έχουμε

$$\mathbf{x} = \sum_{i=0}^{N-1} y(i)\mathbf{a}_i \quad \text{και} \quad y(i) = \mathbf{a}_i^T \mathbf{x} \tag{2.2.5}$$

Ορίζουμε λοιπόν τώρα ένα νέο διάνυσμα στον m-διάστατο υποχώρο

$$\widehat{\mathbf{x}} = \sum_{i=0}^{N-1} y(i)\mathbf{a}_i \tag{2.2.6}$$

στο οποίο προφανώς εμπλέκονται μόνο m απο τα διανύσματα βάσης. Με τον παραπάνω τρόπο δηλαδή ορίζεται η προβολή του $\mathbf{x}$ στον υποχώρο που ορίζουν τα ορθοκανονικά διανύσματα m τα οποία εμπλέκονται στην παραπάνω άθροιση.

Σκοπός μας λοιπόν στο σημείο αυτό είναι να προσεγγίσουμε με όσο το δυνατόν μικρότερο σφάλμα το διάνυσμα $\mathbf{x}$. Η προσέγγισή μας είναι το διάνυσμα $\widehat{\mathbf{x}}$ και θα προκύψει χρησιμοποιώντας την εξίσωση ελαχιστοποίησης μέσου τετραγωνικού σφάλματος. Έχουμε λοιπόν την εξίσωση

$$E[\|\mathbf{x} - \widehat{\mathbf{x}}\|^2] = E\left[\| \textstyle\sum_{i=m}^{N-1} y(i)\mathbf{a}_i\|^2\right] \tag{2.2.7}$$

Απο την παραπάνω εξίσωση στόχος μας τώρα είναι να επιλέξουμε τα ιδιοδιανύσματα τα οποία οδηγούν στο ελάχιστο μέσο τετραγωνικό σφάλμα. Λαμβάνοντας υπόψιν την ορθοκανονικότητα των ιδιοδιανυσμάτων και την παραπάνω εξίσωση καταλήγουμε ότι

$$E\left[\| \textstyle\sum_{i=m}^{N-1} y(i)\mathbf{a}_i\|^2\right] = E\left[\textstyle\sum_i \sum_j (y(i)\mathbf{a}_i^T)(y(j)\mathbf{a}_j)\right] = \tag{2.2.8}$$

$$= \sum_{i=m}^{N-1} E[y^2(i)] = \sum_{i=m}^{N-1} \mathbf{a}_i^T E[\mathbf{x}\mathbf{x}^T]\mathbf{a}_i \tag{2.2.9}$$

και λαμβάνοντας υπόψιν τον ορισμό των ιδιοδιανυσμάτων προκύπτει τελικά ότι

$$E[\|\mathbf{x}-\widehat{\mathbf{x}}\|^2] = \sum_{i=m}^{N-1} \mathbf{a}_i^T \lambda_i \mathbf{a}_i = \sum_{i=m}^{N-1} \lambda_i \tag{2.2.10}$$

Αν επομένως στην παραπάνω εξίσωση επιλέξουμε τα ιδιοδιανύσματα που αντιστοιχούν στις m ιδιοτιμές του μητρώου συσχέτισης τότε το σφάλμα της εξίσωσης ελαχιστοποιείται και μάλιστα ισούτε με το άθροισμα των N-m μικρότερων ιδιοτιμών. Επιπλέον έχει αποδειχθεί ότι αυτό είναι το ελάχιστο μέσο τετραγωνικό σφάλμα σε σύγκριση με οποιαδήποτε άλλη προσέγγιση του $\mathbf{x}$ απο ένα m-διάστατο διάνυσμα. Για τον λόγο αυτό ο μετασχηματισμός Karhunen-Loeve είναι επίσης γνωστός ως *Ανάλυση κυρίων συνιστωσών* (Principal component analysis-PCA).

2.2.2 Συνολική Διασπορά

Έστω $\mathbf{y}$ το μετασχηματισμένο κατα **KL** διάνυσμα του $\mathbf{x}$ και $E[x] = 0$. Τότε απο τον αντίστοιχο ορισμό της διασποράς έχουμε ότι $\sigma^2_{y(i)} \equiv E[y^2(i)] = \lambda_i$. Δηλαδή έχουμε ότι οι διασπορές του μητρώου συσχέτισης εισόδου είναι ίσες με τις διασπορές των μετασχηματισμένων χαρακτηριστικών. Επομένως επιλεγοντας εκείνα τα χαρακτηριστικά $y(i) = \mathbf{a}_i^T\mathbf{x}$ που αντιστοιχούν στις m μεγαλύτερες ιδιοτιμές οδηγούμαστε σε μεγιστοποίηση της αθροιστικής διασποράς $\sum_i \lambda_i$. Συμπεραίνουμε λοιπόν ότι με αυτή την μεθολογία που ακολουθήσαμε, τα m χαρακτηριστικά που έχουν επιλεχθεί διατηρούν το μεγαλύτερο μέρος απο την συνολική διασπορά που σχετίζεται με τις αρχικές τυχαίες ματαβλητές $x(i)$.

2.2.3 Μείωση της διάστασης μέσω PCA

Απο την παραπάνω ανάλυση είναι φανερό ότι η μέθοδος PCA επιτυγχάνει τον *γραμμικό μετασχηματισμό* ενός χώρου υψηλής διάστασης σε έναν χαμηλής διάστασης του οποίου μάλιστα τα στοιχεία είναι στατιστικώς ασυσχέτιστα. Έχοντας υποθέσει ότι $E[x] = 0$ και επίσης ότι οι N-m μικρότερες ιδιοτιμές του μητρώου συσχέτισης είναι μηδέν τότε απο την εξίσωση (2.2.10) συνεπάγεται ότι $\mathbf{x} = \widehat{\mathbf{x}}$. Δηλαδή έχουμε ότι το διάνυσμα $\mathbf{x}$ του αρχικού χώρου διάστασης N βρίσκεται σε έναν m-διάστατο υποχώρο του αρχικού και μάλιστα μπορούμε να το προσδιορίσουμε μέσω του διανύσματος $\widehat{\mathbf{x}}$ με πολύ καλή προσέγγιση. Το γεγονός αυτό εισάγει την έννοια της *εγγενούς διάστασης* (intrinsic dimensionality). Τέλος στην περίπτωση της εγγενούς διάστασης μπορούμε να πούμε ό,τι το X μπορεί να περιγραφεί απο m ελεύθερες παραμέτρους.

2.3 Μετρική πολυδιάσττης κλιμάκωσης (Metric multidimensional scaling - MDS)

Ένας ακόμα πολύ διαδεδομένος αλγόριθμος μείωσης διάστασης είναι ο αλγόριθμος *Μετρική πολυδιάστατης κλιμάκωσης* (Metric multidimensional scaling - MDS). Ο αλγόριθμος αυτός δοθέντος ενός συνόλου $Q \subset \Re^N$, έχει ως στόχο να γίνει προβολή σε χώρο χαμηλότερης διάστασης, $Y \subset \Re^m$, έτσι ώστε τα εσωτερικά γινόμενα να διατηρηθούν κατά βέλτιστο τρόπο. Πρέπει δηλαδή να γίνει η ελαχιστοποίηση της εξίσωσης

$$E = \sum_i \sum_j (\mathbf{x}_i^T \mathbf{x}_j - \mathbf{y}_i^T \mathbf{y}_j)^2 \tag{2.3.1}$$

όπου $\mathbf{y}_i$ είναι η εικόνα του $\mathbf{x}_i$ και το άθροισμα υπολογίζεται ως προς όλα τα σημεία εκπαίδευσης του X. Το πρόβλημα δηλαδή, και σε αυτή την περίπτωση είναι όμοιο με αυτό της μεθόδου PCA, και μπορεί να αποδειχθεί ότι η λύση δίνεται απο την ανάλυση σε ιδιοτιμές-ιδιοδιανύσματα του μητρώου Gram, τα στοιχεία του οποίου ορίζονται ως

$$K(i, j) = \mathbf{x}_i^T \mathbf{x}_j \tag{2.3.2}$$

Ένας εναλλακτικός τρόπος επίλυσης του προβλήματος είναι η απαίτηση να διατηρηθούν, κατά βέλτιστο τρόπο, οι Ευκλείδιες αποστάσεις αντί των εσωτερικών γινομένων. Μπορούμε έτσι, να δημιουργήσουμε ένα μητρώο Gram συμβατό με τις τετραγωνικές Ευκλείδιες αποστάσεις, το οποίο μας οδηγεί στην ίδια λύση όπως και στην προηγούμενη περίπτωση. Προκύπτει μάλιστα, ότι οι λύσεις που προκύπτουν απο τις μεθόδους PCA και MDS είναι ισοδύναμες.

Μια σύντομη απόδειξη της παραπάνω διατύπωσης είναι η εξής. Η μέθοδος PCA εκτελεί την ανάλυση ιδιοτιμών του μητρώου συσχέτισης R_x, το οποίο προσεγγίζεται απο τη σχέση

$$R_x = E[\mathbf{x}\mathbf{x}^T] \approx \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k\mathbf{x}_k^T = \frac{1}{n}X^TX \tag{2.3.3}$$

όπου

$$X^T = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n,] \tag{2.3.4}$$

Απο την άλλη τώραμ το μητρώο Gram μπορεί επίσης να γραφεί ως

$$K = XX^T \tag{2.3.5}$$

Τέλος αποδεικνύεται ότι τα δύο μητρώα X^TX καί XX^T είναι ίδου βαθμού και έχουν τις ίδιες ιδιοτιμές με ιδιοδιανύσματα τα οποία ναι μεν είναι διαφορετικά μεταξύ τους αλλά παρόλα αυτά σχετίζονται.

2.4 Ανάλυση στην βάση των ιδιαζουσών τιμών (SVD)

Η ανάλυση ενός μητρώου με βάση τις ιδιάζουσες τιμές είναι μια απο τις πιο κομψές και ισχυρές μεθόδους γραμμικής άλγεβρας η οποία έχει χρησιμοποιηθεί εκτενώς για την μείωση του βαθμού και της διάστασης σε προβλήματα αναγνώρισης προτύπων και σε εφαρμογές ανάκτησης πληροφορίας.

Δοθέντως ενός μητρώου X, τάξης $l \times n$, βαθμού r με $r \leq$ min{l,n} υπάρχουν ορθοκανονικά μητρώα U και V, τάξης $l \times l$ και $n \times n$ αντίστοιχα ώστε

$$X = U \begin{bmatrix} \Lambda^{\frac{1}{2}} & \mathcal{O} \\ \mathcal{O} & 0 \end{bmatrix} V^H \quad ή \quad Y = \begin{bmatrix} \Lambda^{\frac{1}{2}} & \mathcal{O} \\ \mathcal{O} & 0 \end{bmatrix} = U^H X V \tag{2.4.1}$$

όπου $\Lambda^{\frac{1}{2}}$ είναι το $r \times r$ διαγώνιο μητρώο με στοιχεία $\sqrt{\lambda_i}$ με λ_i οι μη μηδενικές ιδιοτιμές που σχετίζονται με το μητρώο $X^H X$. Με $\mathcal{O}$ συμβολίζουμε το μητρώο μηδενικών τιμών. Απο τα παραπάνω γίνεται φανερό ότι υπάρχουν μητρώα U και V που μετασχηματίζουν το X στην διαγώνια δομή του Y. Αν $\mathbf{u}_i$,$\mathbf{v}_i$ είναι τα διανύσματα στήλης των μητρώων U και V αντίστοιχα τότε η παραπάνω εξίσωση μπορεί να γραφεί στην μορφή

$$X = [u_0, u_1, \ldots, u_{r-1},] \begin{bmatrix} \sqrt{\lambda_0} & & & \\ & \sqrt{\lambda_1} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_{r-1}} \end{bmatrix} \begin{bmatrix} \mathbf{v}_0^H \\ \mathbf{v}_1^H \\ \vdots \\ \mathbf{v}_{r-1}^H \end{bmatrix} \tag{2.4.2}$$

ή

$$X = \sum_{i=0}^{r-1} \sqrt{\lambda_i} \mathbf{u}_i \mathbf{v}_i^H = U_r \Lambda^{\frac{1}{2}} V_r^H \tag{2.4.3}$$

όπου U_r δηλώνει το $l \times r$ μητρώο που αποτελείται απο τις r πρώτες στήλες του U και V_r το $r \times n$ μητρώο που σχηματίζεται χρησιμοποιώντας τις πρώτες r στήλες του V. Επίσης $\mathbf{u}_i$,$\mathbf{v}_i$ είναι τα ιδιοδιανύσματα που αντιστοιχούν στις μη μηδενικές ιδιοτιμές των μητρώων XX^H και $X^H X$ αντίστοιχα. Οι ιδιοτιμές λ_i είναι γνωστές ως *ιδιάζουσες τιμές* (singular values) του X και το ανάπτυγμα της παραπάνω εξίσωσης ως *ανάλυση με βάση τις ιδιάζουσες τιμές* (singular value decomposition - SVD) του X.

2.4.1 Μείωση της διάστασης μέσω SVD

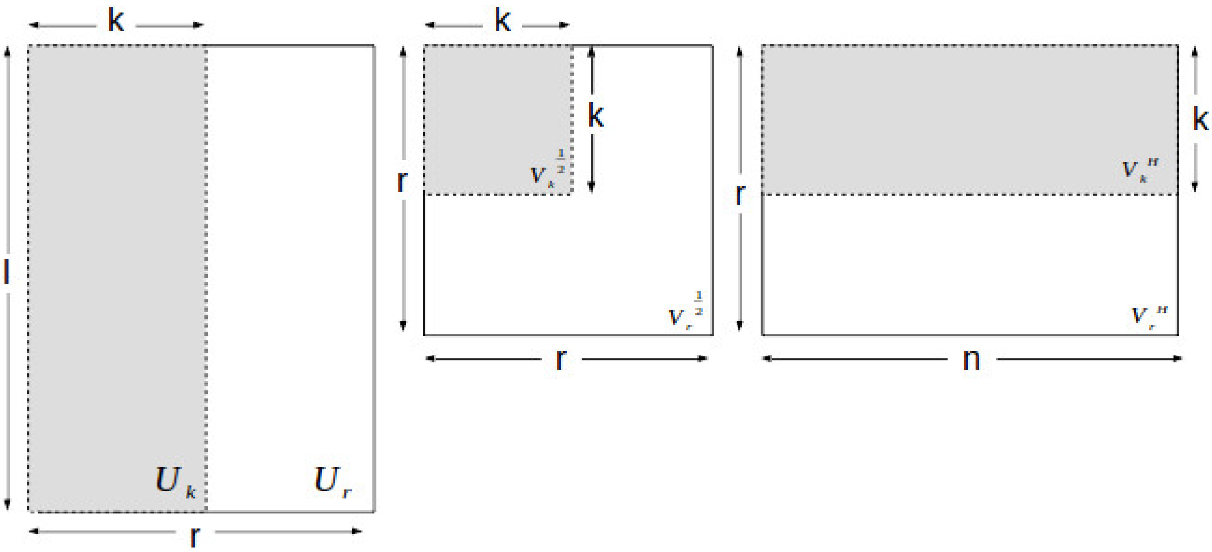
Η μέθοδος SVD έχει χρησιμοποιηθεί εκτενώς για την μείωση της διάστασης του χώρου χαρακτηριστικών σε ένα μεγάλο εύρος εφαρμογών αναγνώρισης προτύπων. Έστω οτι έχουμε την προσέγγιση χαμηλού βαθμού (low rank approximation) $\hat{X}$ του X. Αποδεικνύεται μέσω ελαχιστοποίησης του μέσου τετραγωνικού σφάλματος ότι αν η παραπάνω προσέγγιση σχηματίζεται απο την άθροιση των k μεγαλύτερων ιδιοτιμών τότε το μέσο τετραγωνικό σφάλμα της προσέγγισης είναι το ελάχιστο. Μπορούμε να καταλήξουμε στο συμπέρασμα ότι η μέθοδος SVD οδηγεί στο ελάχιστο τετραγωνικό σφάλμα και επομένως το $\hat{X}$ είναι η καλύτερη *προσέγγιση βαθμού* k του X. Η προσέγγιση αυτή δίνεται απο τον τύπο

$$X \simeq \hat{X} = \sum_{i=0}^{k-1} \sqrt{\lambda_i}\mathbf{u}_i\mathbf{v}_i^H, \quad k \leq r$$

$$= [\mathbf{u}_0, \mathbf{u}_1, \ldots, \mathbf{u}_{k-1}] \begin{bmatrix} \sqrt{\lambda_0}\mathbf{v}_0^H \\ \sqrt{\lambda_1}\mathbf{v}_1^H \\ \vdots \\ \sqrt{\lambda_{k-1}}\mathbf{v}_{k-1}^H \end{bmatrix} = U_k[\mathbf{a}_0, \mathbf{a}_1, \ldots, \mathbf{a}_{n-1},] \tag{2.4.4}$$

όπου ο μητρώο U_k αποτελείται απο τις k πρώτες στήλες του U και τα k-διάστατα διανύσματα $\mathbf{a}_i, i = 0, 1, \ldots, n-1$ είναι τα διανύσματα στήλες της $k \times n$ μήτρας του γινομένου $\Lambda^{\frac{1}{2}}V_k^H$. Το μητρώο V_k^H αποτελείται απο τις k πρώτες γραμμές του V^H και $\Lambda^{\frac{1}{2}}$ είναι διαγώνιο μητρώο με στοιχεία τις τατραγωνικές ρίζες των αντίστοιχων k ιδιαζουσών τιμών.

Στο παρακάτω σχήμα παρουσιάζεται γραφικά ώστε να γίνει καλύτερα κατανοητή η παραπάνω διαδικασία.



Σχήμα 2.1: Διαγραμματική αναπαράσταση των γινομένων των μητρών που χρησιμοποιούνται στην μέθοδο SVD. Στην προσέγγιση του X απο το $\hat{X}$, εμπλέκονται οι πρώτες k στήλες του U_r και οι πρώτες k γραμμές του V_r^H.

Απο την παραπάνω ανάλυση καταλήγουμε στο συμπέρασμα ότι το l-διάστατο διάνυσμα $\mathbf{x}_i$ προσεγγίζεται απο το k-διάστατο διάνυσμα $\mathbf{a}_i$ που βρίσκεται στον υποχώρο που ορίζουν τα $\mathbf{u}_i, i = 0, 1, \ldots, k-1$ (το $\mathbf{a}_i$ είναι στην ουσία η προβολή του $\mathbf{x}_i$ στον υποχώρο αυτόν). Επίσης, λόγω της ορθοκανονικότητας των στηλών $\mathbf{u}_i, i = 0, 1, \ldots, k-1$ του U_k βλέπουμε ότι

$$\|\mathbf{x}_i - \mathbf{x}_j\| \simeq \|U_k(\mathbf{a}_i - \mathbf{a}_j)\| = \|\sum_{m=0}^{k-1} \mathbf{u}_m(a_i(m) - a_j(m))\| = \|\mathbf{a}_i - \mathbf{a}_j\|, \quad i, j = 0, 1, \ldots, n-1 \quad (2.4.5)$$

Αντιλαμβανόμαστε λοιπόν ότι χρησιμοποιώντας την προηγούμενη προβολή και υποθέτωντας ότι η προσέγγιση είναι ικανοποιητική, η Ευκλείδια απόσταση μεταξύ $\mathbf{x}_i$ και $\mathbf{x}_j$ στον υψηλής διάστασης l-διάστατο χώρο διατηρείται (κατά προσέγγιση) κατά την προβολή στον χαμηλότερης διάστασης k-διάστατο χώρο.

2.5 Πρακτική εφαρμογή

Στο σημείο αυτό αξίζει να αναφερθεί ένα απλό παράδειγμα μέσω του οποίου μπορεί να γίνει αντιληπτή η πρακτική εφαρμογή των παραπάνω. Ας θεωρήσουμε λοιπόν ένα σύνολο n προτύπων, όπου το καθένα αναπαρίσταται απο ένα l-διάστατο διάνυσμα χαρακτηριστικών. Τότε, δοθέντως ενός άγνωστου προτύπου στόχος μας είναι να αναζητήσουμε στο σύνολο των γνωστών προτύπων που έχουμε ώστε νε βρούμε αυτό το οποίο παρουσιάζει την μεγαλύτερη ομοιότητα με το άγνωστο για το οποίο θέλουμε να καταλήξουμε σε κάποιο συγκεκριμένο συμπέρασμα. Η διαδικασία αυτή είναι εφικτή υπολογίζοντας την Ευκλείδια απόσταση μεταξύ του άγνωστου προτύπου με όλα τα γνωστά και επιλέγοντας τελικά το ζευγάρι με την μικρότερη απόσταση, δηλαδή αυτό με την μεγαλύτερη ομοιότητα.

Σε περιπτώσεις όπου τόσο ο αριθμός των διαστάσεων όσο και ο αριθμός των δειγμάτων είναι μεγάλος τότε η παραπάνω διαδικασία μπορεί να είναι ιδιαίτερα χρονοβόρα. Προκειμένου λοιπόν να απλοποιήσουμε τους υπολογισμούς μπορούμε να ακολουθήσουμε την παραπάνω διαδικασία που αναλύσαμε ώστε να μειώσουμε τις διαστάσεις του προβλήματός μας. Η διαδικασία έχει ως εξής: Αρχικά σχηματίζουμε το μητρώο δεδομένων X, διάστασης $l \times n$ με στήλες τα n διανύσματα χαρακτηριστικών. Εκτελούμε την μεθοδολογία SVD στο X και αναπαριστούμε κάθε διάνυσμα χαρακτηριστικών $\mathbf{x}_i$ με την χαμηλότερης διάστασης προβολή του, $\mathbf{a}_i$. Το άγνωστο διάνυσμα προβάλλεται στον υποχώρο που ορίζουν οι στήλες του U_k και εκτελούνται οι υπολογισμοί των Ευκλείδιων αποστάσεων στον k-διάστατο χώρο. Επειδή οι Ευκλείδιες αποστάσεις διατηρούνται κατά προσέγγιση, είναι εφικτό να αποφασίσουμε τους κοντινότερους γείτονες των διανυσμάτων εργαζόμενοι στον χώρο χαμηλότερης διάστασης. Σε περιπτώσεις για τις οποίες έχουμε $k \ll l$ επιτυγχάνεται σημαντική εξοικονόμηση στους υπολογισμούς.

Τέλος, αξίζει να αναφερθεί ότι η μεθοδολογία SVD είναι πολύ αποτελεσματική τεχνική μείωσης της διάστασης σε περιπτώσεις όπου τα δεδομένα μπορούν να περιγραφούν επαρκώς μέσω του μητρώου συνδιασποράς, για παράδειγμα περιπτώσεις όταν ακολουθούν κατανομές παρόμοιες με την Gaussian κατανομή.

Κεφάλαιο 3

Αλγόριθμοι μείωσης διαστάσεων

3.1 Γραμμική μείωση διαστάσεων

Όλες οι τεχνικές μείωσης διαστάσεων στις οποίες έχουμε αναφερθεί μέχρι στιγμής είναι κατεξοχήν τεχνικές μείωσης της διάστασης του χώρου των χαρακτηριστικών. Μάλιστα το ιδιαίτερο χαρακτηριστικό τους είναι ότι αποτελούν μεθόδους οι οποίες σέβονται την γραμμικότητα. Η μέθοδος PCA για παράδειγμα η οποία αποτελεί μια απο τις γνωστότερες αλλά και πιο ισχυρές μεθόδους γραμμικής μείωσης διαστάσεων λειτουργεί καλά αν τα σημεία των δεδομένων είναι κατανεμημένα σε ένα υπερεπίπεδο. Επίσης, όπως αναλύθηκε στην ενότητα (2.2) η μέθοδος PCA προβάλλει στις διευθύνσεις μέγιστης διασποράς. Τέλος όπως εξηγήσαμε στο προηγούμενο κεφάλαιο η ανάλυση ιδιοτιμών-ιδιοδιανυσμάτων του μητρώου συσχέτισης αποκαλύπτει την διάσταση του υπερεπιπέδου στο οποίο τα δεδομένα είναι διεσπαρμένα.

Με άλλα λόγια δηλαδή η διάσταση είναι ένα μέτρο του πλήθους των ελεύθερων μεταβλητών που είναι ϋπεύθυνες' για τον τρόπο με τον οποίο μεταβάλλεται ένα σήμα, δηλαδή για την πραγματική πληροφορία την οποία κωδικοποιούν τα δεδομένα.

Παρότι ο αλγόριθμος PCA αποτελεί μία πολύ ισχυρή και ευρέως χρησιμοποιούμενη μέθοδο μείωσης της διάστασης υπάρχουν περιπτώσεις στις οποίες η μέθοδος αποτυγχάνει. Τέτοιες είναι περιπτώσεις κατα τις οποίες ο μηχανισμός παραγωγής των δεδομένων είναι έντονα μη γραμμικός με αποτέλεσμα τα δεδομένα να κείτονται σε πιο περίπλοκες πολλαπλότητες. Ας πάρουμε για παράδειγμα τις εξισώσεις

$$x_1 = rcos\theta, \quad x_2 = rsin\theta$$

Προφανώς απο τις παραπάνω εξισώσεις είναι φανερό ότι το x βρίσκεται στην περιφέρεια κύκλου ακτίνας r. Πρόκειται δηλαδή για πρόβλημα μονοδιάστατης πολλαπλότητας αφού αρκεί μια μόνο μεταβλητή για την περιγραφή των δεδομένων. Η παράμετρος αυτή είναι η απόσταση κατα μήκος της περιφέρειας απο ένα σημείο(αφετηρία) πάνω στην περίμετρο του κύκλου. Αν λοιπόν εφαρμόσουμε την μέθοδο PCA στο παραπάνω σύνολο δεδομένων τότε η απάντηση που θα μας δώσει για την διάσταση των δεδομένων θα είναι, λανθασμένα προφανώς, ίση με δύο.

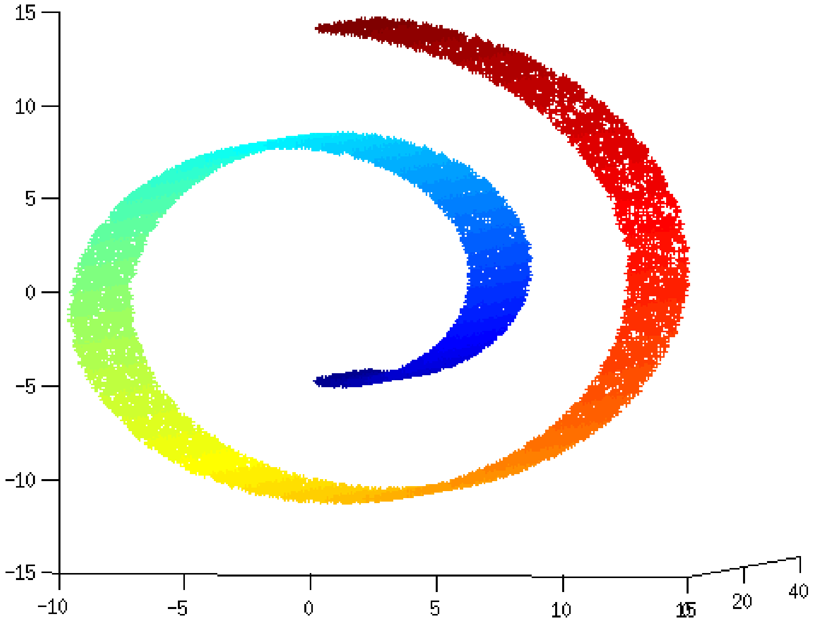
Περιπτώσεις όπως οι παραπάνω απαιτούν αλγορίθμους μείωσης διάστασης και εξαγωγής χαρακτηριστικών οι οποίοι να λαμβάνουν υπόψιν την γεωμετρία του προβλήματος ώστε να μπορούν να εξάγουν αποτελεσματικά συμπεράσματα για την διάσταση των δεδομένων. Στον τομέα της υπολογιστικής όρασης για παράδειγμα, ο οποίος όπως αναφέραμε και παραπάνω αποτελεί βασικό κομμάτι της εν λόγω διατριβής, απαιτούνται κατεξοχήν αλγόριθμοι μη γραμμικής μείωσης διαστάσεων αφού οι εικόνες ή τα χαρακτιριστκά των εικόνων τα οποία αποτελούν τα δεδομένα μας είναι κατα κύριο λόγο μη γραμμικά.

3.2 Μη γραμμική μείωση διαστάσεων

Υπάρχει λοιπόν μια ευρεία γκάμα εφαρμογών οι οποίες απαιτούν αλγορίθμους μη γραμμικής μείωσης διαστάσεων. Αυτό συμβαίνει διότι στις συγκεκριμένες εφαρμογές η γεωμετρική αναπαράσταση των δεδομένων είναι τέτοια ώστε απαιτείται να βρεθεί μια ενσωμμάτωση μικρότερης διάστασης η οποία βρίσκεται 'κρυμμένη' στον χώρο των αρχικών διαστάσεων. Θα πρέπει μάλιστα κατά την διαδικασία αυτή να ληφθούν υπόψιν τα γεωμετρικά χαρακτηριστικά του χώρου των δεδομένων.

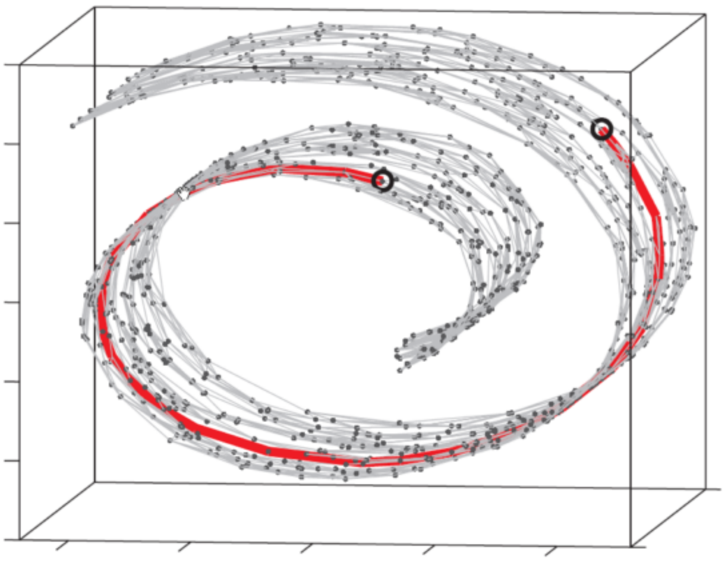
Έχει πολύ μεγάλη σημασία στο σημείο αυτό να κατανοήσουμε τι εννοούμε όταν αναφερόμαστε στα γεωμετρικά χαρακτηριστικά του προβλήματος. Το πιο χαρακτηριστικό και ευρέως χρησιμοποιούμενο παράδειγμα για τον σκοπό αυτό είναι ένα τεχνητό σετ δεδομένων, με την όνομασία Swiss

Roll το οποίο φαίνεται στην παρακάτω εικόνα.



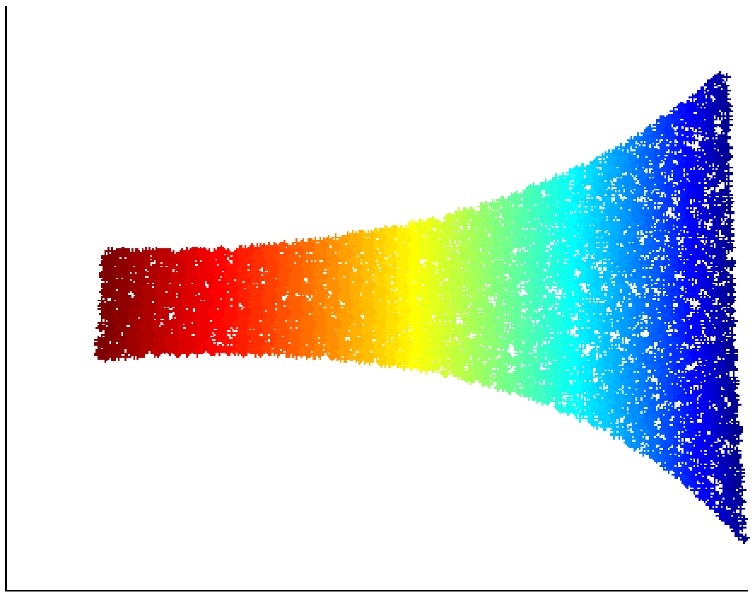
Σχήμα 3.1: Swiss Roll Synthetic Dataset.

Αυτό που αξίζει να παρατηρηθεί λοιπόν στο παραπάνω σετ δεδομένων είναι ότι αν για παράδειγμα διαλέξουμε κάποιο οποιοδήποτε σημείο του απο την κόκκινη περιοχή και προσπαθήσουμε να βρούμε ποιά δεδομένα αποτελούν κοντινότερους γείτονες του σημείου αυτού πιθανότατα θα πέφταμε στην παγίδα, όπως και οι τεχνικές γραμμικής μείωσης διαστάσεων, να πούμε ότι κάποια σημεία απο την μπλέ περιοχή βρίσκονται και αυτά στην γειτονιά του σημείου που διαλέξαμε. Αυτό προφανώς είναι λάθος αφού απο τον χρωματισμό των παραπάνω δεδομένων αντιλαμβανόμαστε ότι στην πραγματικότητα τα μπλέ δεδομένα βρίσκονται πολύ μακριά απο τα κόκκινα. Ο παραπάνω εσφαλμένος συλλογισμός αναπαρίσταται στο παρακάτω γράφημα.



Σχήμα 3.2: Swiss Roll Synthetic Dataset Manifold Learning Path.

Αντιλαμβανόμαστε λοιπόν, μέσω της παραπάνω απεικόνισης ότι θα πρέπει να ληφθεί υπόψιν η γεωμετρία του προβλήματος ώστε σε καμιά περίπτωση υπολογίζοντας κοντινότερες αποστάσεις να συμπεριλάβουμε το αρχικό και το τελικό σημείο ως κοντινούς γείτονες, ενώντάς τα απευθείας μεταξύ τους. Αυτή είναι και η διαφορά των αλγορίθμων μη γραμμικής μείωσης διαστάσεων με αυτούς της γραμμικής. Για να γίνει πλήρως κατανοητός ο τρόπος μείωσης των διαστάσεων του παραπάνω σετ δεδομένων, δίνεται η απεικόνιση των δεδομένων σε χώρο χαμηλής διάστασης μετά απο την εφαρμογή αλγορίθμου μη γραμμικής μείωσης διαστάσεων.



Σχήμα 3.3: Dimensionality Reduction with LLE - 3D (K=16,d=2).

Απο την παραπάνω απεικόνιση μπορούμε να φανταστούμε ότι κάνοντας μείωση των διαστάσεων στην πραγματικότητα 'ξετυλίξαμε' το Swiss Roll και έτσι απο τον αρχικό χώρο των τριών διαστάσεων στην πραγματικότητα η εγγενής διάσταση των δεδομένων είναι ίση με δύο. Στις επόμενες ανότητες θα γίνει παρουσίαση των πιο γνωστών μεθόδων μη γραμμικής μείωσης διαστάσεων καθώς επίσης θα γίνει και η μαθηματική τους ανάλυση.

3.2.1 ISOMAP

Ένας βασικός αλγόριθμος μη γραμμικής μείωσης διαστάσεων είναι ο αλγόριθμος Ισομετρική απεικόνιση (Isometric Mapping - ISOMAP). Ο αλγόριθμος αυτός υιοθετεί την άποψη ότι μόνο οι γεωδαιτικές αποστάσεις μεταξύ όλων των ζευγών των σημείων των δεδομένων μπορούν να αντικατοπτρίσουν την πραγματική δομή της πολλαπλότητας του προβλήματος. Η παραπάνω διατύπωση αντικατοπτρίζει το παράδειγμα που δόθηκε στο γράφημα (3.2), και τονίζει το γεγονός ότι οι Ευκλείδιες αποστάσεις μεταξύ σημείων μιας πολλαπλότητας δεν μπορούν να την αναπαραστήσουν ικανοποιητικά διότι σημεία (στο γράφημα τα δύο σημεία που έχουν επισυμανθεί με μαύρους κύκλους) που είναι απομακρυσμένα μεταξύ τους, σύμφωνα με την γεωδαιτική απόσταση, μπορεί να θεωρηθούν, λανθασμένα, κοντικά ως προς την Ευκλείδια απόστασή τους.

Ουσιαστικά η μέθοδος ISOMAP είναι μια παραλλαγή του αλγορίθμου Multi Dimensional Scaling - MDS, με την διαφορά ότι οι Ευκλείδιες αποστάσεις αντικαθίστανται απο τις αντίστοιχες γεωδαιτικές κατά μήκος της πολλαπλότητας των δεδομένων. Η ουσία του αλγορίθμου είναι να εκτιμηθούν σωστά οι γεωδαιτικές αποστάσεις μεταξύ σημείων τα οποία είναι απομακρυσμένα μεταξύ τους. Ο αλγόριθμος μπορεί να χωριστεί σε δύο βασικά βήματα:

Βήμα-1:
Για κάθε σημείο $x_i, i = 1, 1 \ldots, n$, υπολόγισε τους πλησιέστερους γείτονες και κατασκέυασε έναν γράφο $G(V, E)$ του οποίου οι κορυφές αναπαριστούν πρότυπα εισόδου και οι ακμές συνδέουν τους πλησιέστερους γείτονες. Οι παράμετροι k ή ϵ είναι παράμετροι που καθορίζονται απο τον χρήστη). Στις ακμές ανατίθενται βάρη σύμφωνα με τις αντίστοιχες Ευκλείδιες αποστάσεις (για τους πλησιέστερους γείτονες αυτή είναι μια καλή προσέγγιση της γεωδαιτικής απόστασης).

Βήμα-2:
Υπολόγισε ανα ζεύγος την γεωδαιτική απόσταση για όλα τα ζεύγη κατα μήκος των συντομότερων διαδρομών μέσα στον γράφο. Το πιο σημαντικό σημείο, είναι ότι η γεωδαιτική απόσταση μεταξύ δύο οποιονδήποτε σημείων της πολλαπλότητας μπορεί να προσεγγιστεί μέσω της συντομότερης διαδρομής που ενώνει τα δύο σημείο στο γράφο $G(V, E)$. Ο πιο γνωστός αλγόριθμος υλοποίησης της παραπάνω διαδικασίας είναι ο αλγόριθμος Djikstar με πολυπλοκότητα $\mathcal{O}(n^2 \ln n + n^2 k)$, μέγεθος απαγορευτικό για τις περισσότερες πρακτικές εφαρμογές.

Εφόσον έχουν εκτελεστεί τα δύο αυτά βήματα είμαστε πλέον σε θέση νε εφαρμόσουμε την κλασική μέθοδο MDS. Το πρόβλημα λοιπόν απο εδώ και στο εξής γίνεται ισοδύναμο με την εφαρμογή της ανάλυσης ιδιοδιανυσμάτων του αντίστοιχου μητρώου Gram και την επιλογή των m περισσότερο σημαντικών ιδιοδιανυσμάτων για την αναπαράσταση του χώρου χαμηλής διάστασης. Μετά απο αυτή την αναπαράσταση, οι Ευκλείδιες αποστάσεις μεταξύ των σημείων του χώρου χαμηλής διάστασης ταιριάζουν με τις αντίστοιχες γεωδαιτικές αποστάσεις στην πολλαπλότητα του αρχικού χώρου υψηλής διάστασης. Όπως και στις μεθόδους PCA και MDS η διάσταση m εκτιμάται απο το πλήθος των m περισσότερο σημαντικών ιδιοτιμών. Αποδεικνύεται τέλος ότι η μέθοδος ISO-MAP ασυμπτωτικά ($n \rightarrow \inf$) θα ανακτήσει την αληθινή διάσταση για ένα σύνολο δεδομένων μη γραμμικής πολλαπλότητας.

3.2.2 Laplassian Eigenmaps

Η μέθοδος αυτή στηρίζεται στην υπόθεση ότι τα σημεία των δεδομένων βρίσκονται σε μια λεία πολλαπλότητα $M \supset Q$, της οποίας η εγγενής διάσταση είναι ίση με $m < N$ και είναι ενσωματωμένη στον $\Re^N$, δηλαδή $M \supset \Re^N$. Η διάσταση m δίνεται ως παράμετρος απο τον χρήστη και εξαρτάται απο το σύνολο των δεδομένων για κάθε εφαρμογή. Η κύρια φιλοσοφία πίσω απο την μέθοδο είναι να υπολογιστεί η αναπαράσταση των δεδομένων σε χώρο χαμηλής διάστασης, έτσι ώστε η τοπική πληροφορία γειτνίασης στον χώρο $Q \supset M$ να διατηρείται κατά βέλτιστο τρόπο. Με τον τρόπο αυτό προσπαθούμε να βρούμε μια λύση που αντανακλά τη γεωμετρική δομή της πολλαπλότητας. Για την επίτευξη αυτού απαιτούνται τα παρακάτω βήματα:

Βήμα-1: Κατασκευή ενός γράφου $G = (V, E)$, όπου $V = v_i, i = 1, 2, \ldots, n$ είναι ένα σύνολο κορυφών και $E = \epsilon_{ij}$ το σύνολο των ακμών που συνδέουν κορυφές (v_i, v_j). Κάθε κόμβος v_i του γράφου αντιστοιχεί σε ένα σημείο $\mathbf{x}_i$ του συνόλου των δεδομένων X. Συνδέουμε τις v_i,v_j, δηλαδή εισάγουμε την ακμή ϵ_{ij} μεταξύ των αντίστοιχων κόμβων, αν τα σημεία $\mathbf{x}_i, \mathbf{x}_j$ είναι μεταξύ τους κοντινά. Η μέθοδος ορίζει την εγγύτητα αυτή με δύο τρόπους:

1. $\|\mathbf{x}_i - \mathbf{x}_j\|^2 < \epsilon$, για κάποια παράμετρο ϵ η οποία ορίζεται απο τον χρήστη. Με $\| \cdot \|$ ορίζουμε την πράξη της Ευκλείδιας νόρμας στον χώρο $\Re^N$.

2. Το $\mathbf{x}_j$ είναι μεταξύ των k πλησιέστερων γειτόνων του $\mathbf{x}_i$ ή και αντίστροφα, με το k να είναι και σε αυτή την περίπτωση είσοδος η οποία καθορίζεται απο τον χρήστη. Επίσης οι γείτονες επιλέγονται χρησιμοποιώντας την μετρική της Ευκλείδιας απόστασης στον χώρο $\Re^N$. Η χρήση της Ευκλείδιας απόστασης αιτιολογείται απο την υπόθεση ότι η πολλαπλότητα είναι λεία, γεγονός που μας επιτρέπει να προσεγγίσουμε, τοπικά, τη γεωδαισία της πολλαπλότητας με Ευκλείδιες αποστάσεις.

Για να αποσαφηνιστεί πλήρως η παραπάνω διατύπωση δίνεται το χαρακτηριστικό παράδειγμα όπου θεωρούμε μια σφαίρα ενσωματωμένη στον τρισδιάστατο χώρο, και έστω κάποιος περιορίζεται να ζεί πάνω στην επιφάνεια της σφαίρας. Τότε η συντομότερη διαδρομή απο ένα σημείο της σφαίρας σε ένα άλλο είναι η γεωδαιτική διαδρομή μεταξύ των δύο σημείων. Προφανώς αυτή δεν θα είναι ευθεία γραμμή, αλλά ένα τόξο στην επιφάνεια της σφαίρας. Παρ΄όλα αυτά όμως, αν τα δύο σημεία είναι πολύ κοντά μεταξύ τους, η γεωδαιτική απόσταση μπορεί να προσεγγιστεί απο την Ευκλείδια απόσταση, υπολογισμένη στον τρισδιάστατο χώρο.

Βήμα-2: Κάθε ακμή ϵ_{ij} συσχετίζεται με ένα βάρος $W(i,j)$. Για κόμβους που δεν συνδέονται μεταξύ τους, τα αντίστοιχα βάρη είναι μηδέν. Κάθε βάρος $W(i,j)$ είναι ένα μέτρο της εγγύτητας των αντίστοιχων γειτόνων $\mathbf{x}_i, \mathbf{x}_j$. Μια τυπική επιλογή είναι

$$W(i,j) = \begin{cases} \exp(\|\frac{\mathbf{x}_i - \mathbf{x}_j}{\sigma^2}\|) & ,if \quad v_i, v_j \quad neighbors \\ 0 & ,not \quad neighbors \end{cases}$$

με σ^2, παράμετρος η οποία ορίζεται και αυτή απο τον χρήστη. Σχηματίζουμε το μητρώο βαρών W, μεγέθους $(n \times n)$, το οποίο έχει για στοιχεία τα βάρη $W(i,j)$. Σημειώνουμε ότι το W είναι συμμετρικό και αραιό αφού στην πράξη προκύπτει ότι πολλά απο τα στοιχεία του είναι μηδενικά.

Βήμα-3: Ορίζεται το διαγώνιο μητρώο D με στοιχεία $D_{ij} = \sum_j W(i,j), i = 1, 2, \ldots, n$, καθώς και το μητρώο $L = D - W$. Το τελευταίο είναι γνωστό ως το μητρώο Laplace του γράφου $G = (V, E)$. Εφαρμόζεται η γενικευμένη ανάλυση σε ιδιοτιμές και ιδιοδιανύσματα

$$\Lambda\mathbf{v} = \lambda D\mathbf{v}$$

Έστω $0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_m$, οι $m+1$ μικρότερες ιδιοτιμές. Αγνοείται η ιδιοτιμή $\mathbf{v}_0$ που αντιστοιχεί στην ιδιοτιμή $\lambda_0 = 0$ και επιλέγονται τα υπόλοιπα m ιδιοδιανύσματα $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m,$. Στην εκτελείται η απεικόνιση

$$\mathbf{x}_i \in \Re^N \mapsto \mathbf{y}_i \in \Re^m, i = 1, 2, \ldots, n$$

όπου

$$\mathbf{y}_i^T = [\mathbf{v}_1(i), \mathbf{v}_2(i), \ldots, \mathbf{v}_m(i)], i = 1, 2, \ldots, m$$

Όπως έχουμε αναλύσει στην αντίστοιχη εντότητα η πολυπλοκότητα υπολογισμού ιδιοτιμών και ιδιοδιανυσμάτων είναι, γενικά, της τάξης $\mathcal{O}(n^3)$. Ωστόσο για αραιά μητρώα, όπως στην συγκεκριμένη περίπτωση το L, μπορούν να εφαρμοστούν αποτελεσματικές τεχνικές με αποτέλεσμα την μείωση της πολυπλοκότητας σε τάξη μικρότερη απο $\mathcal{O}(n^2)$. Η πιο γνωστή και αποτελεσματική τεχνική για τον σκοπό αυτό είναι ο αλγόριθμος Lanczos.

Ο αλγόριθμος Laplassian Eigenmaps ο οποίος αναλύθηκε παραπάνω, ανήκει στην ίδια κατηγορία (*μέθοδοι μείωσης διαστάσεων που βασίζονται σε γράφους*) με τον αλγόριθμο **Locally Linear Embeddings - LLE** ο οποίος αποτελεί και βασικό αντικείμενο της εν λόγω διατριβής. Οι δύο αλγόριθμοι έχουν πολύ κοινή λογική και μεθοδολογία και γι αυτό στο επόμενο κεφάλαιο στο οποίο γίνεται αναλυτική μαθηματική ανάλυση του LLE θα αποσαφηνιστούν και τα παραπάνω βήματα του Laplassian Eigenmaps καθώς τα βήματα τους είναι πανομοιότυπα.

Κεφάλαιο 4

Ο αλγόριθμος Locally Linear Embeddings - LLE

Ο αλγόριθμος LLE ανήκει στην κατηγορία αλγορίθμων μη γραμμικής μείωσης διαστάσεων με την χρήση γράφων και αποτελεί μια απο τις αποτελεσματικότερες αλλά και γρηγορότερες τεχνικές αυτού του είδους. Όπως αναφέραμε και στο προηγούμενο κεφάλαιο, βασική υπόθεση της μεθόδου είναι ότι τα δεδομένα μας βρίσκονται σε μια αρκετά λεία πολλαπλότητα, διάστασης m, και η οποία είναι ενσωματωμένη στον υποχώρο του $\Re^N$, με $m < N$. Η υπόθεση για το λείο της πολλαπλότητας μας επιτρέπει να υποθέσουμε επιπλέον ότι, με δεδομένη την ύπαρξη αρκετών δεδομένων και ότι η πολλαπλότητα είναι ≪καλά≫ δειγματοληπτημένη, τα κοντινά σημεία βρίσκονται πάνω (ή κοντά) σε ένα ≪τοπικό≫ γραμμικό τμήμα της πολλαπλότητας.

4.1 Ο αλγόριθμος ως τεχνική μη γραμμικής μείωσης διαστάσεων

Δεδομένης της αποτελεσματικότητας του αλγορίθμου να ανακαλύπτει τον χώρο μειωμένης διάστασης στον οποίο βρίσκεται ενσωματωμένη η πληροφορία ενός προβλήματος, ο αλγόριθμος έχει χρησιμοποιηθεί με επιτυχία σε αρκετές πρακτικές εφαρμογές. Ιδιαίτερο ενδιαφέρον παρουσιάζουν νέες μελέτες κυρίως απο τον χώρο της Ιατρικής [1] [2]. Απο τις αναφορές αυτές είναι φανερό ότι ο ρόλος της μείωσης των διαστάσεων μπορεί να καθορίσει σε μεγάλο βαθμό την βελτίωση του αποτελέσματος ταξινόμησης. Στις συγκεκριμένες περιπτώσεις στόχος είναι να γίνει σωστή πρόβλεψη για το αν κάποιος ασθενής πάσχει απο μια συγκεκριμένη ασθένεια ή βρίσκεται στην ευπαθή ομάδα με μεγάλη πιθανότητα να του παρουσιαστεί στο μέλλον. Φαίνεται οτι ο αλγόριθμος LLE είναι ένα πολύ ισχυρό εργαλείο το οποίο μπορεί να υλοποιήσει την μείωση των διαστάσεων σε τέτοιου είδους εφαρμογές και μάλιστα επιφέροντας σημαντικά και ουσιαστικά αποτελέσματα. Ο αλγόριθμος επίσης, έχει χρησιμοποιηθεί και σε εφαρμογές ταξινόμησης με σετ δεδομένων ευρέως διαδεδομένα στον χώρο της αναγνώρισης προτύπων [3] [4] [5], όπως το σετ δεδομένων με χειρόγραφα ψηφία MNIST.

4.2 Μαθηματική ανάλυση και υλοποίηση του αλγορίθμου Locally Linear Embeddings σε κώδικα MATLAB

Ο αλγόριθμος LLE[6] αποτελεί το κύριο κομμάτι της εν λόγω διατριβής και η υλοποίηση του έχει στηριχθεί στον αλγόριθμο της παραπάνω αναφοράς. Ο ψευδοκώδικας είναι διαθέσιμος στην παρακάτω τοποθεσία LLE Algorithm Pseudocode. Παρ' όλα αυτά στην συγκεκριμένη υλοποίηση έχουν γίνει συγκεκριμένες βελτιστοποιήσεις σε κάποια βήματα του αλγορίθμου, όπως για παράδειγμα η χρήση του αλγορίθμου κοντινότερων γειτόνων, υλοποιημένο σε CUDA, με σκοπό την μείωση του χρόνου εκτέλεσης του συγκεκριμένου βήματος.

4.2.1 Βήμα-1: Εύρεση του πίνακα γειτνίασης

Κατά το πρώτο βήμα του αλγορίθμου γίνεται ο υπολογισμός των κοντινότερων γειτόνων για κάθε σημείο X_i του συνόλου των δεδομένων. Στο βήμα αυτό ο χρήστης επιλέγει ανάλογα με την κάθε εφαρμογή έναν αριθμό γειτόνων K και χρησιμοποιεί κάποιον αλγόριθμο υπολογισμού κοντινότερων γειτόνων για κάθε ένα απο τα σημεία του δείγματος. Με τον τρόπο αυτό έχει υπολογιστεί ο τετραγωνικός πίνακας $N \times N$, ο οποίος δίνει πληροφορία για κάθε σημείο του δείγματος ως προς τους K κοντινότερους γείτονές του.

Ο τρόπος υπολογισμού του πίνακα αυτού στην συγκεκριμένη υλοποίηση γίνεται μέσω της συνάρτησης knnsearch του MATLAB για την σειριακή υλοποίηση και με την συνάρτηση gpu_knn για την παράλληλη υλοποίηση. Η συνάρτηση αυτή αντιπροσωπεύει την κύρια συνάρτηση gpuknnHeap του πακέτου knn-toolbox, η οποία με την σειρά της αποτελεί το πέρασμα απο τον κώδικα MATLAB στην συνάρτηση πυρήνα υπολογισμού κοντινότερων γειτόνων με τη χρήση παράλληλης υλοποίησης σε CUDA γραμμένη στην γλώσσα προγραμματισμού C. Η υλοποίηση αυτή χρησιμοποιεί την Ευκλείδια απόσταση ως μέθοδο προσδιορισμού των κοντινότερων γειτόνων. Παρ΄ όλα αυτά στο βήμα αυτό μπορούν να χρησιμοποιηθούν και άλλες μετρικές γειτνίασης όπως για παράδειγμα ο προσδιορισμός κοντινών γειτόνων με τη χρήση σφαίρας ακτίνας ϵ. Επίσης, μια άλλη γνωστή μέθοδος επίλυσης του βήματος αυτού η οποία βελτιώνει τον χρόνο εκτέλεσης είναι η χρήση KD-trees.

4.2.2 Βήμα-2: Εύρεση του πίνακα βαρών W

Στο δεύτερο αυτό βήμα του αλγορίθμου στόχος είναι να υπολογιστεί ο πίνακας βαρών $W(i,j), i,j = 1,2,\ldots,n$, μέσω των οποίων είναι εφικτή η ανακατασκευή του κάθε δείγματος X_i μέσω των βαρών που αντιστοιχούν στους κοντινούτερους γείτονές του. Πιο απλά στο βήμα αυτό θέλουμε να προσδιορίζουμε κάθε σημείο X_i του δείγματός μας, ελαχιστοποιώντας την συνάρτηση κόστους

$$\arg\min_{w} E_w = \sum_{i=1}^{n} \|\mathbf{X_i} - \sum_{j=1}^{n} W(i,j)\mathbf{X}_j\|^2 \tag{4.2.1}$$

η οποία στην πραγματικότητα αυτό που προσπαθεί να ανακαλύψει είναι οι κοντινότεροι γείτονες j του σημείου X_i οι οποίοι ασκούν την σημαντικότερη επιρροή πάνω του ως προς την ανακατασκευή του. Η ελαχιστοποίηση της συνάρτησης αυτής γίνεται εφαρμόζοντας τον αλγόριθμο ελαχίστων τετραγώνω Least squares εξασφαλίζοντας παράλληλα κάποιες απαραίτητες ιδιότητες για τον πίνακα βαρών W. Καταρχήν θα πρέπει να ισχύει ότι το κάθε δείγμα X_i θα πρέπει να μπορεί αν ανακατασκευαστεί μόνο απο τους κοντινότερους γείτονές του γεγονός που θέτει τον περιορισμό $W(i,j) = 0$ στην περίπτωση κατά την οποία το j στοιχείο δεν είναι γείτονας του i. Επίσης θα πρέπει τα στοιχεία κάθε γραμμής του μητρώου βαρών W να αθροίζονται στην μονάδα, δηλαδή $\sum_{j=1}^{n} W(i,j) = 1$, ώστε να εξασφαλιστεί η αμεταβλητότητα κατά την μεταφορά. Με τους περιορισμούς αυτούς λοιπόν εξασφαλίζεται ότι τα βάρη τα οποία ελαχιστοποιούν την παραπάνω συνάρτηση κόστους είναι αμετάβλητα κατά την περιστροφή, την μεταφορά και την κλιμάκωση.

Η παραπάνω διαδικασία υλοποιείται με τον παρακάτω κώδικα σε MATLAB.

```
W = zeros(K,N);
for ii=1:N
    z = X(:,neighborhood(:,ii))-repmat(X(:,ii),1,K);
    C = z'*z;                                    % local covariance
    C = C + eye(K,K)*tol*trace(C);               % regularlization (K>D)
    W(:,ii) = C\ones(K,1);                       % solve Cw=1
    W(:,ii) =W(:,ii)/sum(W(:,ii));               % enforce sum(w)=1
end;
```

Για να γίνει κατανοητή η παραπάνω διαδικασία ακολουθούμε τον εξής συλλογισμό. Ας πάρουμε για παράδειγμα ένα σημείο x το οποίο έχει K κοντινούς γείτονες n_j και βάρη ανακατασκευής w_j για τα οποία ισχύει η συνθήκη $\sum_j w_j = 1$. Τότε μπορούμε να γράψουμε την συνάρτηση κόστους ως

$$\epsilon = \mid \overrightarrow{x} - \sum_j w_j \overrightarrow{n}_j \mid^2 = \mid \sum_j w_j (\overrightarrow{x} - \overrightarrow{n}_j) \mid^2 = \sum_{jk} w_j w_k C_{jk} \tag{4.2.2}$$

Στην παραπάνω σχέση χρησιμοποιήσαμε το μητρώο Gram το οποίο ορίζεται ως

$$C_{jk} = (\overrightarrow{x} - \overrightarrow{n}_j) \cdot (\overrightarrow{x} - \overrightarrow{n}_k) \tag{4.2.3}$$

Εκ κατασκευής για τον πίνακα Gram έχουμε ότι είναι συμμετρικός και θετικά ημιορισμένος. Σύμφωνα με τα παραπάνω λοιπόν τα βέλτιστα βάρη ανακατασκευής w_j της συνάρτησης κόστους μπορούν να υπολογιστούν, αφού μέσω του πολλαπλασιαστή Lagrange εξασφαλίσουμε τη συνθήκη $\sum_j w_j = 1$, μέσω της επίλυσης του συστήματος

$$w_j = \frac{\sum_k C_{jk}^{-1}}{\sum_{lm} G_{lm}^{-1}} \tag{4.2.4}$$

Όπως είχαμε αναφέρει στην παράγραφο (2.3) οι πίνακες $X_T X$(πίνακας συνδιασποράς) και $X X_T$(πίνακας Gram) έχουν τις ίδιες ιδιοτιμές και ιδιοδιανύσματα τα οποία σχετίζονται μεταξύ τους. Για τον λόγο αυτό μπορεί να παραληφθεί η αντιστροφή του πίνακα Gram, όπως φαίνεται και στην υλοποίηση που παρατέθηκε παραπάνω, λύνοντας το σύστημα $\sum_j C_{jk} w_k = 1$ και έπειτα απαιτώντας τον περιορισμό $\sum_j w_j = 1$ ο οποίος υλοποιείται με την τελευταία γραμμή του παραπάνω κώδικα. Επίσης βλέπουμε ότι στην υλοποίηση του κώδικα δεν υπολογίζεται ο πίνακας Gram αλλά αυτός της συνδιασποράς και στην συνέχεια ακολουθείται η διαδικασία που αναλύθηκε παραπάνω. Τελευταία διευκρίνηση για την παραπάνω διαδικασία, η γραμμή 5 του κώδικα, στην οποία γίνεται κανονικοποίηση του πίνακα συνδιασποράς. Αυτό απαιτείται στην περίπτωση για την οποία ο πίνακας συνδιασποράς προκύπτει μοναδιαίος ή πολύ κοντά σε αυτόν, οπότε και δεν υπάρχει μοναδική λύση του συστήματος.

Καταλήγουμε λοιπόν μέσω της παραπάνω διαδικασίας στον υπολογισμό του μητρώου βαρών W για το σύνολο των δεδομένων. Ο τρόπος μάλιστα με τον οποίο έγινε ο υπολογισμός αυτός εξασφαλίζει το γεγονός ότι η εσωτερική ενσωματωμένη γεωμετρία η οποία υπάρχει στην γειτονιά ενός σημείο X_i του συνόλου των δεδομένων θα εξακολουθεί να υπάρχει και στον χώρο της μειωμένης διάστασης. Το γεγονός αυτό εξασφαλίζεται απο την ανεξαρτησία των βαρών ως προς την περιστροφή, την μεταφορά και την κλιμάκωση αλλά και απο το γεγονός ότι οι γείτονες του σημείου X_i στον χώρο αρχικών διαστάσεων D θα εξακολουθούν να αποτελούν γείτονες του σημείου Y_i (προβολή του X_i απο τον χώρο υψηλής διάστασης στο σημείο Y_i χαμηλής διάστασης). Αυτό συμβαίνει επίσης, διότι όπως θα δούμε παρακάτω με τα βάρη με τα οποία γίνεται ανακατασκεύη του X_i τα ίδια θα χρησιμοποιηθούν και για την κατασκευή του Y_i στον χώρο μειωμένης διάστασης. Συνέπεια λοιπόν των παραπάνω είναι ότι τα βάρη w_j που υπολογίστηκαν δεν εξαρτώνται απο το εκάστοτε σημείο αλλά κωδικοποιούν πληροφορία σχετική με τα εγγενή χαρακτηριστικά κάθε γειτονιάς τα οποία και διατηρούνται κατά την ενσωμάτωση των δεδομένων στον χώρο χαμηλότερης διάστασης.

4.2.3 Βήμα-4: Επιλογή των τελικών διαστάσεων με τη χρήση του πίνακα W

Στο τελευταίο βήμα του αλγορίθμου πραγματοποιείται η μείωση των διαστάσεων των δειγμάτων απο τον χώρο υψηλής διάστασης D σε έναν χαμηλότερης d. Η διαδικασία αυτή πραγματοποιείται όπως αναφέραμε και παραπάνω χρησιμοποιώντας τον πίνακα των βαρών W και τα οποία έχουν την ιδιότητα ότι ανταναкλούν τις εγγενείς ιδιότητες της τοπικής γεωμετρίας στην οποία υπόκεινται τα δεδομένα. Η λύση λοιπόν προκύπτει επιλύοντας και πάλι ένα πρόβλημα ελαχιστοποίησης το οποίο ορίζεται ως

$$\arg\min_{w} E_y = \sum_{i=1}^{n} \|\mathbf{Y_i} - \sum_{j=1}^{n} W(i,j)\mathbf{Y}_j\|^2 \tag{4.2.5}$$

Και σε αυτή την περίπτωση απαιτούμε την διατήρηση των συνθηκών, $\sum_i Y_i = 0$ ώστε να εξασφαλιστεί η αμεταβλητότητα ως προς την μεταφορά, και $\frac{1}{N}\sum_i Y_i Y i^T = I$ η οποία εξασφαλίζει ότι οι διαστάσεις d θα είναι δευτέρου βαθμού ασυσχέτιστες , οτι τα βάρη ανακατασκευής για τις διαστάσεις d θα υπολογιστούν σε κοινή κλίμακα και ότι αυτή η κλίμακα θα είναι μοναδιαίου βαθμού. Ο πίνακας I συμβολίζει τον μοναδιαίο πίνακα διάστασης $d \times d$.

Η λύση της εξίσωσης (4.2.5) για τα άγνωστα στοιχεία $y_i, i = 1, 2, \ldots, n$, είναι ισοδύναμη με την εύρεση των $d + 1$ μικρότερων ιδιοτιμών του τετραγωνικού πίνακα M ο οποίος προκύπτει απο την σχέση

$$E_y = \sum_{i=1}^{n} \|\mathbf{Y_i} - \sum_{j=1}^{n} W(i,j)\mathbf{Y}_j\|^2 = \mid (I - W)Y \overset{2}{\min} = Y_T M Y \tag{4.2.6}$$

και ισούτε με

$$M = (I - W)^T (I - W) \tag{4.2.7}$$

Ο πίνακας αυτός έχει διαστάσεις $N \times N$, όπου N το πλήθος των δεδομένων εισόδων. Παρ'όλα αυτά ο πίνακας αυτός στην πράξη προκύπτει αραιός (sparce matrix) γεγονός το οποίο απλοποιεί σημαντικά τους υπολογισμούς, ιδιαίτερα για μεγάλες τιμές του N. Στην συνέχεια χρησιμοποιώντας τον πολλαπλασιαστή Lagrange καταλήγουμε στην επίλυση της εξίσωσης

$$(M - \Lambda)Y^T = 0 \tag{4.2.8}$$

όπου Λ είναι ο διαγώνιος πίνακας των πολλαπλασιαστών Lagrange. Η υλοποίηση της παραπάνω διαδικασίας στην συγκεκριμένη υλοποίηση σε κώδικα MATLAB είναι η παρακάτω

```
% M=eye(N,N); % use a sparse matrix with storage for 4KN nonzero elements
M = sparse(1:N,1:N,ones(1,N),N,N,4*K*N);
for ii=1:N
    w = W(:,ii);
    jj = neighborhood(:,ii);
    M(ii,jj) = M(ii,jj) - w';
    M(jj,ii) = M(jj,ii) - w;
    M(jj,jj) = M(jj,jj) + w*w';
end;
```

Η παραπάνω επίλυση του προβλήματος ανάλυσης ιδιοτιμών μας οδηγεί στον προσδιορισμό των ιδιοδιανυσμάτων τα οποία αποτελούν και λύσεις του πίνακα M. Επίσης τα ιδιοδιανύσματα με τις μικρότερες ιδιοτιμές είναι αυτά τα οποία ελαχιστοποιούν την συνάρτηση κόστους την οποία και θέλαμε να επιλύσουμε. Στο σημείο αυτό λοιπόν είμαστε σε θέση να προσδιορίσουμε τις τελικές διαστάσεις οι οποίες αντιπροσωπεύουν τα δεδομένα μας στον νέο χώρο μειωμένης διάστασης. Οι τελικές αυτές διαστάσεις d λαμβάνουν υπόψιν τους περιορισμούς οι οποίοι έχουν τεθεί και έτσι με τον τρόπο αυτό εξασφαλίζεται η διατήρηση των γεωμετρικών χαρακτηριστικών της κάθε γειτονιάς για όλα τα σημεία X_i του αρχικού συνόλου δεδομένων μεγέθους N.

Σημαντικό σημείο στην παραπάνω διαδικασία αποτελεί το γεγονός ότι δεν λαμβάνουμε υπόψιν την μικρότερη ιδιοτιμή της λύσης του παραπάνω συστήματος και αυτό διότι ισούτε με το μηδέν. Η ιδιοτιμή αυτή αντιπροσωπεύει το μοναδιαίο διάνυσμα το οποίο εξασφαλίζει ότι το σύνολο των δεδομένων έχει μηδενική μέση τιμή, εξασφαλίζοντας έτσι τον περιορισμό ως προς την αμεταβλητότητα κατά την μεταφορά.

Η παραπάνω διαδικασία απόρριψης την μηδενικής ιδιοτιμής και τελικά της επιλογής των τελικών διαστάσεων οι οποίες και αντιπροσωπεύουν τα αρχικά δεδομένα στον χώρο μειωμένης διάστασης υλοποιούνται με τον παρακάτω κώδικα MATLAB

```
options.disp = 0; options.isreal = 1; options.issym = 1;
[Y, eigenvals] = eigs(M,d+1,0,options);

Y = Y(:,end-1:-1:end-d)'*sqrt(N); % bottom evect is [1,1,1,1...] with eval 0
```

Βλέπουμε ότι η επίλυση της εξίσωσης γίνεται με την χρήση της συνάρτησης *eigs* του MATLAB η οποία επιλύει το πρόβλημα εύρεσης ιδιοτιμών και ιδιοδιανυσμάτων χρησιμοποιώντας τον αλγόριθμο Lanczos[7] ο οποίος βελτιστοποιεί σε μεγάλο βαθμό την επίλυση του προβλήματος, απο την στιγμή που ο τετραγωνικός πίνακας M του συστήματος είναι αραιός και θετικά ημιορισμένος.

Κεφάλαιο 5

Πειράματα

Στην εργασία αυτή πραγματοποιήθηκε μια σειρά πειραμάτων τα οποία έχουν ως στόχο τόσο την διερεύνηση του αλγορίθμου ως προς την απόδοσή του σε εφαρμογές αναγνώρισης προτύπων όσο και στο πώς επιδρούν στο αποτέλεσμα του αλγορίθμου οι δύο παράμετροι που δέχεται σαν είσοδο. Οι παράμετροι αυτοί είναι ο αριθμός k των κοντινών γειτόνων σύμφωνα με τους οποίους κατασκευάζεται ο πίνακας γειτνίασης του πρώτου βήματος του αλγορίθμου και η δεύτερη παράμετρος είναι ο αριθμός d των τελικών διαστάσεων και οι οποίες καθορίζουν ουσιαστικά το αποτέλεσμα του τελευταίου βήματος. Ο σχεδιασμός των πειραμάτων έγινε με τρόπο ώστε να γίνει φανερό το πως επηρεάζουν την συμπεριφορά του αλγορίθμου οι παράμετροι αυτοί αλλά επίσης δόθηκε έμφαση στο να βρεθεί ο βέλτιστος συνδυασμός των δύο ανάλογα με το σετ δεδομένων κάθε πειράματος.

Εφαρμόζονται επίσης δύο τεχνικές με τις οποίες μπορεί κάποιος να αποφύγει το τεράστιο υπολογιστικό κόστος που απαιτείται. Συγκεκριμένα το τελευταίο βήμα του αλγορίθμου το οποίο είναι και το πιο απαιτητικό έχει πολυπλοκότητα $\mathcal{O}n^3$ στην γενική περίπτωση ενώ στην συγκεκριμένη περίπτωση λόγω του αραιού μητρώου M είναι της τάξης $\mathcal{O}n^2$. Αντιλαμβανόμαστε λοιπόν ότι ακόμα και για ένα σχετικά μικρό σετ δεδομένων, για τα δεδομένα του κλάδου της μηχανικής μάθησης, το πρόβλημα που έχουμε να αντιμετωπίσουμε έχει απαγορευτικές διαστάσεις.

Ένα άλλο πρόβλημα που συναντάει κανείς κατά την εφαρμογή του αλγορίθμου σε κάποιο σετ δεδομένων είναι ο εξής περιορισμός. Ας υποθέσουμε ότι έχουμε ένα σύνολο δεδομένων μεγέθους N, απο τα οποία για κάποιον αριθμό $N1$ απο αυτά γνωρίζουμε την ετικέτα τους. Με τον όρο ετικέτα εννοούμε την τελική κλάση στην οποία ανήκει το κάθε δείγμα. Για τα υπόλοιπα δείγματα, έστω μεγέθους $N2$ δεν γνωρίζουμε την ετικέτα τους και είναι αυτά τα δείγματα για τα οποία θέλουμε να εξάγουμε το συμπέρασμα. Το συμπέρασμα αυτό είναι φυσικά η τελική απόφαση ως προς σε ποιά κλάση θα πρέπει να ταξινομηθεί το καθένα απο αυτά. Προφανώς η παραπάνω απόφαση προκύπτει λαμβάνοντας υπόψιν την πληροφορία την οποία μας δίνει το σύνολο των δεδομένων $N1$ τα οποία στον χώροτ της μηχανικής μάθησης αναφέρονται ως το σύνολο των δεδομένων εκπάιδευσης (train data). Τα υπόλοιπα δείγματα $N2$ αναφέρονται ως το σύνολο των δεδομένων αξιολόγησης (test data).

Στο συγκεκριμένο λοιπόν έστω οτι τα δείγματα του αρχικού χώρου έχουν αρχική διάσταση μεγέθους D και μέσω του αλγόριθμου μείωσης των διαστάσεων θέλουμε να βρεθούμε σε έναν νέο χώρο διάστασης d, προφανώς με $d < D$. Στην περίπτωση αυτή λοιπόν ο πιο απλός συλλογισμός που θα μπορούσε να κάνει κάποιος είναι να τρέξει τον αλγόριθμο LLE πάνω στο σετ δεδομένων εκπαίδευσης ώστε να έχει ένα σύνολο δεδομένων μεγέθους $N1$, διάστασης d. Με τον ίδιο ακριβώς τρόπο θα μπορούσε να έχει και το δεύτερο σετ δεδομένων, τα δεδομένα αξιολόγησης, μεγέθους $N2$ και αυτά διάστασης d. Έπειτα για την ταξινόμηση των αποτελεσμάτων θα μπορούσε να εφαρμοστεί ο αλγόριθμος ταξινόμησης κοντινότερων γειτόνων ανάμεσα στο σετ αξιολόγησης με το σετ εκπαίδευσης. Έπειτα ανάλογα με την κλάση στην οποία ανήκει ο κοντινότερος γείτονας απο το σετ εκπαίδευσης για κάθε ένα στοιχείο των δεδομένων αξιολόγησης θα καταλήγαμε στην τελική απόφαση για την κλάση στην οποία ανήκει κάθε ένα απο τα δεδομένα του σετ $N2$. Προφανώς ο αλγόριθμος κοντινότερων γειτόνων θα εφασρμοστεί στον χώρο μειωμένης διάστασης μεγέθους d χρησιμοποιώντας για παράδειγμα την μετρική της Ευκλείδιας απόστασης μεταξύ των σημείων.

Αν λοιπόν εφαρμόσουμε την παραπάνω διαδικασία για κάποιο σετ δεδομένων, θα παρατηρήσουμε ότι το τελικό αποτέλεσμα της ταξινόμησής μας έχει πολύ μικρή επιτυχία. Αυτό συμβαίνει διότι, οι δύο υποχώροι οι οποίοι προέκυψαν απο το τελικό βήμα του αλγορίθμου LLE, κατά το οποίο υπολογίστηκε ο νέος χώρος μειωμένης διάστασης για κάθε ένα απο τα δύο σύνολα δεδομένων, έχουν διαφορετική διανυσματική βάση και δεν μπορούν σε καμιά περίπτωση να συσχετιστούν μεταξύ τους ώστε να μπορέσουμε απο τα δεδομένα του ενός να καταλήξουμε σε κάποιο ορθό συμπέρασμα για τα δεδομένα του άλλου. Ο παραπάνω λοιπόν περιορισμός μας αναγκάζει να εφαρμόσουμε τον αλγόριθμο μείωσης των διαστάσεων στο σύνολο των δεδομένων, δηλαδή δίνοντας σαν είσοδο στον αλγόριθμο το σύνολο των δεδομένων μεγέθους $N = N1 + N2$. Με τον τρόπο αυτό θα καταλήγαμε σε ένα νέο σετ δεδομένων μεγέθους N αλλά διάστασης $d < D$. Τέλος σε αυτό το σετ δεδομένων τώρα μπορούμε κάλλιστα να εφαρμόσουμε τον αλγόριθμο εύρεσης κοντινότερων γειτόνων για κάθε ένα απο τα δεδομένα αξιολόγησης ως πρός τα δεδομένα εκπαίδευσης, φυσικά στον χώρο d διστάσεων, και έτσι να καταλήξουμε στην ορθή ταξινόμηση των δειγμάτων $N2$ ως προς την κλάση στην οποία ανήκουν.

5.1 Μέθοδοι αντιμετώπισης της πολυπλοκότητας του προβλήματος

Όπως αντιλαμβανόμαστε απο την παραπάνω ανάλυση, η διαδικασία αυτή δεν είναι καθόλου πρακτική και μάλιστα δεν δίνει την δυνατότητα για λήψη αποφάσεων και ταξινόμησης δειγμάτων σε πραγματικό χρόνο. Αυτό διότι, για κάθε δείγμα αξιολόγησης που μας έρχεται ως είσοδος κάποια συγκεκριμένη χρονική στιγμή, και για το οποίο θέλουμε να καταλήξουμε σε κάποιο συμπέρασμα ως προς την κλάση στην οποία ανήκει, θα πρέπει να το ενσωματώνουμε στο σετ των δεδομένων εκπαίδευσης και στην συνέχεια να εκτελούμε τον αλγόριθμο LLE. Αντιλαμβανόμαστε λοιπόν ότι η συγκεκριμένη διαδικασία δεν προσφέρεται σε καμιά περίπτωση για πρακτικές εφαρμογές κατα τις οποίες μάλιστα ο στόχος μας είναι να γίνει μείωση των διαστάσεων ώστε να μπορούμε να λαμβάνουμε ταχύτερα και ακριβέστερα αποτελέσματα. Το γεγονός αυτό μάλιστα αντιτίθεται στην συνολική φιλοσοφία της μείωσης των διαστάσεων κατά την οποία η μείωση των διαστάσεων μπορεί να επιταχύνει σε πολύ μεγάλο βαθμό τους απαραίτητους υπολογισμούς.

5.1.1 Μέθοδος-1: Προβολή στον χώρο των δεδομένων εκπαίδευσης

5.1.2 Μέθοδος-2: Δημιουργία υποχώρων και πλειοψηφική απόφαση ταξινόμησης

5.2 Σχεδιασμός και οργάνωση των πειραμάτων

5.2.1 Σετ δεδομένων

5.2.2 Πειράματα

5.3 Αποτελέσματα

Κεφάλαιο 6

Συμπεράσματα

Bibliography

[1] Xin Liu, Duygu Tosun, Michael W. Weiner, and Norbert Schuff. Locally linear embedding (lle) for mri based alzheimer's disease classification. 2013.

[2] Hualei Shen, Dacheng Tao, and Dianfu Ma. Multiview locally linear embedding for effective medical image retrieval. 2013.

[3] L.J.P. van der Maaten, E. O. Postma, and H. J. van den Herik. Dimensionality reduction: A comparative review. 2008.

[4] Olga Kouropteva, Oleg Okun, and Matti Pietikäinen. Supervised locally linear embedding algorithm for pattern recognition. 2003.

[5] Dick de Ridder and Robert P.W. Duin. Locally linear embedding for classification. 2002.

[6] Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. 2000.

[7] Jane K. Cullum and Ralph A. Willoughby. *Lanczos Algorithms for Large Symmetric Eigenvalue Computations, Vol. 1.* 2002.