

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 από 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 ότι ο ανθρώπινος εγκέφαλος είναι ένα τρομερά περίπλοκο σύστημα εκατομμυρίων νευρώνων συνδεδεμένων μεταξύ τους οι οποίοι είναι σε θέση να εκτελούν 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 ανάλογα με αυτά ενός ανθρώπου.

Χαρακτηριστικά παραδείγματα τέτοιων εφαρμογών με τα οποία καταπιάνεται και η εργασία αυτή είναι η χαρακτηριστικών 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 δύο σετ δεδομένων με εικόνες ψηφία αριθμών και σε ένα με ιατρικά δεδομένα απο καρκινοπαθείς και μη ασθενείς. Αφού γίνει μείωση των διαστάσεων που πρέπει να ληφθούν υπόψιν για την εξαγωγή συμπεράσματος εφαρμόζεται ο ταξινομητής κοντινότερων γειτόνων 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 κ.Νικόλαο Πιτσιάνη επίκουρο καθηγητή του τμήματός μου ο οποίος μου έδωσε το ερέθισμα καθώς και 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, τους φίλους και την κοπέλα μου διότι ο καθένας με τον τρόπο του βοήθησαν στην αντιμετώπιση των δυσκολιών που συνάντησα καθ'όλη 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

1	Εισαγωγή	13
1.1	Αναγνώριση προτύπων και μηχανική μάθηση	13
1.2	Ερεθίσματα απο τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου	14
1.2.1	Μάθηση με επίβλεψη - χωρίς επίβλεψη - με ημιεπίβλεψη	15
1.3	Μείωση της διάστασης των δεδομένων	17
2	Μαθηματικό και θεωρητικό υπόβαθρο	19
2.1	Διανύσματα βάσης	19
2.1.1	Διάνυσμα εικόνας	20
2.1.2	Ορθοκανονικά ιδιοδιανύσματα	20
2.2	Ο μετασχηματισμός Karhunen-Loeve - PCA	22
2.2.1	Προσέγγιση μέσου τετραγωνικού σφάλματος - MSE	23
2.2.2	Συνολική Διασπορά	25
2.2.3	Μείωση της διάστασης μέσω PCA	25
2.3	Μετρική πολυδιάσττης κλιμάκωσης (Metric multidimensional scaling - MDS) . . .	26
2.4	Ανάλυση στην βάση των ιδιζουσών τιμών (SVD)	27
2.4.1	Μείωση της διάστασης μέσω SVD	28

2.5	Πρακτική εφαρμογή	30
3	Αλγόριθμοι μείωσης διαστάσεων	32
3.1	Γραμμική μείωση διαστάσεων	32
3.2	Μη γραμμική μείωση διαστάσεων	33
3.2.1	ISOMAP	36
3.2.2	Laplacian Eigenmaps	38
3.2.3	LLE	38
4	Ανάλυση του αλγορίθμου Locally Linear Embeddings	39
4.1	Βήμα-1: Εύρεση του πίνακα γειτνίασης	39
4.2	Βήμα-2: Κατασκευή του Laplacian	39
4.3	Βήμα-3: Επίλυση του προβλήματος εύρεσης ιδιτιμών και ιδιοδιανυσμάτων	39
4.4	Βήμα-4: Επιλογή των τελικών διαστάσεων	39
5	Πειράματα	40
5.1	Σετ δεδομένων	40
5.2	Σχεδιασμός και οργάνωση των πειραμάτων	40
5.3	Αποτελέσματα	40
6	Συμπεράσματα	41
	Βιβλιογραφία	41

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

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	30
3.1	Swiss Roll Synthetic Dataset.	34
3.2	Swiss Roll Synthetic Dataset Manifold Learning Path.	35
3.3	Dimensionality Reduction with LLE - 3D (K=16,d=2).	36

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 σκετ δεδομένων χρειάζεται για κάποιο λόγο να ταξινομηθεί. Στις μέρες μας η ανάγκη διαχείρισης αλλά και ανάκτησης πληροφοριών μέσω ηλεκτρονικών υπολογιστών αποκτά τεράστια σπουδαιότητα καταρχήν διότι ο όγκος των πληροφοριών αυξάνεται ραγδαία με ρυθμό αδύνατο να διαχειριστεί ο άνθρωπος και επίσης διότι η ανάπτυξη της τεχνολογίας μας παρέχει πολύ ισχυρά υπολογιστικά συστήματα με τη χρήση των οποίων μπορούμε να δημιουργήσουμε πολύπλοκα μοντέλα εξόρυξης γνώσης .

Αντίστοιχοι κλάδοι στους οποίους έχει τεράστια σημασία η αναγνώριση προτύπων είναι οι επιστημονικοί κλάδοι 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 αλυσίδας του γενετικού 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. Αυτό 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 παράδειγμα 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 αναγνωρίσει 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 ότι 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 ότι το μοτίβο του ψηφίου 1 είναι πολύ κοντά σε αυτά των ψηφίων επτά και τέσσερα αλλά σε καμιά περίπτωση δεν θα λέγαμε ότι 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 ότι απαιτούνται πολύ πιο σύνθετοι υπολογισμοί και θα πρέπει να συνδιάσουμε ένα πλήθος από 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 ότι 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 αλγόριθμους με σκοπό να εφαρμοστούν σε μοντέλα εξόρυξης γνώσης.

1.2.1 Μάθηση με επίβλεψη - χωρίς επίβλεψη - με ημιεπίβλεψη

Ένα πολύ εύλογο ερώτημα το οποίο προκύπτει από 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 από 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 όποτε ξαναεμφανιστούν μπροστά του. Τα ερεθίσματα αυτά είναι είτε εικόνες, είτε ήχοι είτε ερεθίσματα τα οποία μπορεί να προέρχονται από 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 συνδυασμός τεχνικών μάθησης και συνεχούς εκπαίδευσης. Οι τεχνικές αυτές στον χώρο της τεχνητής νοημοσύνης αναφέρονται ως τεχνικές μάθησης με επίβλεψη, χωρίς επίβλεψη και με ημι-επίβλεψη. Θα μπορούσε κάποιος αρχικά να υποστηρίξει ότι ο ανθρώπινος εγκέφαλος 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 από την πολύ νεαρή του ηλικία 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 μεταξύ του. Παρόλα αυτά μετά από κάποιο σημείο ο ανθρώπινος εγκέφαλος αποκτά δυνατότητες με τις οποίες μπορεί αξιολογεί και να μαθαίνει μόνος του πολύ σύνθετα πράγματα αναλύοντάς τα σε απλούστερα προβλήματα 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.

Από 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 από 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. Παρόλα είναι πολύ δύσκολο να εφαρμοστεί στον τομέα της τεχνητής νοημοσύνης και αυτό διότι ακόμα δεν είμαστε σε θέση να δώσουμε εξηγήσεις για τον 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 οδηγήσουμε να μάθει συγκεκριμένα μοντέλα τα οποία να μπορεί να χρησιμοποιήσει στην συνέχεια με σκοπό της 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 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 προβλημάτων.

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 εμφανίζουν 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 απομάχρυνσης 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 $N \times 1$,

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

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

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

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

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

Οι στήλες του A , $\mathbf{a}_i = 0, 1, \dots, N-1$ 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:

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

2.1.1 Διάνυσμα εικόνας

Αν πάρουμε για παράδειγμα μια εικόνα, το σύνολο των δειγμάτων εισόδου είναι μια δυδιάστατη ακολουθία $X(i, j), i, j = 0, 1, \dots, N-1$, 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

$$\mathbf{x}^T = \left[X(0, 0), \dots, X(0, N-1), \dots, X(N-1, 0), \dots, X(N-1, N-1) \right] \quad (2.1.4)$$

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

$$Y = U^H X V \quad (2.1.5)$$

ή

$$X = U Y V^H \quad (2.1.6)$$

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

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

όπου \mathbf{u}_i είναι τα διανύσματα στήλης του Y και ν_j τα διανύσματα στήλης του * . Η παραπάνω εξίσωση είναι ένα ανάπτυγμα του μητρώου Ξ ως προς τις $N \times 2$ εικόνες βάσης. Τέλος κάθε ένα απο τα γινόμενα $\mathbf{u}_i \nu_j$ είναι ένα μητρώο $N \times N$

$$\mathbf{u}_i \nu_j = \begin{bmatrix} u_{i0} \nu_{j0}^* & \dots & u_{i0} \nu_{jN-1}^* \\ \vdots & \vdots & \vdots \\ u_{iN-1} \nu_{j0}^* & \dots & u_{iN-1} \nu_{jN-1}^* \end{bmatrix} \quad (2.1.8)$$

Στην περίπτωση κατα την οποία το Ψ είναι διαγώνιο τότε έχουμε

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

με αποτέλεσμα το πλήθος των μητρώων-εικόνων βάσης 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

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

και τελικά

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

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 λεξιλογιαφικά όπως περιγράφηκε 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

$$E[y(i)y(j)] = 0, i \neq j \quad (2.2.1)$$

Έστω

$$\mathbf{y} = A^T \mathbf{x} \quad (2.2.2)$$

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

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

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 $\mathbf{a}_i, i = 0, 1, \dots, N - 1$ του R_x τότε το R_y είναι διαγώνιο.

$$R_y = A^T R_x A = \Lambda \quad (2.2.4)$$

Το Λ είναι διαγώνιο μητρώο με διαγώνια στοιχεία τις αντίστοιχες ιδιοτιμές $\lambda_i, i = 0, 1, \dots, N - 1$ 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 τετραγωνικού σφάλματος. Ας πάρουμε ξανά τις εξισώσεις (2.1.1) και (2.1.2) τότε έχουμε

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

Ορίζουμε λοιπόν τώρα ένα νέο διάνυσμα στον m -διάστατο υποχώρο

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

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 $\hat{\mathbf{x}}$ και θα προκύψει χρησιμοποιώντας την εξίσωση ελαχιστοποίησης μέσου τετραγωνικού σφάλματος. Έχουμε λοιπόν την εξίσωση

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

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

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

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

και λαμβάνοντας υπόψιν τον ορισμό των ιδιοδιανυσμάτων προκύπτει τελικά ότι

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

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} από 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 κατά **KL** διάνυσμα του \mathbf{x} και $E[x] = 0$. Τότε από τον αντίστοιχο ορισμό της διασποράς έχουμε ότι $\sigma_{y(i)}^2 \equiv E[y^2(i)] = \lambda_i$. Δηλαδή έχουμε ότι οι διασπορές του μητρώου συσχέτισης εισόδου είναι ίσες με τις διασπορές των μετασχηματισμένων χαρακτηριστικών. Επομένως επιλεγώντας 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 μεθοδολογία που ακολουθήσαμε, τα m χαρακτηριστικά που έχουν επιλεχθεί διατηρούν το μεγαλύτερο μέρος από την συνολική διασπορά που σχετίζεται με τις αρχικές τυχαίες μεταβλητές $x(i)$.

2.2.3 Μείωση της διάστασης μέσω PCA

Από 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 από την εξίσωση (2.2.10) συνεπάγεται ότι $\mathbf{x} = \hat{\mathbf{x}}$. 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 $\hat{\mathbf{x}}$ με πολύ καλή προσέγγιση. Το γεγονός αυτό εισάγει την έννοια της εγγενούς διάστασης (intrinsic dimensionality). Τέλος στην περίπτωση της εγγενούς διάστασης μπορούμε να πούμε ότι το X μπορεί να περιγραφεί από m ελεύθερες παραμέτρους.

2.3 Μετρική πολυδιάστατης κλιμάκωσης (Metric multidimensional scaling - MDS)

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

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

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 από την ανάλυση σε ιδιοτιμές-ιδιοδιανύσματα του μητρώου Gram, τα στοιχεία του οποίου ορίζονται ως

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

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

λύσεις που προκύπτουν από 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 από τη σχέση

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

όπου

$$Q^T = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \quad (2.3.4)$$

Από την άλλη τώρα το μητρώο Gram μπορεί επίσης να γραφεί ως

$$K = X X^T \quad (2.3.5)$$

Τέλος αποδεικνύεται ότι τα δύο μητρώα $X^T X$ και $X X^T$ είναι ίδιου βαθμού και έχουν τις ίδιες ιδιοτιμές με ιδιοδιανύσματα τα οποία να μεν είναι διαφορετικά μεταξύ τους αλλά παρόλα αυτά σχετίζονται.

2.4 Ανάλυση στην βάση των ιδιζουσών τιμών (SVD)

Η ανάλυση ενός μητρώου με βάση τις ιδιζουσες τιμές είναι μια από 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

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

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, ν_i είναι τα διανύσματα στήλης των μητρώων U και V αντίστοιχα τότε η παραπάνω εξίσωση μπορεί να γραφεί στην μορφή

$$X = [u_0, u_1, \dots, u_{r-1},] \begin{bmatrix} \sqrt{\lambda_0} & & & \\ & \sqrt{\lambda_1} & & \\ & & \ddots & \\ & & & \sqrt{\lambda_{r-1}} \end{bmatrix} \begin{bmatrix} \nu_0^H \\ \nu_1^H \\ \vdots \\ \nu_{r-1}^H \end{bmatrix} \quad (2.4.2)$$

ή

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

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, ν_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* (σινγκυλαρ αλυες) 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 από 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 από τον τύπο

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

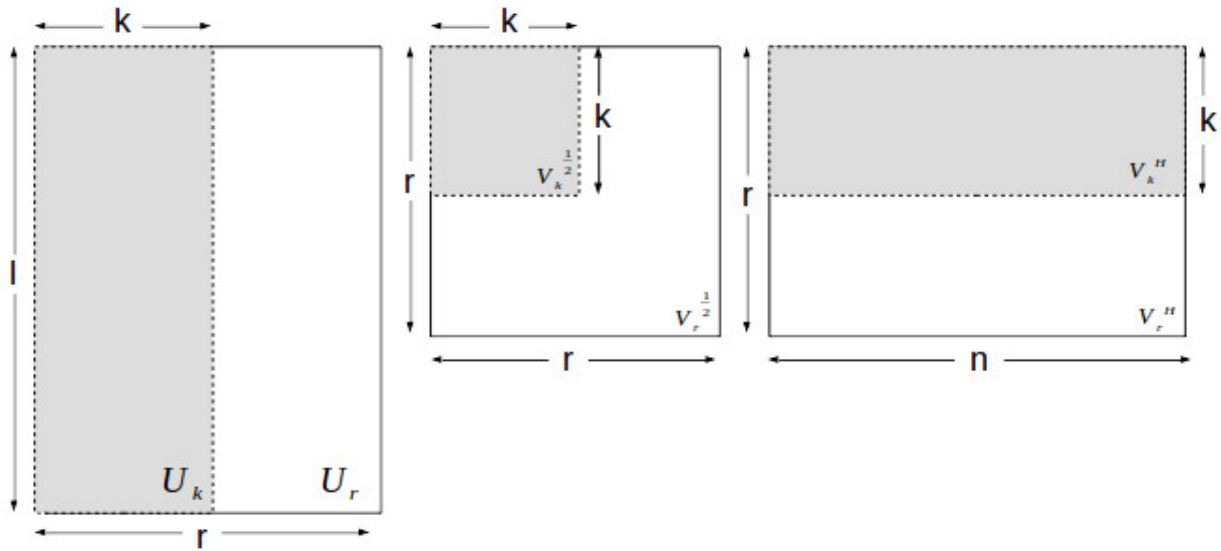
$$= [\mathbf{u}_0, \mathbf{u}_1, \dots, \mathbf{u}_{k-1}] \begin{bmatrix} \sqrt{\lambda_0} \nu_0^H \\ \sqrt{\lambda_1} \nu_1^H \\ \vdots \\ \sqrt{\lambda_{k-1}} \nu_{k-1}^H \end{bmatrix} = U_k [\mathbf{a}_0, \mathbf{a}_1, \dots, \mathbf{a}_{n-1},] \quad (2.4.4)$$

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

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

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

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



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

Αντιλαμβανόμαστε λοιπόν ότι χρησιμοποιώντας την προηγούμενη προβολή και υποθέτοντας ότι η προσέγγιση είναι ικανοποιητική, η Ευκλείδεια 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 από ένα l -διάστατο διάνυσμα χαρακτηριστικών. Τότε, δοθέντως ενός άγνωστου προτύπου στόχος μας είναι να αναζητήσουμε στο σύνολο των γνωστών προτύπων που έχουμε ώστε να 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 Ευκλείδεια 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 Ευκλείδειων αποστάσεων στον k -διάστατο χώρο. Επειδή οι Ευκλείδειες 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 από τις γνωστότερες αλλά και πιο ισχυρές μεθόδους γραμμικής μείωσης διαστάσεων λειτουργεί καλά αν τα σημεία των δεδομένων είναι κατανομημένα 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 υπεύθυνες 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 κατά τις οποίες ο μηχανισμός παραγωγής των δεδομένων είναι έντονα μη γραμμικός με αποτέλεσμα τα δεδομένα να κείτονται σε πιο περίπλοκες πολλαπλότητες. Ας πάρουμε για πα-

ράδειγμα τις εξισώσεις

$$x_1 = r \cos \theta, \quad x_2 = r \sin \theta$$

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.

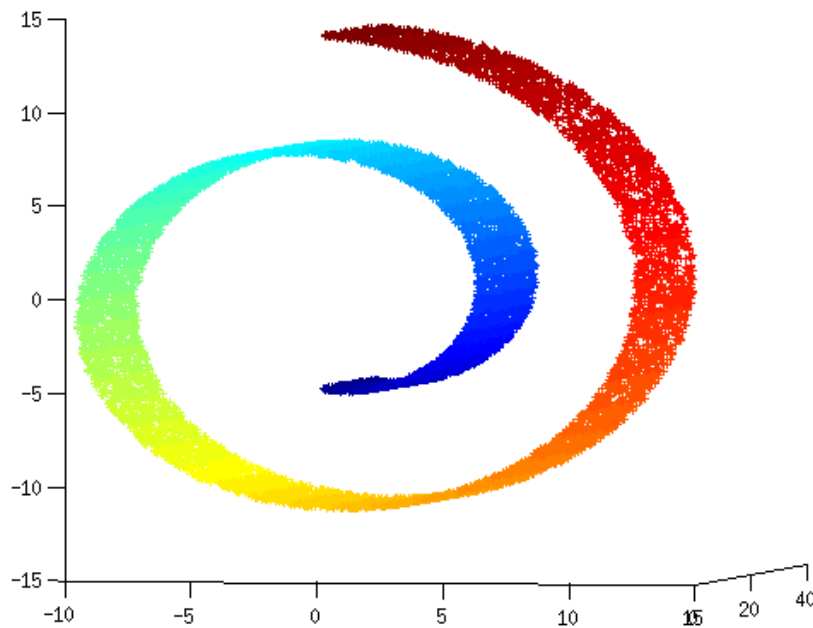
Περιπτώσεις όπως οι παραπάνω απαιτούν αλγορίθμους μείωσης διάστασης και εξαγωγής χαρακτηριστικών οι οποίοι να λαμβάνουν υπόψιν την γεωμετρία του προβλήματος ώστε να μπορούν να εξάγουν αποτελεσματικά συμπεράσματα για την διάσταση των δεδομένων. Στον τομέα της υπολογιστικής όρασης για παράδειγμα, ο οποίος όπως αναφέραμε και παραπάνω αποτελεί βασικό κομμάτι της εν λόγω διατριβής, απαιτούνται κατεξοχήν αλγόριθμοι μη γραμμικής μείωσης διαστάσεων αφού οι εικόνες ή τα χαρακτηριστικά 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 ενσωμάτωση μικρότερης διάστασης η οποία βρίσκεται 'κρυμμένη' στον χώρο των αρχικών διαστάσεων. Θα πρέπει μάλιστα κατά την διαδικασία αυτή να ληφθούν προφανώς υπόψιν τα γεωμετρικά χαρακτηριστικά του χώρου των δεδομένων.

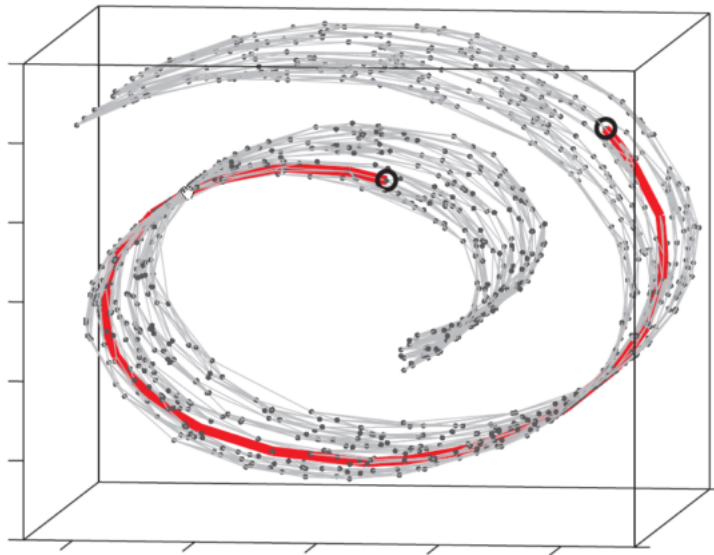
Έχει πολύ μεγάλη σημασία στο σημείο αυτό να κατανοήσουμε τι εννοούμε όταν αναφερόμαστε στα γεωμετρικά χαρακτηριστικά του προβλήματος. Το πιο χαρακτηριστικό και ευρέως χρησιμοποιο-

ύμενο παράδειγμα για τον σκοπό αυτό είναι ένα τεχνητό σετ δεδομένων, με την ονομασία Swiss Roll το οποίο φαίνεται στην παρακάτω εικόνα.



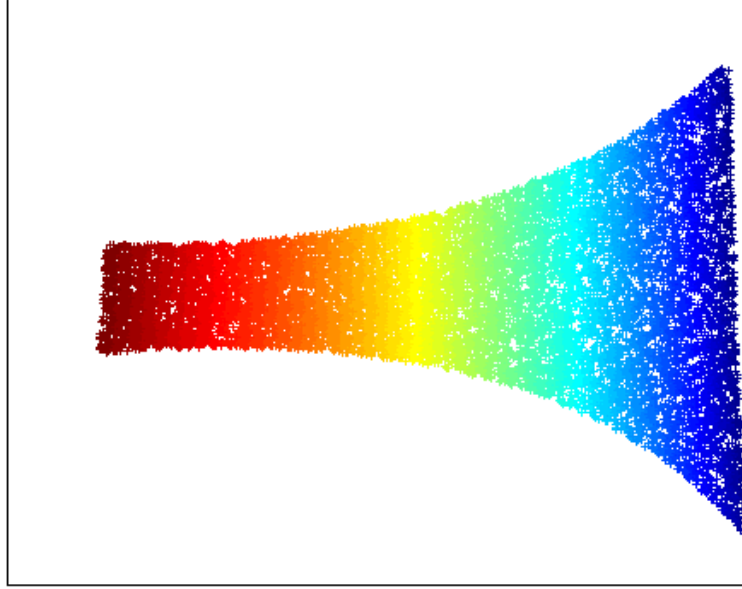
Σχήμα 3.1: Swiss Roll Synthetic Dataset.

Αυτό που αξίζει να παρατηρηθεί λοιπόν στο παραπάνω σετ δεδομένων είναι ότι αν για παράδειγμα διαλέξουμε κάποιο οποιοδήποτε σημείο του από 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 από την μπλέ περιοχή βρίσκονται και αυτά στην γειτονιά του σημείου που διαλέξαμε. Αυτό προφανώς είναι λάθος αφού από τον χρωματισμό των παραπάνω δεδομένων αντιλαμβανόμαστε ότι στην πραγματικότητα τα μπλέ δεδομένα βρίσκονται πολύ μακριά από τα κόκκινα. Ο παραπάνω εσφαλμένος συλλογισμός αναπαρίσται 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 από 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 Ευκλείδειες 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 Ευκλείδεια απόστασή τους.

Ουσιαστικά η μέθοδος ISOMAP είναι μια παραλλαγή του αλγορίθμου Multi Dimensional Scaling - MDS, με την διαφορά ότι οι Ευκλείδειες αποστάσεις αντικαθίστανται από 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 $x_i, i = 1, 1 \dots, n$, υπολόγισε τους πλησιέστερους γείτονες και κατασκεύασε 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 από τον χρήστη). Στις ακμές ανατίθενται βάρη σύμφωνα με τις αντίστοιχες Ευκλείδειες 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 κατά 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 σημεία στο γράφο $G(V, E)$. Ο πιο γνωστός αλγόριθμος υλοποίησης της παραπάνω διαδικασίας είναι ο αλγόριθμος Dijkstra 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 να εφαρμόσουμε την κλασική μέθοδο MDS. Το πρόβλημα λοιπόν από 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 από αυτή την αναπαράσταση, οι Ευκλείδειες 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 από το πλήθος των m περισσότερο σημαντικών ιδιοτιμών. Αποδεικνύεται τέλος ότι η μέθοδος ISO-MAP ασυμπτωτικά ($n \rightarrow \infty$) θα ανακτήσει την αληθινή διάσταση για ένα σύνολο δεδομένων μη γραμμικής πολλαπλότητας.

3.2.2 Laplassian Eigenmaps

3.2.3 LLE

Κεφάλαιο 4

Ανάλυση του αλγορίθμου Locally Linear Embeddings

- 4.1 Βήμα-1: Εύρεση του πίνακα γειτνίασης
- 4.2 Βήμα-2: Κατασκευή του Laplacian
- 4.3 Βήμα-3: Επίλυση του προβλήματος εύρεσης ιδιτιμών και ιδιοδιανυσμάτων
- 4.4 Βήμα-4: Επιλογή των τελικών διαστάσεων

Κεφάλαιο 5

Πειράματα

5.1 Σετ δεδομένων

5.2 Σχεδιασμός και οργάνωση των πειραμάτων

5.3 Αποτελέσματα

Κεφάλαιο 6

Συμπεράσματα