ΠΕΤΡΑΚΗΣ ΚΩΝΣΤΑΝΤΙΝΟΣ ΑΜ:1041589 ΥΔΑ

Unsupervised Dimensionality Reduction: Overview and Recent Advances

JOHN A. LEE AND MICHAEL VERLEYSEN

Τι περιλαμβάνει η εργασία

Συνοπτική περιγραφή των κυριότερων μεθόδων

Προτείνεται μια ταξινόμηση αυτών με βάση διάφορα κριτήρια

Με βάση ποια κριτήρια μπορεί να γίνει η αξιολόγηση των μεθόδων?

Ορισμός του Προβλήματος

Curse of dimensionality: όλες εκείνες οι ανεπιθύμητες ιδιότητες των χώρων υψηλής διαστατικότητας (έννοιες όπως κοντά ή μακριά χάνουν το νόημα τους)

Η βασική ιδέα του πεδίου του DR είναι πως εάν η αναπαράσταση των δεδομένων είναι δύσκολη σε χώρους πολλών διαστάσεων, ίσως μια (σχεδόν) ισοδύναμη αναπαράσταση αυτών σε ένα χώρο λιγότερων διαστάσεων θα μπορούσε να βελτιώσει την αναγνωσιμότητα των δεδομένων.

Η αναπαράσταση στο χώρο χαμηλότερης διαστατικότητας θα πρέπει να έχει κάποιο νόημα. Απώτερος στόχος του DR είναι η αναπαράσταση όμοιων αντικειμένων κοντά το ένα με το άλλο, διατηρώντας ταυτόχρονα μεγάλες αποστάσεις για αυτά που είναι ανόμοια.

Στην πράξη προσπαθούμε να διατηρήσουμε κάποιες ιδιότητες των δεδομένων όπως:

- Ανά ζεύγη αποστάσεις (pairwise distances)
- Ομοιότητες
- Ranks
- Hard ή soft γειτονιές

Εφαρμογές

Η μείωση της διαστατικότητας χρησιμοποιείται κυρίως κατά τη φάση προ-επεξεργασίας σαν ένα βήμα ‘συμπίεσης’ των δεδομένων ή απομάκρυνσης τυχόν θορύβου που αυτά περιέχουν, με την ελπίδα πως η απλοποιημένη αναπαράσταση που θα προκύψει θα επιταχύνει οποιαδήποτε μεταγενέστερη επεξεργασία (π.χ. κάποιος αλγόριθμος Μηχανικής Μάθησης) ή θα βελτιώσει τα αποτελέσματα της .

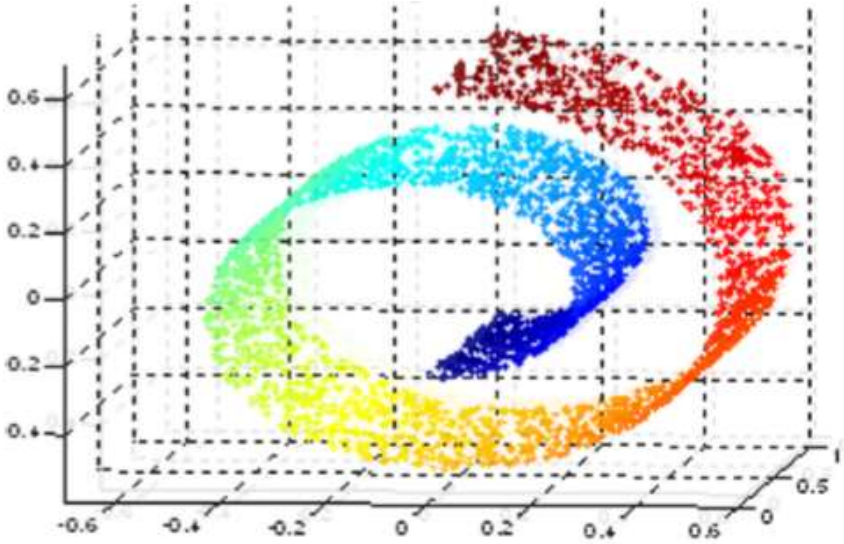
Επίσης πολύ σημαντική είναι και η χρήση της για την οπτικοποίηση των δεδομένων.

Εισαγωγή (LDR-NLDR)

Linear DR: γίνεται η υπόθεση ότι τα δεδομένα είναι κατανεμημένα εντός ενός (σχεδόν) γραμμικού υπόχωρου

NLDR (manifold learning): θεωρούμε ότι τα δεδομένα δειγματοληπτούνται από ένα smooth manifold. Στόχος είναι να ενσωματωθεί (re-embed) το manifold σε ένα χώρο όσο το δυνατόν πιο χαμηλής διαστατικότητας.

Οι διάφορες μέθοδοι διαφέρουν ως προς τις ιδιότητες των δεδομένων ή του manifold που προσπαθούν να διατηρήσουν.



Principal Component Analysis

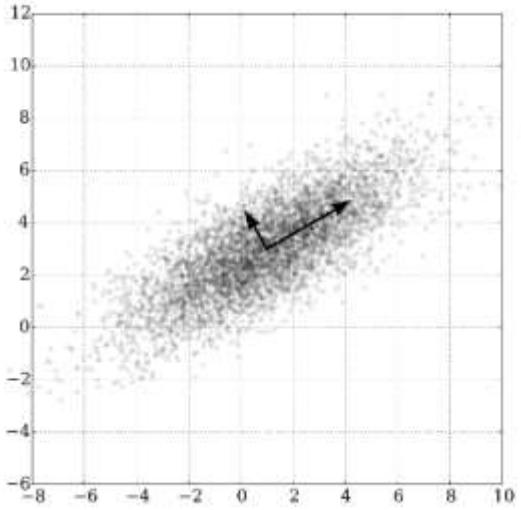
$\Xi = [\xi_i]_{1 \le i \le N}$: δεδομένα στον αρχικό χώρο

$X = [x_i]_{1 \le i \le N}$: δεδομένα στον χώρο χαμηλότερης διάστασης (ελεύθεροι παράμετροι)

PCA: Αναζήτηση των κατευθύνσεων (principal components) ως προς τις οποίες τα δεδομένα παρουσιάζουν την μεγαλύτερη διασπορά.

- Αναζήτηση για το ολικό ελάχιστο της $\min_X \|C_{\Xi\Xi} - C_{XX}\|_2^2$ όπου $C_{\Xi\Xi}$ και C_{XX} τα μητρώα συνδιασποράς των δεδομένων στον αρχικό και στον νέο χώρο αντίστοιχα.
- Λύση αποτελούν τα κυρίαρχα ιδιοδιανύσματα του $C_{\Xi\Xi}$.

‘Crowding’ effect



Multidimensional Scaling

Στόχος η διατήρηση των ανά ζεύγη αποστάσεων από τον αρχικό στον τελικό χώρο.

- $\delta_{ij} = \|\xi_i - \xi_j\|_2$, $d_{ij} = \|x_i - x_j\|_2$

Classical MDS : Δοθέντος του μητρώου των τετραγωνικών ευκλείδειων αποστάσεων $\Delta=[\delta_{ij}]_{1\leq i,j\leq N}$ υπολογίζεται, με double centering, το Gram μητρώο $\mathrm{G}_{\Xi\Xi} = -\frac{1}{2}\mathrm{H}\Delta\mathrm{H}$, όπου $H = I - \frac{1}{N}1^T 1$. Οι συντεταγμένες των σημείων στο χώρο χαμηλότερης διάστασης m είναι $X = E_m \Lambda_m^{1/2}$

- Λύση (με φασματική παραγοντοποίηση) της $\min_X \|\mathrm{G}_{\Xi\Xi} - G_{\mathrm{XX}}\|_2^2$ (strain function)

Metric MDS : Γενίκευση του classical MDS.

- Εφαρμογή αλγορίθμων βελτιστοποίησης για την λύση της $\min_X \sum_{i<j} w_{ij}(\delta_{ij} - d_{ij})^2$ (stress function)

Επεκτάσεις MDS

Sammon NLM: Χρησιμοποιεί τη ‘stress function’ με $w_{ij} = \frac{1}{\delta_{ij}}$.

- Δίνει περισσότερο βάρος στις κοντινές αποστάσεις (ο παρονομαστής μειώνει την εστίαση σε μεγάλες αποστάσεις)
- Ανέχεται intrusions

Curvilinear Component Analysis: όμοια, αλλά με $w_{ij} = f(d_{ij}/\sigma)$, όπου σ πλάτος της γειτονιάς και f φθίνουσα.

- Εδώ τα βάρη εξαρτώνται από τις αποστάσεις στο χώρο χαμηλότερης διάστασης
- Έχει την ικανότητα να ‘σχίζει’ τα manifolds και να ξεδιπλώνει βρόγχους σε αυτά.
- Επιτρέπει extrusions

Στο NLDR συμβαίνουν 2 ειδών λάθη:

- Tearing errors (extrusions): κοντινά σημεία αναπαρίστανται μακριά
- Flattening errors (intrusions): απομακρυσμένα σημεία γίνονται εσφαλμένα κοντινοί γείτονες

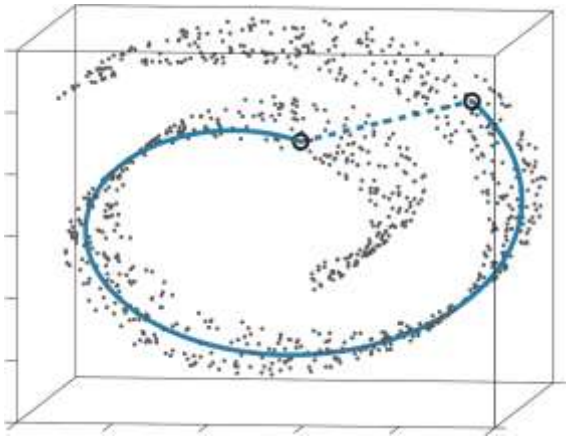
Isomap

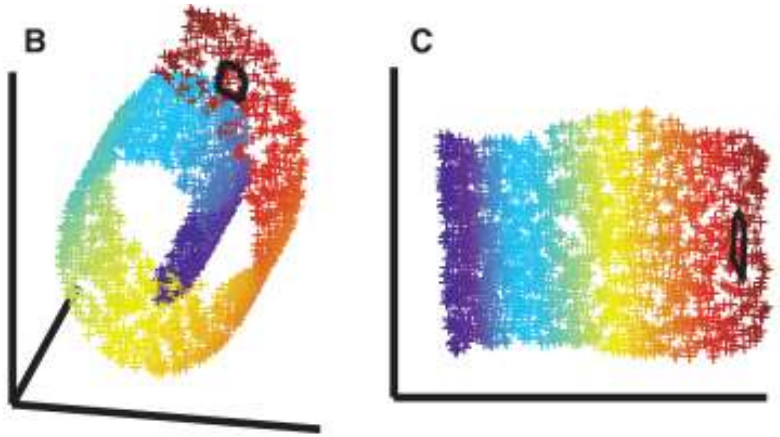
Isomap: Συνίσταται στην εφαρμογή classical metric MDS στο μητρώο που περιέχει τις ανά ζεύγη geodesic αποστάσεις.

Οι geodesic distances προσεγγίζονται από τα μήκη των συντομότερων μονοπατιών (shortest paths) μεταξύ των σημείων στον Ευκλείδειο γράφο των δεδομένων

- K-ary neighborhood graph
- e-balls graph

Βρίσκει ακριβή λύση μέσω φασματικής παραγοντοποίησης.





Διατήρηση των αποστάσεων σε χώρους χαρακτηριστικών

Οι παρακάτω μέθοδοι μπορούμε να πούμε ότι ενεργούν σε χώρους χαρακτηριστικών.

Kernel PCA: Αντιστοίχιση των δεδομένων σε ένα νέο χώρο χαρακτηριστικών και εφαρμογή PCA σε αυτό το νέο χώρο.

- Δεν απαιτείται κανένας υπολογισμός στο χώρο χαρακτηριστικών με την εφαρμογή του 'kernel trick' $k(\xi_i, \xi_j) = \langle \Phi(\xi_i), \Phi(\xi_j) \rangle$ στον classical metric MDS. Το Gram μητρώο κατασκευάζεται εφαρμόζοντας τον kernel k σε ζεύγη διανυσμάτων του dataset Ξ.
- η προβολή σε ένα χώρο υψηλότερης διάσταση μπορεί να απλοποιήσει μη-γραμμικά διαχωρίσιμα δεδομένα

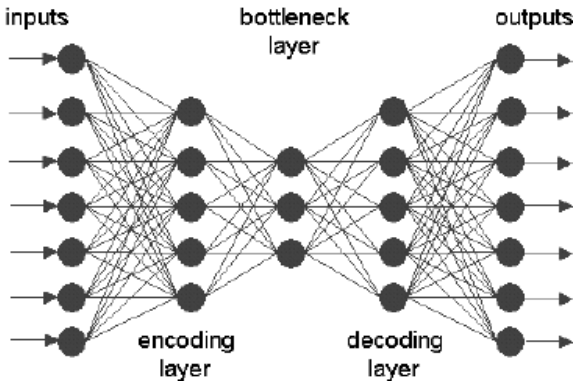
Laplacian eigenmaps: Οι συντεταγμένες X του χώρου χαμηλότερης διάστασης δίνονται από τα ιδιοδιανύσματα του Laplacian μητρώου L=D-W, που αντιστοιχούν στις μικρότερες ιδιοτιμές.

- ισοδυναμεί με την εφαρμογή του αλγορίθμου MDS σε commute time distances

Νευρωνικά Δίκτυα

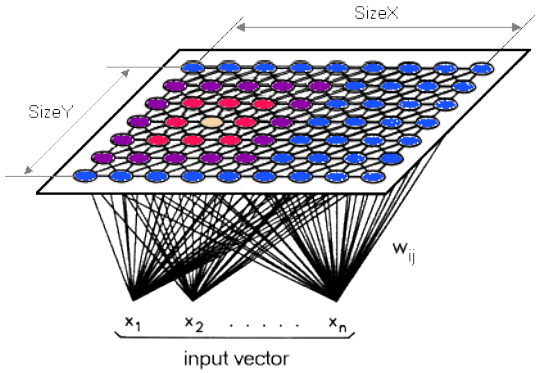
Auto-associative ΤΝΔ μπορούν να χρησιμοποιηθούν για NLDR.

- Η χρήση γραμμικών συναρτήσεων ενεργοποίησης σε όλους τους νευρώνες έχει τα ίδια αποτελέσματα με την PCA. (εκτελείται TLS regression προσπαθώντας να ταιριάξουν έναν γραμμικό υπόχωρο δια μέσω των δεδομένων.)



SOM: Αυτό-οργανούμενοι χάρτες χαρακτηριστικών Kohonen

- $\gamma_k \leftarrow \gamma_k + \alpha \mathrm{K}_\sigma(d(g_k, g_l))(\xi_i - \gamma_k)$,
- $l = \underset{k}{\operatorname{argmin}} \|\xi_i - \gamma_k\|_2$, d: απόσταση στο πλέγμα, K: Gaussian-Mexican hat



Διατήρηση ομοιότητας και SNE

Η διατήρηση ομοιότητας είναι ίσως καλύτερη από την διατήρηση των αποστάσεων. Η ομοιότητα ορίζεται συνήθως σαν φθίνουσα συνάρτηση της απόστασης.

Η χρήση της ομοιότητας δίνει περισσότερο βάρος στην διατήρηση τοπικών ιδιοτήτων (K-ary neighborhoods) από ότι global ιδιοτήτων.

Κανονικοποιημένες ανά ζεύγη ομοιότητες (pdf): $\pi_{ij} = \frac{\gamma(\delta_{ij}^2)}{\sum_{k<l}\gamma(\delta_{kl}^2)}$ και $p_{ij} = \frac{g(d_{ij}^2)}{\sum_{k<l} g(d_{kl}^2)}$, όπου γ και g φθίνουσες συναρτήσεις.

- KL divergence: $D(X;\Xi) = \sum_{i<j}\pi_{ij}\log(\pi_{ij}/p_{ij})$ (πόσο απέχουν οι 2 κατανομές)
- Είναι μέτρο σχετικής εντροπίας (αν είναι 0 οι κατανομές είναι ταυτόσημες)

t-SNE: εστιάζει στις αποστάσεις μεταξύ γειτονικών σημείων και επιτρέπει μεγάλες διακυμάνσεις στις μεγάλες αποστάσεις.

Κατηγοριοποίηση των μεθόδων

Linear DR vs NLDR
- Linear DR: PCA, classical metric MDS

Με βάση τις ιδιότητες που προσπαθούν να διατηρήσουν
- διατήρηση αποστάσεων
- διατήρηση εσωτερικών γινομένων
- διατήρηση ομοιότητας
- διατήρηση rank

Το DR ανέρχεται σε πρόβλημα TLS (άρα χρειάζεται βελτιστοποίηση).Με βάση τη βελτιστοποίηση που χρησιμοποιείται
- spectral methods: PCA,MDS
  - Dense: KPCA, Isomap (χρησιμοποιοπούν τα κυρίαρχα ιδιοδιανύσματα πυκνών μητρώων)
  - Sparse : LLE και Laplacian Eigenmaps (χρησιμοποιούν non-trivial trailing ιδιοδιανύσματα αραιών θετικά ημιορισμένων μητρώων)
- Γενική βελτιστοποίηση (π.χ. SGD): auto-associative ΤΝΔ

Παραμετρικές και Μη-Παραμετρικές Τεχνικές:
- PCA, classical MDS, autoassiative networks αποτελούν παραμετρικά μοντέλα (μειώνουν τη διάσταση νέων δεδομένων που δεν ήταν στο training set).
- Sammon NLM, CCA, SOM, SNE είναι μη παραμετρικές

Hard DR vs Soft DR:
- PCA και classical MDS από πολλές διαστάσεις μπορούν να πάνε σε πολύ λίγες.
- Οι NLDR τεχνικές αποτυγχάνουν (πλην του SNE) αν προσπαθήσουμε να πάμε κάτω από την εγγενή (intrinsic) διαστατικότητα των δεδομένων.

Αξιολόγηση των αποτελεσμάτων

Το DR συνδέεται με TLS regression, άρα μπορεί να χρησιμοποιηθεί το τετραγωνικό σφάλμα ανακατασκευής σαν μέτρο αξιολόγησης.

- Πρόβλημα: πρέπει να ξαναπροβάλουμε τα δεδομένα στο αρχικό χώρο (μόνο PCA και auto-associative ΤΝΔ μπορούν)

Stress function: υπολογισμός της τιμής της για οποιαδήποτε αντιστοίχιση.

- είναι διαφορίσιμη
- αλλά είναι η διατήρηση των αποστάσεων καλό κριτήριο? (όχι για περίπλοκα manifolds)

rank preservation: πιο κοντά στη βασική αρχή του DR (δηλ. διατήρηση μικρών αποστάσεων μεταξύ γειτονικών σημείων δεδομένων και μεγάλων αποστάσεων μεταξύ μακρινών σημείων δεδομένων)

- rank του ξ_j ως προς το ξ_i : $\rho_{ij} = \left|\{k\colon\ \delta_{ik} < \delta_{ij}\}\right|$
- rank του x_j ως προς το x_i: $r_{ij} = \left|\{k\colon\ d_{ik} < d_{ij}\}\right|$

Rank errors: $\rho_{ij} - r_{ij}$

- $\rho_{ij} - r_{ij} > 0$: intrusion (το j διάνυσμα είναι intruder στην K-γειτονιά του i)
- $\rho_{ij} - r_{ij} < 0$: extrusion

Ευχαριστώ!