Μετασχηματισμοί Δεδομένων: Γένεση Χαρακτηριστικών και Μείωση Αριθμού Διαστάσεων

ECE-TEL830 ΑΝΑΓΝΩΡΙΣΗ ΠΡΟΤΥΠΩΝ

Αθανάσιος Κούτρας

Αναπληρωτής Καθηγητής

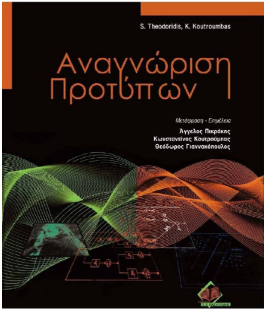
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών,
Παν. Πελοποννήσου

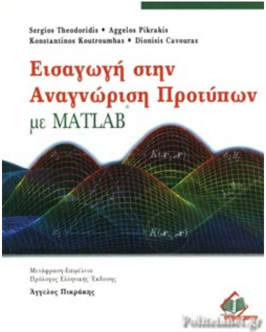
20 Μαΐου 2023

Περίγραμμα διάλεξης

Υλικό μελέτης

- Theodoridis S., Pikrakis, A., Koutroumbas K., Cavouras, D., **"Εισαγωγη στην αναγνώριση προτύπων με MATLAB",ΚΕΦΑΛΑΙΟ** 3
- Theodoridis S., Koutroumbas K., **"Αναγνώριση προτύπων",ΚΕΦΑΛΑΙΟ** 5, 6





Εισαγωγή

- Στην διάλεξη αυτή θα παρουσιαστούν γραμμικές και μη γραμμικές τεχνικές μετασχηματισμών που χρησιμοποιούνται για τη γένεση ενός συνόλου χαρακτηριστικών.
- οι νέες παράμετροι, μπορούν να προκύψουν είτε από ένα σύνολο μετρήσεων, είτε από ένα υπάρχον σύνολο διαθέσιμων χαρακτηριστικών.
- ο σκοπός είναι να δημιουργηθούν νέα χαρακτηριστικά που θα κωδικοποιούν την πληροφορία ταξινόμησης με περισσότερο συμπαγή τρόπο σε σχέση με τα αρχικά χαρακτηριστικά.
- αυτό θα μπορούσε να είναι μείωση των χαρακτηριστικών ή διαφορετικά μείωση του αριθμού των διαστάσεων.

Ανάλυση σε κύριες συνιστώσες - PCA

- η **ανάλυση σε κύριες συνιστώσες** (Principal Component Analysis - PCA) είναι ίσως η πιο δημοφιλής τεχνική μείωσης του αριθμού των διαστάσεων των χαρακτηριστικών.
- ξεκινά από ένα αρχικό σύνολο l χαρακτηριστικών που αποτελούν τα στοιχεία ενός διανύσματος $x \in R^l$ με στόχο την δημιουργία ενός νέου συνόλου χαρακτηριστικών μέσω ενός γραμμικού μετασχηματισμού

$$\boxed{\mathbf{y} = \mathbf{A}^T\mathbf{x}}$$

- το αποτέλεσμα του μετασχηματισμού αυτού είναι ότι τα στοιχεία του y είναι **ασυσχέτιστα**.
- στο δεύτερο στάδιο, θα επιλεγούν τα πιο σημαντικά από τα στοιχεία αυτά.

Βήματα εφαρμογής της PCA

Τα βήματα που ακολουθούμε για την εφαρμογή της PCA είναι τα παρακάτω:

1. εκτίμηση του μητρώου συνδιασποράς **S**. Υποθέτουμε ότι η μέση τιμή είναι ίση με 0, σε διαφορετική περίπτωση απλώς την αφαιρούμε από τα δεδομένα. Στην περίπτωση αυτή ο πίνακας **συνδιασποράς** και **αυτοσυσχέτισης** συμπίπτουν.
2. ανάλυση του S στα **ιδιοδιανύσματα** του. Υπολογίζονται οι l **ιδιοτιμές** και **ιδιοδιανύσματα** $\lambda_i, \alpha_i \in R^l, i = 0, 1, 2 \cdots, l-1$
3. ταξινόμηση ιδιοτιμών σε **φθίνουσα** σειρά $\lambda_0 \geq \lambda_1 \geq \cdots, \geq \lambda_{i-1}$
4. επιλογή των m **μεγαλύτερων** ιδιοτιμών. Για την καλύτερη επιλογή του m, διαλέγουμε αυτό για το οποίο το διάστημα μεταξύ λ_{m-1} και λ_m είναι πολύ μεγάλο.
5. οι ιδιοτιμές που επιλέγουμε είναι γνωστές ως **κύριες συνιστώσες**.

Βήματα εφαρμογής της PCA (συνέχεια)

- χρησιμοποίηση των αντίστοιχων ιδιοδιανυσμάτων (στήλες) $\alpha_i, i = 1, 2, \cdots, m-1$ για την δημιουργία του **πίνακα μετασχηματισμού**

$$\mathbf{A} = [\alpha_0 \quad \alpha_1 \quad \alpha_2 \quad \cdots \quad \alpha_{m-1}]$$

- μετασχηματισμός κάθε $l-$διάστατου διανύσματος x του αρχικού χώρου σε ένα $m-$διάστατο διάνυσμα y μέσω του μετασχηματισμού

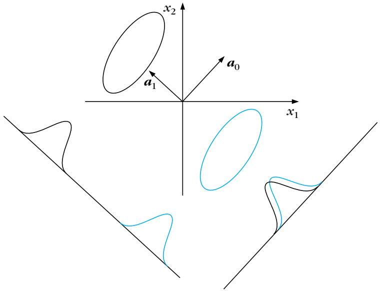
$$\mathbf{y} = \mathbf{A}^T \mathbf{x}$$

Μετασχηματισμός PCA

- μπορεί να δειχτεί ότι η συνολική διασπορά των στοιχείων του x ισούται με το άθροισμα των ιδιοτιμών.
- μετά τον μετασχηματισμό, η διασπορά του κάθε μετασχηματισμένου στοιχείου ισούται με λ_i
- συνεπώς επιλέγοντας τις συνιστώσες που αντιστοιχούν στις m μεγαλύτερες ιδιοτιμές, διατηρείται η μέγιστη διασπορά.
- σε πολλές περιπτώσεις ο μετασχηματισμός PCA δεν οδηγεί αναγκαστικά στην μέγιστη **διαχωρισιμότητα** κλάσεων στον **χαμηλότερης διάστασης** χώρο.
- αυτό είναι λογικό καθώς η μείωση του αριθμού διαστάσεων δεν βελτιστοποιείται ως προς την διαχωρισιμότητα των κλάσεων.

Παράδειγμα μετασχηματισμού PCA

- στο σχήμα φαίνεται ένα παράδειγμα με διανύσματα δύο κλάσεων που ακολουθούν την Gaussian κατανομή με το ίδιο μητρώο συνδιασποράς.
- το σχήμα φαίνονται τα ιδιοδιανύσματα α_0 και α_1 που αντιστοιχούν στην μεγαλύτερη και την μικρότερη ιδιοτιμή αντίστοιχα.
- η προβολή στο α_0 έχει ως αποτέλεσμα να συμπίπτουν σχδόν οι δύο κλάσεις.
- η προβολή στο α_1 διατηρεί την διαχωρισιμότητα των κλάσεων.



Σχήμα: Παράδειγμα μετασχηματισμού PCA ο οποίος δεν είναι η καλύτερη επιλογή στην αναγνώριση προτύπων.

Ανάλυση σε ιδιάζουσες τιμές - SVD

- για έναν πίνακα X ο οποίος έχει διαστάσεις $l \times N$, υπάρχουν μοναδιαίοι τετραγωνικοί πίνακες U και V με διαστάσεις $l \times l$ και $N \times N$ ώστε:

$$X = U \begin{bmatrix} \Lambda^{1/2} & \mathbf{O} \\ \mathbf{O} & 0 \end{bmatrix} V^T$$

- δηλαδή: υπάρχουν ορθοκανονικά μητρώα U και V που μετασχηματίζουν το X στην ειδική διαγώνια δομή του Y
- ο πίνακας Λ είναι τετραγωνικός πίνακας με διαστάσεις $r \times r$ που περιλαμβάνει τις μη-μηδενικές ιδιοτιμές (ιδιάζουσες τιμές)του XX^T ως ακολούθως:

$$\Lambda^{1/2} = \begin{bmatrix} \sqrt{\lambda_0} & & & \\ & \sqrt{\lambda_1} & & \\ & & \dots & \\ & & & \sqrt{\lambda_{r-1}} \end{bmatrix}$$

- με βάση το παραπάνω, μπορούμε να γράψουμε το X ως:

$$\boxed{X = \sum_{i=0}^{r-1} \sqrt{\lambda_i} u_i v_i^T}$$

με u_i, v_i τα αντίστοιχα ιδιοδιανύσματα των $XX^T, X^T X$ αντίστοιχα.

Ανάλυση σε ιδιάζουσες τιμές - SVD

- αν στο προηγούμενο άθροισμα κρατήσουμε $m \leq r$ όρους:

$$X = \sum_{i=0}^{m-1} \sqrt{\lambda_i} u_i v_i^T$$

  τότε το X είναι η καλύτερη προσέγγιση τάξης m του X.
- η τεχνική αυτή ονομάζεται ανάλυση σε ιδιάζουσες τιμές ή **Singular Value Decomposition - SVD**
- η τεχνική αυτή βασίζεται στην **καθολική πληροφορία** που είναι διεσπαρμένη **σε όλα τα διανύσματα δεδομένων** του X
- είναι πολύ αποτελεσματική σε περιπτώσεις όπου τα δεδομένα μπορούν να περιγραφούν επαρκώς μέσω του μητρώου συνδιασποράς (π.χ. όταν ακολουθούν κατανομές παρόμοιες με την Gaussian κατανομή).

Ανάλυση Γραμμικής Διάκρισης κατά Fisher

- στην μέθοδο PCA, η μείωση του αριθμού των διαστάσεων επιτυγχάνεται χωρίς επίβλεψη.
- στην μέθοδο που θα παρουσιάσουμε εδώ, ο υποχώρος στον οποίο γίνεται η προβολή προκειμένου να μειωθεί ο αριθμός των διαστάσεων είναι αποτέλεσμα **επιτηρούμενης** διαδικασίας.
- για την περίπτωση δύο κλάσεων, ο στόχος είναι να βρούμε μια μοναδική διεύθυνση w τέτοια ώστε οι προβολές y των $l-$διάστατων διανυσμάτων χαρακτηριστικών $x \in R^l$ να μεγιστοποιούν τον **λόγο διάκρισης Fisher**

$$\boxed{FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}}$$

- ο στόχος είναι μετά την προβολή στην διεύθυνση w οι δύο κλάσεις να είναι όσο το δυνατόν **απομακρυσμένες** και οι **διασπορές** τους να είναι όσο το δυνατό **μικρότερες**.

Ανάλυση Γραμμικής Διάκρισης κατά Fisher

- η διεύθυνση w είναι το μέγιστο ιδιοδιάνυσμα του γινομένου των πινάκων $S_w^{-1}S_b$, όπου για ισοπίθανες κλάσεις, ο πίνακας

$$\boxed{S_w = \frac{1}{2}(S_1 + S_2)}$$

ονομάζεται **πίνακας σκέδασης εντός κλάσεων** (within class scatter matrix).

- ο πίνακας S_b ονομάζεται **πίνακας σκέδασης μεταξύ κλάσεων** (between class sctter matrix) και είναι ίσος με

$$\boxed{S_b = \frac{1}{2}(m_1 - m_0)(m_1 - m_0)^T + \frac{1}{2}(m_2 - m_0)(m_2 - m_0)^T}$$

όπου m_0 είναι η συνολική μέση τιμή των δεδομένων στον αρχικό χώρο.

- σε αυτή την περίπτωση, δεν είναι απαραίτητο να υπολογίσουμε ιδιοτιμές και η λύση προκύπτει απευθείας από την σχέση

$$\boxed{w = S_w^{-1}(m_1 - m_2)}$$

- η τεχνική αυτή ονομάζεται **ανάλυση γραμμικής διάκρισης** ή Linear Discriminant Analysis - LDA

Ανάλυση σε κύριες συνιστώσες με χρήση πυρήνων

- όλες οι προηγούμενες μέθοδοι μείωσης διαστατικότητας είναι γραμμικές.
- πρώτα κατασκευάζουν έναν υπο-χώρο χαμηλής διάστασης τον οποίο για παράδειγμα τον ορίζουν οι m κυρίαρχες διευθύνσεις του αρχικού χώρου.
- η επιλογή των κυρίαρχων διευθύνσεων εξαρτάται από τη μέθοδο που χρησιμοποιούμε κάθε φορά.
- σε δεύτερο βήμα, όλα τα διανύσματα που μας ενδιαφέρουν, προβάλονται σε αυτόν τον υποχώρο **με γραμμικό τρόπο**.
- αυτό μπορεί να εφαρμοστεί καλά στην περίπτωση που τα δεδομένα μας βρίσκονται στον χώρο R^l σε μια γραμμική πολλαπλότητα.
- δεν είναι πάντα έτσι όμως: τα δεδομένα μας μπορεί να κατανέμονται γύρω από μια πολλαπλότητα χαμηλότερης διάστασης που δεν είναι όμως γραμμική (π.χ. γύρα από κύκλο, σφαίρα κλπ)

Ανάλυση σε κύριες συνιστώσες με χρήση πυρήνων

- σε αυτή την περίπτωση μπορούμε να χρησιμοποιήσουμε την **Ανάλυση σε Κύριες Συνιστώσες με χρήση Πυρήνων** (kernel PCA)
- σύμφωνα με αυτή τη μέθοδο, τα δεδομένα αρχικά απεικονίζονται σε ένα χώρο υψηλής διάστασης μέσω μιας μη γραμμικής απεικόνισης

$$x \in R^l \rightarrow \phi(x) \in H$$

- στη συνέχεια εφαρμόζεται η PCA στον νέο χώρο H. Στον χώρο αυτό (RKHS), τα εσωτερικά γινόμενα που πρέπει να υπολογιστούν, μπορούν να εκφραστούν με τη χρήση πυρήνα.
- παρόλο που εφαρμόζεται η (γραμμική) PCA στον χώρο H, λόγω της μη γραμμικής φύσης της συνάρτησης απεικόνισης $\phi(\cdot)$, η συνολική μέθοδος ισοδυναμεί με μη γραμμική μέθοδο στον αρχικό χώρο.

Signal & Image Processing, Pattern Recognition Group (SIPPRE)
www.sippre-group.com