Γραμμική Αλγεβρα II

PCA
(παράδειγμα)

15 Μαιου 2017

Μας δίνόνται τα παρακάτω δεδομένα.

Βάρος (lb)	120	125	125	135	145
Υψος (inch)	61	60	64	68	72

Να γίνει η ανάλυση κυρίων συνιστωσών (PCA).

Λυση:
Ευκολα μπορούμε να υπολογίσουμε το διάνυσμα των μέσων ως εξης

$$M = \frac{1}{5}\begin{bmatrix}\sum x_i \\ \sum y_i\end{bmatrix} = \begin{bmatrix}130 \\ 65\end{bmatrix}$$

Μετά αφαιρώντας τον μέσο απο κάθε x_i έχουμε

$$\hat{x_1} = x_1 - M = \begin{bmatrix}-10 \\ -4\end{bmatrix}, \quad \hat{x_2} = x_2 - M = \begin{bmatrix}-5 \\ -5\end{bmatrix}$$

$$\hat{x_3} = x_3 - M = \begin{bmatrix}-5 \\ -1\end{bmatrix}, \quad \hat{x_4} = x_4 - M = \begin{bmatrix}5 \\ 3\end{bmatrix}, \quad \hat{x_5} = x_5 - M = \begin{bmatrix}15 \\ 7\end{bmatrix},$$

Αρα τα δεδομένα μας γίνονται ως εξης

$$B = \begin{bmatrix}10 & -5 & -5 & 5 & 15 \\ -4 & -5 & -1 & 3 & 7\end{bmatrix}.$$

Απ΄ τον πίνακα B προκύπτει ο πίνακας διακύμανσης ως εξής

$$\Sigma = \frac{1}{N-1}BB^\top = \begin{bmatrix}100 & 47.5 \\ 47.5 & 25\end{bmatrix}$$

οπου παρατηρούμε οτι η συνόλικη διακύμανση είναι ίση με 125. Αρα η μετάβλητη βάρος έχουμε οτι περιγράφει το $\frac{100}{125}\%$ της πληροφορίας που υπάρχει στα δεδομένα ενω η μετάβλητη υψος περιγράφει το $\frac{25}{125}\%$.
Ας αρχίσουμε τώρα την ανάλυση σε κύριες συνιστώσες, υπολογίζοντας τις ιδιοτιμές του πίνακα Σ. Εχουμε ότι

$$\det(\Sigma - \lambda I) = 0 \Rightarrow \cdots \Rightarrow \lambda_1 = 123, \lambda_2 = 2$$

Ετσι προκύπτει οτι η μετάβλητη βάρος έχει την μεγαλύτερη ιδιοτιμή ($\lambda_1 = 123$) και αρα είναι η μεγαλύτερη συνιστώσα. Το αντιστοιχο ιδιοδιάνυσμα της προκύπτει οτι είναι το $u_1 = [0.9 \;\; 0.436]^\top$. Αξίζει να σημειώσουμε οτι η συνόλικη διακύμανση δεν έχει μεταβληθεί ($\lambda_1 + \lambda_2 = 125$). Ενδείκτικα, αναφέρουμε οτι ένας δείκτης που θα λάμβανε υπόψιν του το βάρος και το υψός που έχουμε διαθέσιμα μέσο αυτού του δείγματος είναι ο

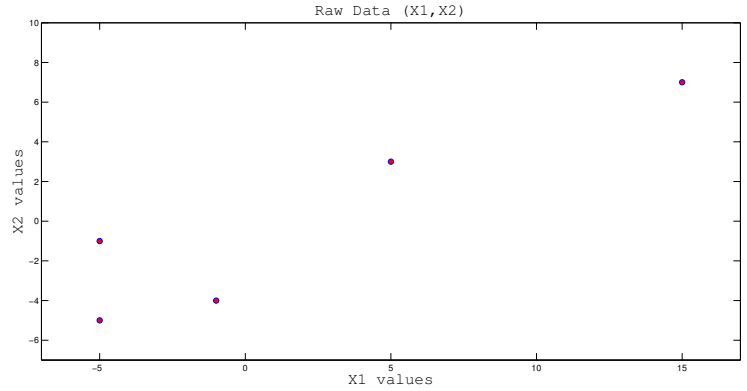
$$y = 0.9x_1 + 0.436x_2$$

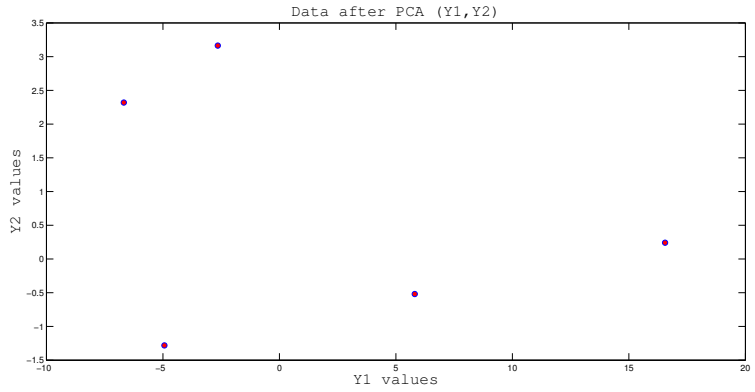
Ο νέος πίνακας διακύμανσης θα είναι

$$\hat{\Sigma} = \begin{bmatrix} 123 & 0 \\ 0 & 2 \end{bmatrix}$$

Ετσι τώρα διαπιστώνουμε οτι τελικώς, μέτα απο την ανάλυση κυρίων συνιστωσών, η πρώτη συνιστώσα έχει το $\frac{123}{125} = 98.4\ \%$ της πληρόφοριας ένω η δευτερη συνιστώσα έχει το $\frac{2}{125} = 1.6\ \%$ της πληρόφοριας. Σημειώνουμε οτι μέτα την ανάλυση κυρίων συνιστωσών, ο νεός πινακας διακυμάνσεων, δεν περιέχει συνδιακυμάνσεις, αρα οι μετάβλητες μας έγιναν ασυσχέτιστες. Στη συνέχεια, χρησιμοποιώντας τον ορθογώνιο πίνακα P που περιέχει τα ιδιοδιανύσματα του Σ σε στήλες, θέτουμε $X = PY \rightarrow Y = P^T X$ και παίρνουμε τα νέα δεδομένα Y, τα οποία οπως ειπαμε ειναι ασυσχέτιστα. Τα παραπάνω φαίνονται στις εικόνες που ακολουθούν:

Σχήμα 1: Τα δεδομένα πριν την Ανάλυση κύριων συνιστωσών





Σχήμα 2: Τα δεδομένα μετα την Ανάλυση κύριων συνιστωσών

Φαίνεται καθαρά οτι τα δεδομένα έχουν χωριστεί σε δυο ομάδες που βρίσκονται πάνω σε δυο ευθείες.