Unsupervised Learning

Supervised Learning

predict outcome variable Y using a set of p features $X_1, X_2, X_3, \dots X_p$ measured on n observations

train the program using well labeled data to predict outcomes for unforeseen data

Unsupervised Learning

set of tools intended to explore only a set of p features $X_1, X_2, X_3, \dots X_p$

performed as part of an exploratory data analysis to discover interesting things about the features

no simple goal for the analysis, not trying to predict anything

looking for relationship between variables and observations

Principal Component Analysis (PCA)

dataset contains n observations with measurements on a set of p variables $X_1, X_2, X_3, \dots X_p$

PCA produces a low-dimensional representation of a data set that contains as much of the variation as possible

Input

data matrix X where the columns are centered to have a mean of 0

p columns represent variables, e.g. age, weight, blood pressure

n rows represent different subjects

usually there are many more rows than columns

Output

data matrix Y whose columns are principal components

can have at most as many principal components as p, assuming $p < n$

Principal Components (PCs)

new variables created from orthogonal transformations of the columns of X

mutually uncorrelated

sorted by decreasing variance

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ x_{31} & x_{32} & x_{33} & \dots & x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{bmatrix} = [x_1 \quad x_2 \quad x_3 \quad \dots \quad x_p] \Rightarrow Y = [Y_1 \quad Y_2 \quad Y_3 \quad \dots \quad Y_p]$$

$$x_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ \vdots \\ x_{n1} \end{bmatrix}, x_2 = \begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \\ \vdots \\ x_{n2} \end{bmatrix}, x_3 = \begin{bmatrix} x_{13} \\ x_{23} \\ x_{33} \\ \vdots \\ x_{n3} \end{bmatrix}, \dots x_p = \begin{bmatrix} x_{1p} \\ x_{2p} \\ x_{3p} \\ \vdots \\ x_{np} \end{bmatrix}$$

column vectors

$$\sum_i x_{i1} = \sum_i x_{i2} = \sum_i x_{i3} = \cdots = \sum_i x_{ip} = 0$$

input variables X_j are centered around 0

$$Y_j = \sum_i \Theta_{ij} x_i$$

$$\sum_j^p \Theta_{ji}{}^2 = 1$$

Θ_{j1} = loading vector of the 1st principal component
loadings are normalized so sum of squares is equal to 1 so variance isn't too large

First Principal Component

$$Y_1 = \Theta_{11}x_1 + \Theta_{21}x_2 + \Theta_{31}x_3 + \cdots + \Theta_{p1}x_p = \Theta_{11}\begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ \vdots \\ x_{n1} \end{bmatrix} + \Theta_{21}\begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \\ \vdots \\ x_{n2} \end{bmatrix} + \Theta_{31}\begin{bmatrix} x_{12} \\ x_{22} \\ x_{32} \\ \vdots \\ x_{n2} \end{bmatrix} + \cdots + \Theta_{p1}\begin{bmatrix} x_{1p} \\ x_{2p} \\ x_{3p} \\ \vdots \\ x_{np} \end{bmatrix}$$

$$= \begin{bmatrix} \Theta_{11}x_{11} & \Theta_{21}x_{12} & \Theta_{31}x_{13} & \dots & \Theta_{p1}x_{1p} \\ \Theta_{11}x_{21} & \Theta_{21}x_{22} & \Theta_{31}x_{23} & \dots & \Theta_{p1}x_{2p} \\ \Theta_{11}x_{31} & \Theta_{21}x_{32} & \Theta_{31}x_{33} & \dots & \Theta_{p1}x_{3p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \Theta_{11}x_{n1} & \Theta_{21}x_{n2} & \Theta_{31}x_{n3} & \dots & \Theta_{p1}x_{np} \end{bmatrix} = \begin{bmatrix} y_{11} \\ y_{21} \\ y_{31} \\ \vdots \\ y_{n1} \end{bmatrix}$$

first principal component = normalized the linear combination of the vectors $x_1, x_2, x_3, \dots x_p$ that has the largest variance

PCA calculates the loadings that maximizes sample variance of the first PC

$y_{11}, y_{21}, y_{31}, \dots, y_{n1}$ are the scores of the first PC

$$y_{i1} = \Theta_{11}x_{i1} + \Theta_{21}x_{i2} + \Theta_{31}x_{i3} + \cdots + \Theta_{p1}x_{ip}$$

$$\overline{Y_1} = \frac{1}{n}\sum_i^n y_{i1} = \frac{1}{n}\sum_i^n \left(\Theta_{11}x_{i1} + \Theta_{21}x_{i2} + \Theta_{31}x_{i3} + \cdots + \Theta_{p1}x_{ip}\right) = \frac{1}{n}\sum_i^n\sum_j^p \Theta_{j1}x_{ij}$$

$$= \frac{1}{n}\sum_j^p \Theta_{j1}\left(\sum_i^n x_{ij}\right) = 0$$

average of y_{i1} is 0 because input variables are centered around 0

$$V(Y_1) = \frac{1}{n}\sum_i^n (y_{i1} - \overline{Y_1})^2 = \frac{1}{n}\sum_i^n (y_{i1})^2 = \frac{1}{n}\sum_j^p \left(\sum_i^n \Theta_{j1}x_{ij}\right)^2 = 0$$

PCA calculates Θ_{j1} to maximize $V(Y_1)$

Geometry of PCA

Θ_1 has elements $\Theta_{11}, \Theta_{21}, \Theta_{31}, \dots, \Theta_{p1}$

Θ_1 defines a direction in feature space along which the data varies the most

n data points $x_1, x_2, x_3, \dots x_n$ projected onto that direction becomes principal component scores $y_{11}, y_{21}, y_{31}, \dots, y_{n1}$

second PC is orthogonal to the first PC because they are uncorrelated