Unsupervised Learning

Supervised Learning

predict outcome variable using a set of features measured on 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 features

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 observations with measurements on a set of variables

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

Input

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

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

rows represent different subjects

usually there are many more rows than columns

Output

data matrix whose columns are principal components

can have at most as many principal components as , assuming

Principal Components (PCs)

new variables created from orthogonal transformations of the columns of

mutually uncorrelated

sorted by decreasing variance

column vectors

input variables are centered around 0

= 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

first principal component = normalized the linear combination of the vectors that has the largest variance

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

are the scores of the first PC

average of is 0 because input variables are centered around 0

PCA calculates to maximize

**Geometry of PCA**

has elements

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

data points projected onto that direction becomes principal component scores

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