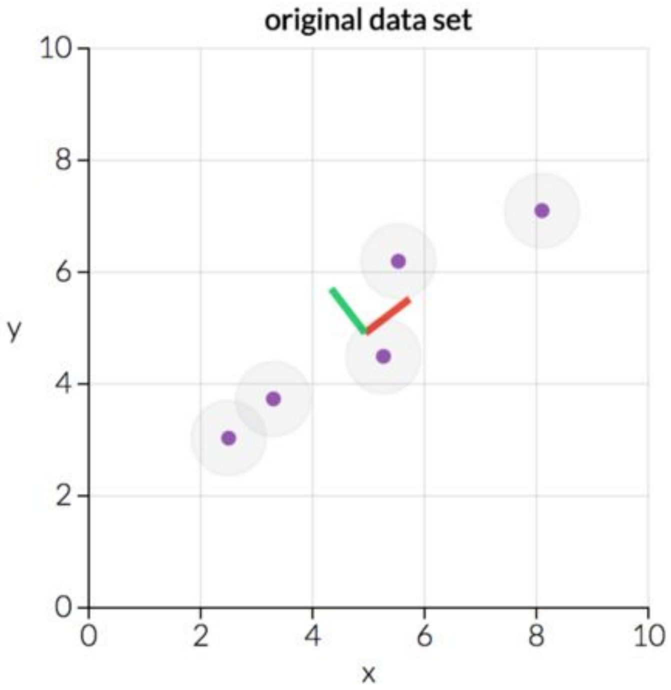
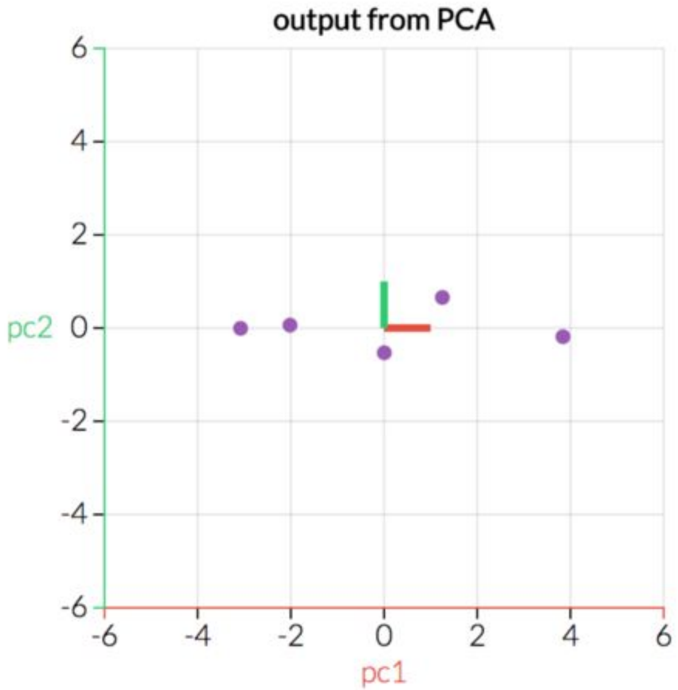
Intro to Machine Learning

Recitation - Homework 6

21st November 2022

Principal Component Analysis (PCA)





PCA - Main Ideas

- Reducing the number of variables of a dataset while preserving as much information as possible
- Principal Components: The new variables that are constructed as linear combinations of the original variables

PCA: Step-by-Step

- Step 1: Centering the dataset

  Corresponds to subtracting the mean from each of sample based on the feature values

$$z = value - mean$$

PCA: Step-by-Step

- Step 2: Compute the covariance matrix

  Compute the covariance matrix of the centered matrix calculated in the previous step

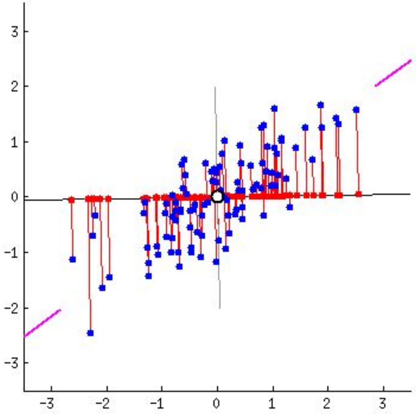
  The covariance matrix is a *p x p* symmetric matrix (where *p* is the number of dimensions) that has as entries the covariances associated with all possible pairs of the initial variables.

$$\begin{bmatrix} Cov(x,x) & Cov(x,y) & Cov(x,z) \\ Cov(y,x) & Cov(y,y) & Cov(y,z) \\ Cov(z,x) & Cov(z,y) & Cov(z,z) \end{bmatrix}$$

PCA: Step-by-Step

- Step 3: Compute the eigenvectors and eigenvalues of the covariance matrix to identify the principal components

Eigenvectors of the covariance matrix are actually the directions of the axes where there is the most amount of variance (i.e., maximum information) — and so, these are also called the “principal components”



PCA: Step-by-Step

- Step 4: Sort the eigenvectors in order of their eigenvalues (in descending order) to get the principal components in order of their significance

$$v1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix} \qquad \lambda_1 = 1.284028$$

$$v2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix} \qquad \lambda_2 = 0.04908323$$

PCA: Step-by-Step

- Step 5: Subsample the principal components

$$v1 = \begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix} \qquad \lambda_1 = 1.284028$$

$$v2 = \begin{bmatrix} -0.7351785 \\ 0.6778736 \end{bmatrix} \qquad \lambda_2 = 0.04908323$$

$$\begin{bmatrix} 0.6778736 & -0.7351785 \\ 0.7351785 & 0.6778736 \end{bmatrix}$$

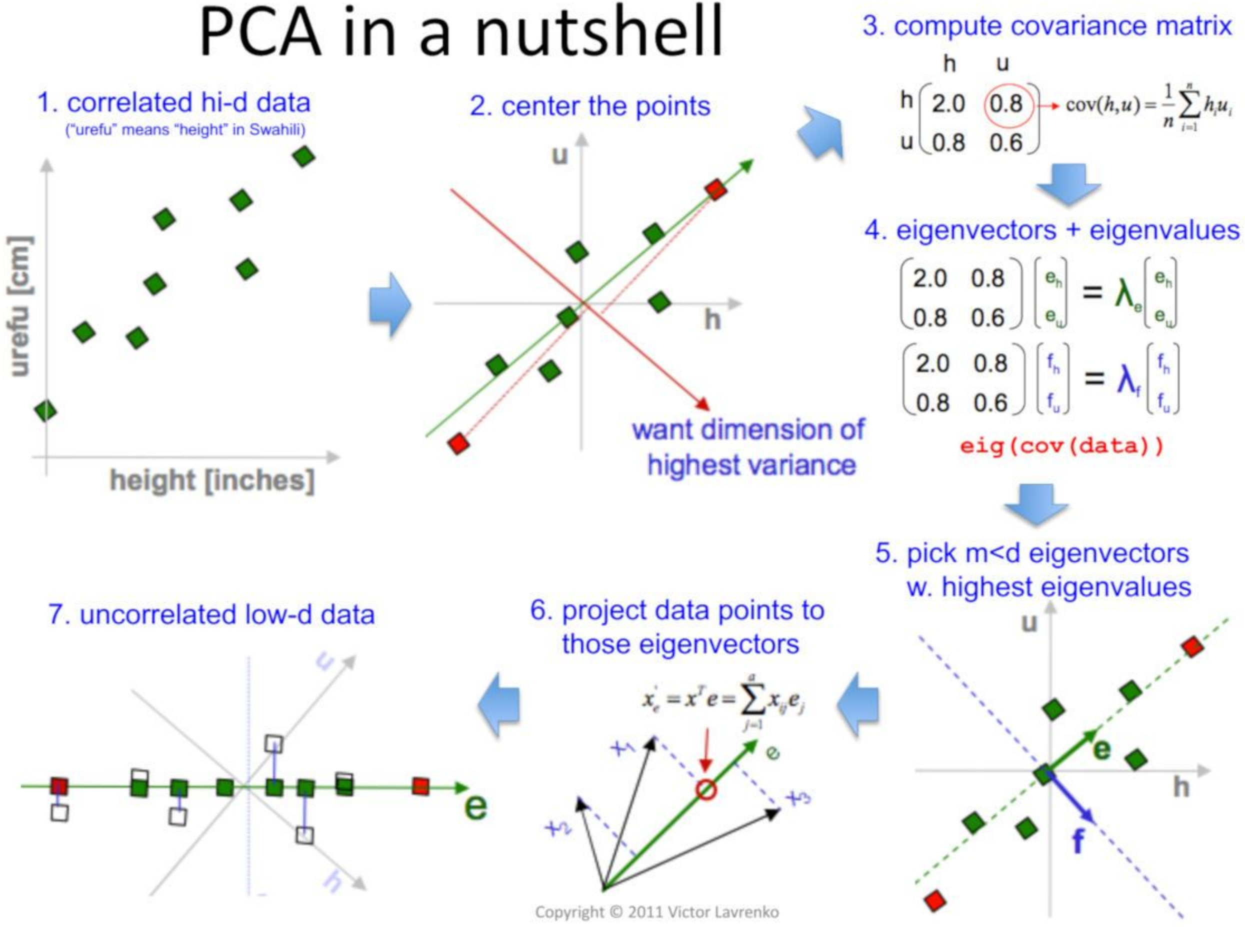
$$\begin{bmatrix} 0.6778736 \\ 0.7351785 \end{bmatrix}$$

PCA: Step-by-Step

- Step 6: Transform the data along the principal component(s) axes

  This can be achieved by multiplying the transpose of the original data set by the transpose of the eigenvector subset

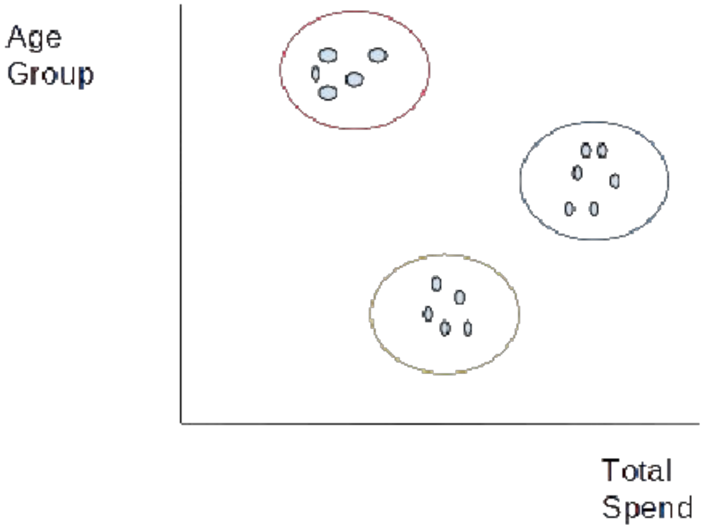
PCA in a nutshell



Source: https://devopedia.org/principal-component-analysis

K-Means clustering

K–Means clustering is a centroid based algorithm where we calculate the distance between each data point and a centroid to assign it to a cluster. The goal is to identify the K number of groups in the dataset.



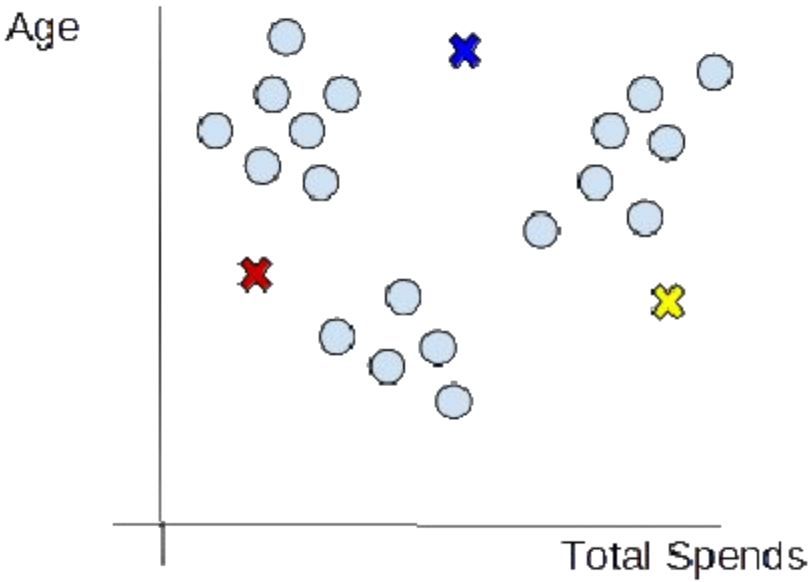
K-Means clustering: Step-by-Step

- Step 1: Choosing the “K” — the number of clusters

  Given: K = 10 (in the homework)

K-Means clustering: Step-by-Step

- Step 2: Initializing centroids

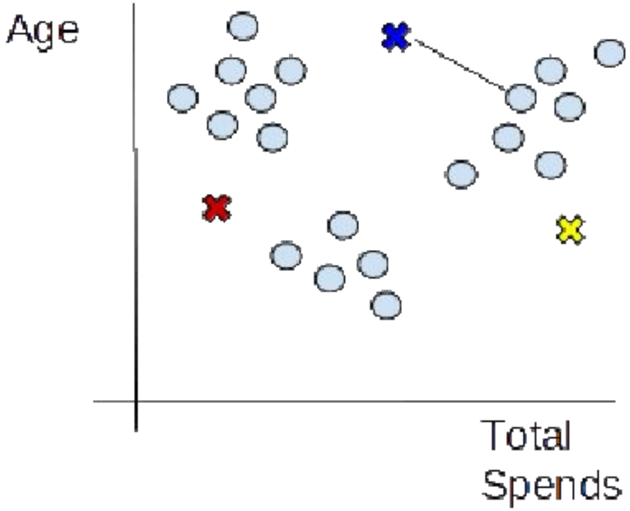


K-Means clustering: Step-by-Step

- Step 3: Assign data points to the nearest cluster

  In this step, we first calculate the distance between the data point X and centroid C using Euclidean Distance function and then choose the cluster for data points where the distance between each data point and the centroid is minimum

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

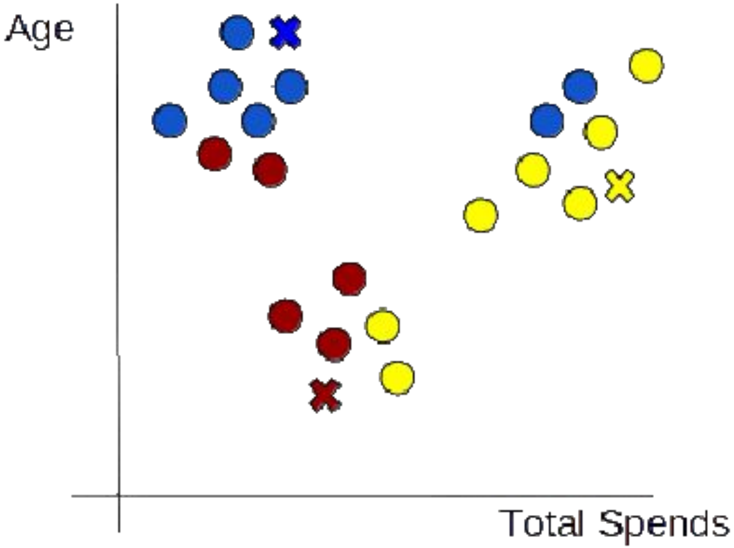


K-Means clustering: Step-by-Step

- Step 4: Recompute the cluster centroids

  Now that we have new members in and definitions of each cluster, we will recompute the centroids by calculating the average of all data points of that cluster

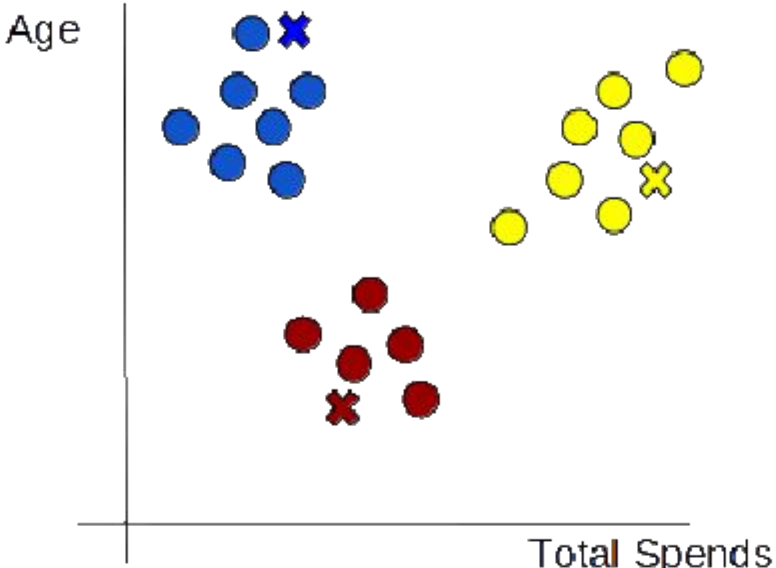
$$Ci = \frac{1}{|Ni|} \sum xi$$



K-Means clustering: Step-by-Step

- Step 5: Repeat steps 3-4 until convergence

  Convergence here implies the state where with any two iterations, the centroids remain the same



Homework 6