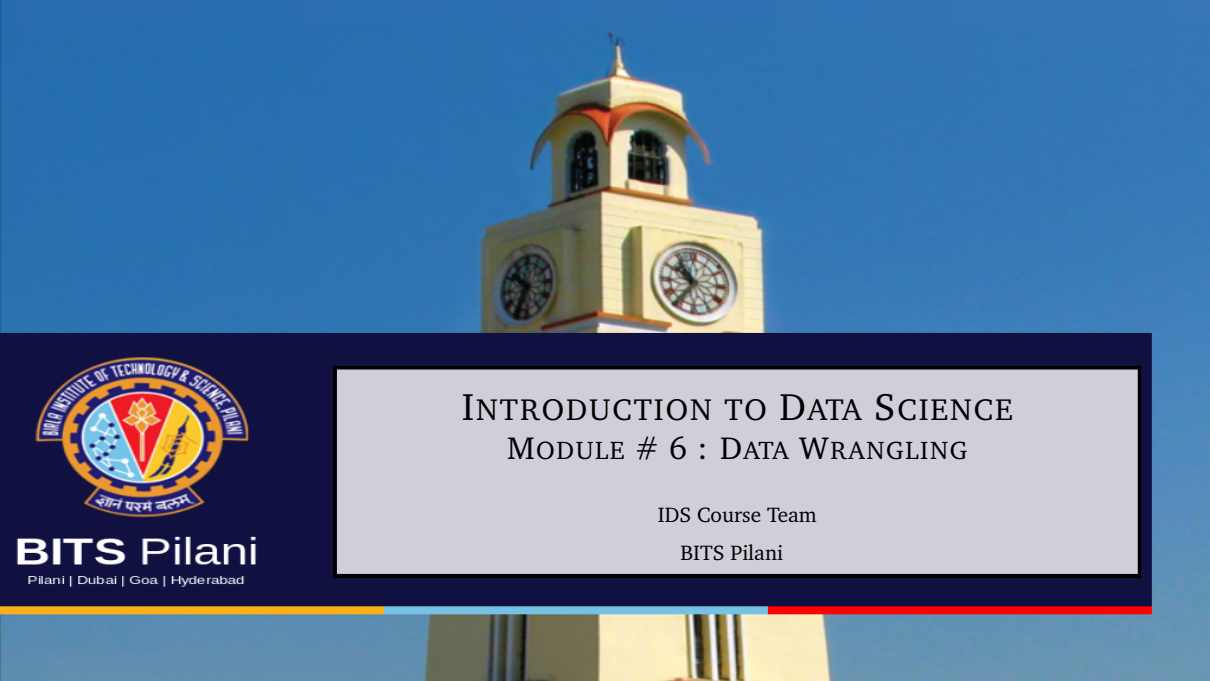


BITS Pilani
Pilani | Dubai | Goa | Hyderabad

Introduction to Data Science
Module # 6 : Data Wrangling

IDS Course Team

BITS Pilani

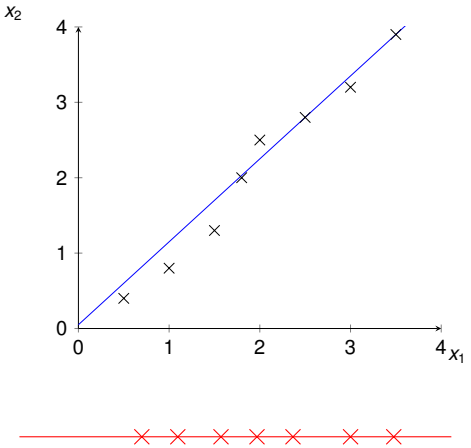
The instructor is gratefully acknowledging the authors who made their course materials freely available online.

Table of Contents

1 Dimensionality Reduction

Dimensionality Reduction – Motivation

- Data Compression
  - Reduce data from 2D to 1D
  - Reduce data from 3D to 2D
  - Reduce data from n-D to 2D
- Example:
  - x_1 can be pilot skill
  - x_2 can be pilot engagement
  - Line represents pilot aptitude.
  - How to come with that line?
  - $x^{(m)} \in \mathbb{R}^2 \rightarrow z^{(m)} \in \mathbb{R}^1$



Dimensionality Reduction – Motivation

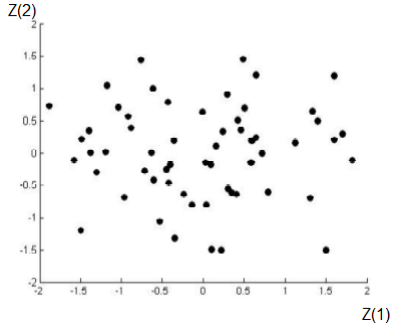
Data Visualization – $X^{(i)} \in \mathbb{R}^{50}$

Country	GDP trillions of US $	Per capita GDP	Human development index	Life expectancy	Poverty index Gini as %	Mean household income thousands of US $	
Canada	1.57	39.17	0.91	80.7	32.6	67.29	...
China	5.88	7.54	0.69	73.0	46.9	10.22	...
India	1.63	3.41	0.55	64.7	36.8	0.74	...
Russia	1.48	19.84	0.75	65.5	39.9	0.72	...
Singapore	0.22	56.7	0.87	80.0	42.5	67.1	...
USA	14.53	46.7	0.91	78.3	40.8	84.3	...
...	...	...	...	...	...	...	...

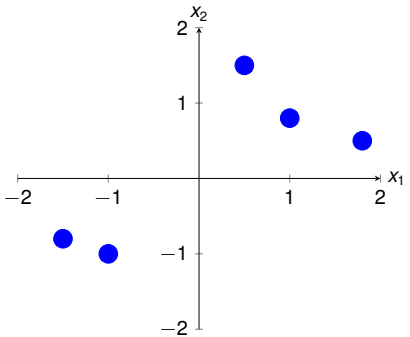
Dimensionality Reduction – Motivation

Data Visualization – reduce data from $X^{(i)} \in \mathbb{R}^{50}$ to $Z^{(i)} \in \mathbb{R}^2$

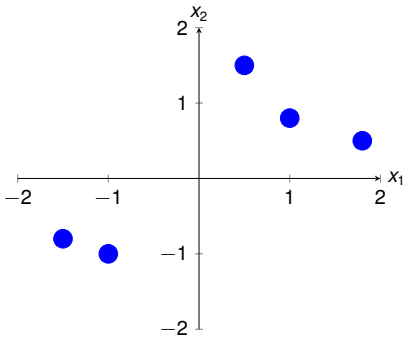
Country	z1	z2
Canada	1.6	1.2
China	1.7	0.3
India	1.6	0.2
Russia	1.4	0.5
Singapore	1.4	0.5
USA	0.5	1.7
...	...	...

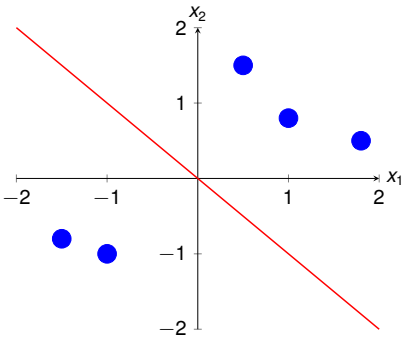


Principal Component Analysis (PCA)

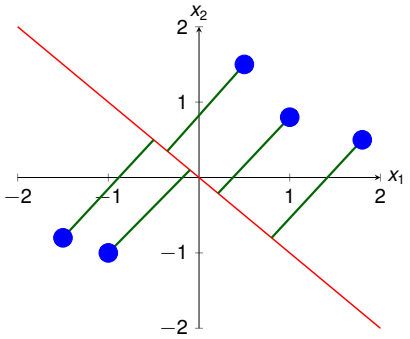


Principal Component Analysis (PCA)





Principal Component Analysis (PCA)



- Reduce from 2-D to 1-D
  Find a direction or a vector $u^{(1)} \in \mathbb{R}^n$ onto which to project the data so as to minimize the projection error.
- Reduce from n-D to k-D
  Find k vectors $u^{(1)}, u^{(2)} \ldots u^{(k)}$ onto which to project the data so as to minimize the projection error.

Andrew Ng

Principal Component Analysis (PCA) Algorithm

- Training set: $x^{(1)}, x^{(2)} \ldots x^{(m)}$
- Pre-processing — Perform Feature scaling.

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)}$$

  - Replace each $x_j^{(i)}$ with $x_j^{(i)} - \mu_j$.
  - If different features are on different scales, scale the features to have comparable range of values.

Andrew Ng

Principal Component Analysis (PCA) Algorithm

- Given Scaled Training set: X
- Compute the "covariance matrix" *Sigma*.

$$Sigma = \frac{1}{m} \sum_{i=1}^{m} X^\top X$$

- Compute the "eigen vectors" of covariance matrix.

$$[U, S, V] = svd(Sigma)$$

- Select first k eigen vectors.

$$Ureduce = U(:, 1 : k)$$

- Compute the PCA component corresponding to the eigen vectors

$$Z = Ureduce^\top X$$

Principal Component Analysis (PCA) Algorithm

- Choosing the number of principal components k.
- Compute the "eigen vectors" of covariance matrix.

$$[U, S, V] = svd(Sigma)$$

S is a diagonal matrix.

- Pick the smallest value of k for which

$$\frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{m} S_{ii}} \geq 0.99 \qquad \text{99\% of variance retained}$$

Andrew Ng

PCA Example – 1

Perform PCA on $X = \begin{bmatrix} 4 & -2 \\ -1 & 3 \end{bmatrix}$

- Step 1: Split as different features.

$$x_0 = \begin{bmatrix} 4 \\ -1 \end{bmatrix} \qquad x_1 = \begin{bmatrix} -2 \\ 3 \end{bmatrix}$$

- Step 2: Compute the "covariance matrix" *Sigma*.

$$Sigma = \mathbb{E}[xx^\top] - \bar{x}\bar{x}^\top$$

$$\bar{x} = \frac{1}{m}\sum_{k=0}^{m} x_k$$

$$= \frac{1}{2}\begin{bmatrix} 4 \\ -1 \end{bmatrix} + \begin{bmatrix} -2 \\ 3 \end{bmatrix} = \frac{1}{2}\begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

PCA Example – 2

- Step 2: Compute the "covariance matrix" *Sigma*.

$$\mathbb{E}[xx^\top] = \frac{1}{m}\sum_{k=0}^{m} x_k x_k^\top$$

$$= \frac{1}{2}\left\{\begin{bmatrix}4\\-1\end{bmatrix}\begin{bmatrix}4 & -1\end{bmatrix} + \begin{bmatrix}-2\\3\end{bmatrix}\begin{bmatrix}-2 & 3\end{bmatrix}\right\}$$

$$= \frac{1}{2}\left\{\begin{bmatrix}16 & -4\\-4 & +1\end{bmatrix} + \begin{bmatrix}4 & -6\\-6 & 9\end{bmatrix}\right\} = \begin{bmatrix}2 & -1\\-1 & 1\end{bmatrix}$$

$$Sigma = \mathbb{E}[xx^\top] - \bar{x}\bar{x}^\top$$

$$= \begin{bmatrix}2 & -1\\-1 & 1\end{bmatrix} - \begin{bmatrix}1\\1\end{bmatrix}\begin{bmatrix}1 & 1\end{bmatrix} = \begin{bmatrix}1 & -2\\-2 & 0\end{bmatrix}$$

PCA Example – 3

- Step 3: Compute the eigen values of covariance matrix.

$$| Sigma - \lambda \mathcal{I} | = 0$$

$$det \left\{ \begin{bmatrix} 1 & -2 \\ -2 & 0 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right\} = 0$$

$$det \left\{ \begin{bmatrix} 1-\lambda & -2 \\ -2 & -\lambda \end{bmatrix} \right\} = 0$$

$$(1-\lambda)(-\lambda) - 4 = 0$$

$$\lambda^2 - \lambda - 4 = 0$$

$$\lambda = \frac{+1 \pm \sqrt{1+16}}{2} = 2.56, (-1.56)$$

PCA Example – 4

- Step 4: Compute the eigen values of covariance matrix.

$$| Sigma - \lambda \mathcal{I} | = 0$$

$$det \left\{ \begin{bmatrix} 1 & -2 \\ -2 & 0 \end{bmatrix} - \begin{bmatrix} \lambda & 0 \\ 0 & \lambda \end{bmatrix} \right\} = 0$$

$$det \left\{ \begin{bmatrix} 1-\lambda & -2 \\ -2 & -\lambda \end{bmatrix} \right\} = 0$$

$$(1-\lambda)(-\lambda) - 4 = 0$$

$$\lambda^2 - \lambda - 4 = 0$$

$$\lambda = \frac{+1 \pm \sqrt{1+16}}{2} = 2.56, (-1.56)$$

PCA Example – 5

- Step 5: Compute the eigen vectors corresponding to the eigen values.
  For $\lambda_0 = 2.56$

$$(Sigma - \lambda_0 \mathcal{I})\phi_0 = 0$$

$$\left\{ \begin{bmatrix} 1 & -2 \\ -2 & 0 \end{bmatrix} - \begin{bmatrix} 2.56 & 0 \\ 0 & 2.56 \end{bmatrix} \right\} \begin{bmatrix} \phi_{00} \\ \phi_{01} \end{bmatrix} = 0$$

$$\left\{ \begin{bmatrix} -1.56 & -2 \\ -2 & -2.56 \end{bmatrix} \right\} \begin{bmatrix} \phi_{00} \\ \phi_{01} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$-1.56\phi_{00} - 2\phi_{01} = 0$$

$$-2\phi_{00} - 2.56\phi_{01} = 0$$

$$\text{Solving} \quad \phi_0 = \begin{bmatrix} -1.28 \\ 1 \end{bmatrix}$$

PCA Example – 6

- Step 5: Compute the eigen vectors corresponding to the eigen values.
  For $\lambda_1 = -1.56$

$$(Sigma - \lambda_1 \mathcal{I})\phi_1 = 0$$

$$\left\{ \begin{bmatrix} 1 & -2 \\ -2 & 0 \end{bmatrix} - \begin{bmatrix} -1.56 & 0 \\ 0 & -1.56 \end{bmatrix} \right\} \begin{bmatrix} \phi_{10} \\ \phi_{11} \end{bmatrix} = 0$$

$$\left\{ \begin{bmatrix} -0.56 & -2 \\ -2 & 1.56 \end{bmatrix} \right\} \begin{bmatrix} \phi_{10} \\ \phi_{11} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$-1.56\phi_{10} - 2\phi_{11} = 0$$

$$-2\phi_{10} + 1.56\phi_{11} = 0$$

$$\text{Solving} \quad \phi_1 = \begin{bmatrix} 0.78 \\ 1 \end{bmatrix}$$

PCA Example – 7

- Step 6: Normalize the eigen vectors.

$$\frac{\phi_{00}}{\| \phi_{00} \|} = \frac{\phi_{00}}{\sqrt{\phi_{00}^2 + \phi_{01}^2}} = -0.788$$

$$\frac{\phi_{01}}{\| \phi_{10} \|} = \frac{\phi_{01}}{\sqrt{\phi_{00}^2 + \phi_{01}^2}} = 0.6156$$

$$\frac{\phi_{10}}{\| \phi_{10} \|} = \frac{\phi_{10}}{\sqrt{\phi_{10}^2 + \phi_{11}^2}} = 0.615$$

$$\frac{\phi_{11}}{\| \phi_{11} \|} = \frac{\phi_{11}}{\sqrt{\phi_{10}^2 + \phi_{11}^2}} = 0.7885$$

PCA Example –8

- Step 7: PCA transformed vector T .

$$T = \begin{bmatrix} \phi_{00} & \phi_{10} \\ \phi_{01} & \phi_{11} \end{bmatrix} = \begin{bmatrix} -0.788 & 0.615 \\ 0.6156 & 0.7885 \end{bmatrix}$$

- Step 8: Verify if $TT^{\top} = \mathcal{I}$.

- Introduction to Data Mining, by Tan, Steinbach and Vipin Kumar (T3)
- The Art of Data Science by Roger D Peng and Elizabeth Matsui (R1)
- Introducing Data Science by Cielen, Meysman and Ali
- Data Science - Concepts and Practice by Vijay Kotuand BalaDeshpande
- Data mining: Concepts and techniques, by Han, J., Kamber, M., and Pei, J. (2012). (3rd ed.)

THANK YOU