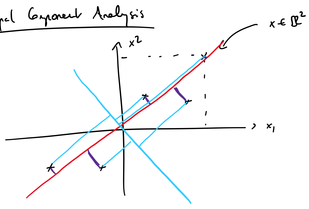
Video 5

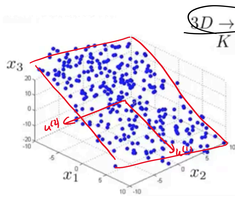
Thursday, September 17, 2020 12:17 PM

PCA : Principal Component Analysis

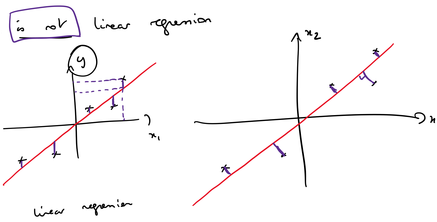


Reduce from 2D to 1D : Finds a direction (a vector $\in \mathbb{R}^2$) onto which to project the data so as to minimize the projection error.

Reduce from nD to kD : Find k vectors $u^{(1)}, \dots, u^{(k)} \in \mathbb{R}^n$



PCA is not linear regression



Linear regression

PCA algorithm Input : training set $x^{(1)}, \dots, x^{(m)}$

Preprocessing :

Mean normalization :

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

$$x_j^{(i)} \leftarrow x_j^{(i)} - \mu_j$$

Feature Scaling : Scale features to have comparable ranges of values.

Scale : s_j

$$x_j^{(i)} \leftarrow \frac{x_j^{(i)} - \mu_j}{s_j}$$

Algorithm : (nD → kD)

1. Compute covariance matrix :

$$\Sigma = \frac{1}{m} \sum_{i=1}^{m} \underbrace{\left(x^{(i)}\right)}_{n \times 1} \underbrace{\left(x^{(i)}\right)^T}_{1 \times n}$$

$n \times n$

2. Compute eigenvectors of Σ

$$U, S, V = \text{singular value decomposition}(\Sigma)$$

$$U = \begin{bmatrix} | & | & | & & | \\ u^{(1)} & u^{(2)} & u^{(3)} & \cdots & u^{(n)} \\ | & | & | & & | \end{bmatrix}$$

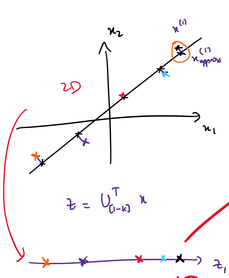
k

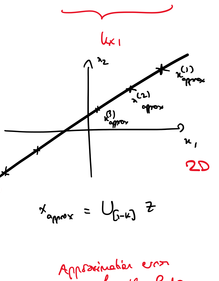
Project $x \in \mathbb{R}^n \longrightarrow z \in \mathbb{R}^k$

$$z = \begin{bmatrix} | & | & & | \\ u^{(1)} & u^{(2)} & \cdots & u^{(k)} \\ | & | & & | \end{bmatrix}^T x = \begin{bmatrix} - & u^{(1)} & - \\ & \vdots & \\ - & u^{(k)} & - \end{bmatrix} x$$

$k \times n$ $n \times 1$

$k \times 1$





Approximation error comes from the fact that we compressed data 2D → 1D

How do we choose k?

Average squared projection error : $\frac{1}{m} \sum_{i=1}^{m} \| x^{(i)} - x^{(i)}_{approx} \|^2$

Total variation in the data : $\frac{1}{m} \sum_{i=1}^{m} \| x^{(i)} \|^2$

Choose k such that :

$$\frac{\frac{1}{m} \sum_{i=1}^{m} \| x^{(i)} - x^{(i)}_{approx} \|^2}{\frac{1}{m} \sum_{i=1}^{m} \| x^{(i)} \|^2} \leq 0.01 \quad (1\%)$$

$\Longrightarrow$ 99% of the variance in the data is retained.

Applications / Tips on PCA:

- Data compression
- Visualization
- Speed-up learning : If apply PCA to train validation test