Sid

Dimensionality Reduction :

Data Compression $\{x^{(1)}, x^{(2)}, \dots, x^{(m)}\}$ $x^{(i)} \in \mathbb{R}^n$

$\downarrow$

Data Visualization $\{z^{(1)}, z^{(2)}, \dots, z^{(m)}\}$ $z^{(i)} \in \mathbb{R}^k$

$k \leq n$ (2 or 3)

Principal Component Analysis

- Reduce n-dim $\rightarrow$ k-dim : Find k vectors $u^{(1)} \dots u^{(k)}$ to project data & minimize $\underbrace{\text{projection error}}$
  - $\longrightarrow$ avg. of distances of features to projection plane/line
- $\neq$ linear regression (squared error vs shortest distance

PCA algorithm

- Data preprocessing (feature scaling/mean normalization)

$$\mu_j = \frac{1}{m} \sum_{i=1}^{m} x_j^{(i)} \implies \text{replace } x_j^{(i)} \text{ w/ } x_j^{(i)} - \mu_j$$

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

- Compute covariance matrix :

$$\underset{\Sigma}{[\ \ ]} = \frac{1}{m} \sum_{i=1}^{m} \underset{n\times 1}{(x^{(i)})} \underset{1\times n}{(x^{(i)})^T}$$

$$\overset{[\ \ ]}{\text{Sigma}} = \frac{1}{m} * X' * X ;$$

- Compute eigenvectors of matrix

$$[U, S, V] = \text{svd}\,(\underbrace{\text{Sigma}}_{\to\, n\times n})$$

$$U = \begin{bmatrix} | & | & & | \\ u^{(1)} & u^{(2)} & \dots & u^{(n)} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n\times n}$$

$x \in \mathbb{R}^n \to z \in \mathbb{R}^{\overbrace{k}^{n}}$ $\quad$ $z = Ureduce' * x;$

$$z^{(i)} = \underbrace{\begin{bmatrix} | & | & & | \\ u^{(1)} & u^{(2)} & \cdots & u^{(k)} \\ | & | & & | \end{bmatrix}}_{\substack{u \times k \\ (U_{reduce})}}^T \cdot x^{(i)} = \underbrace{\underbrace{\begin{bmatrix} — (u^{(1)})^T — \\ \vdots \\ — (u^{(k)})^T — \end{bmatrix}}_{k \times n} \underset{n \times 1}{x^{(i)}}}_{k \times 1}$$

```
Sigma = (1/m) * X' * X; % compute the covariance matrix
[U,S,V] = svd(Sigma);   % compute our projected directions
Ureduce = U(:,1:k);     % take the first k directions
Z = X * Ureduce;        % compute the projected data points
```

Reconstruction from compressed representation

$$\underset{\mathbb{R}^n}{X_{approx}} = \underbrace{\underbrace{U_{reduce}}_{n \times k} \cdot \underbrace{z}_{k \times 1}}_{n \times 1}$$

Choosing # of Principal components (k)

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

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

Choose k such that: (smallest)

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

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

$$S = \begin{bmatrix} S_{11} & & & 0 \\ & S_{22} & & \\ & & \ddots & \\ 0 & & & S_{nn} \end{bmatrix}$$

$$1 - \frac{\sum_{i=1}^{k} S_{ii}}{\sum_{i=1}^{n} S_{ii}}$$

(99% of variance retained)