lecture 3 : data preprocessing

Principal component analysis

given a dataset $X \in \mathbb{R}^{3\times 2}$, project the data using PCA

feature A ↓ feature B ↓

$$X = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \end{bmatrix}$$

#1 center X

$$\mu_A = \frac{2+1+0}{3} = 1 \qquad \mu_B = \frac{4+3+0}{3} = \frac{7}{3}$$

$$X_{center} = X - \mu = \begin{bmatrix} 1 & 5/3 \\ 0 & 2/3 \\ -1 & -7/3 \end{bmatrix}$$

#2 calculate covariance matrix $\Sigma = \frac{1}{n-1} X_{center}^T X_{center}$

$$\Sigma = \frac{1}{3-1} \cdot \begin{bmatrix} 1 & 5/3 \\ 0 & 2/3 \\ -1 & -7/3 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & -1 \\ 5/3 & 2/3 & -7/3 \end{bmatrix}$$

$$\Sigma = \frac{1}{2} \begin{bmatrix} 2 & 4 \\ 4 & 26/3 \end{bmatrix} = \begin{bmatrix} 1 & 2 \\ 2 & 13/3 \end{bmatrix}$$

#3 compute eigenvalues & eigenvectors
(eigenvalue decomposition)

$\Sigma = \lambda \cdot V$ → eigenvectors

λ ↙ eigenvalues

$\det(\Sigma - \lambda I) = 0$

$$\Sigma - \lambda I = \begin{bmatrix} 1-\lambda & 2 \\ 2 & 13/3 - \lambda \end{bmatrix}$$

$$\det(\Sigma - \lambda I) = (1-\lambda)\left(\frac{13}{3} - \lambda\right) - (2)(2)$$

$$= \frac{13}{3} + \lambda^2 - \frac{16}{3}\lambda - 4$$

$$= \lambda^2 - \frac{16}{3}\lambda + \frac{1}{3}$$

$$\Rightarrow \lambda_1 = \frac{-\sqrt{61} + 8}{3} \quad \text{and} \quad \lambda_2 = \frac{\sqrt{61} + 8}{3}$$

$= 0.06325$ $= 5.27008$

(too long for him to ask us for v_1 and v_2 so im skipping to the end)

$$v_1 = \begin{bmatrix} \frac{-\sqrt{61} - 5}{6} \\ 1 \end{bmatrix} \qquad v_2 = \begin{bmatrix} \frac{\sqrt{61} - 5}{6} \\ 1 \end{bmatrix}$$

$\Rightarrow \Sigma v_1 = \lambda_1 v_1$ and $\Sigma v_2 = \lambda_2 v_2$

#4 choose K largest eigenvalues (lets just do K=1)

$$\lambda_2 = \frac{\sqrt{61} + 8}{3} \quad \text{and} \quad v_2^T = \begin{bmatrix} \frac{\sqrt{61} - 5}{6} & 1 \end{bmatrix}$$

#5 project original data to k-dimensional space

$$Y = \underset{\text{center}}{X}_{3\times 2} \cdot \underset{}{W_k}_{2\times 1} \quad \text{s.t.} \quad Y \in \mathbb{R}^{n \times K} \ (3\times 1)$$

$$W_k \in \mathbb{R}^{d \times K} \quad W_k = \begin{bmatrix} v_1 & v_2 & \cdots \end{bmatrix} \quad (d\times 1,\ d\times 1)$$

$$W_k = \begin{bmatrix} \frac{\sqrt{61} - 5}{6} \\ 1 \end{bmatrix}$$

$$Y = \begin{bmatrix} 1 & 5/3 \\ 0 & 2/3 \\ -1 & -7/3 \end{bmatrix} \cdot \begin{bmatrix} \frac{\sqrt{61} - 5}{6} \\ 1 \end{bmatrix}$$

$$Y = \begin{bmatrix} 2.13504 \\ 2/3 \\ -2.80171 \end{bmatrix}$$

→ new data w/ one feature

back to #3 but with singular value decomposition

singular values

$$X_{\text{center}} = U \quad \Sigma \quad V^T$$

left singular vectors

right singular vectors (PC's)

$$X \in \mathbb{R}^{n \times d} \qquad U \in \mathbb{R}^{n \times n} \qquad \Sigma \in \mathbb{R}^{n \times d} \qquad V^T \in \mathbb{R}^{d \times d}$$

from the same eigenvalue decomposition process

$$\Sigma = \begin{bmatrix} \sqrt{\lambda_1} & 0 \\ 0 & \sqrt{\lambda_2} \end{bmatrix} \qquad V = \begin{bmatrix} \vec{v}_1 & \vec{v}_2 \end{bmatrix}$$

↳ PC's so essentially its the same thing

$$U = X_{\text{center}}^{-1} \Sigma V^T$$

lecture 4: similarity measures

Pearson correlation

$A = [1, 2, 3, 4, 5]$ $\quad B = [2, 4, 6, 8, 10]$

$$p(A,B) = \frac{cov(A,B)}{\sigma_A \sigma_B}$$

$$\bar{A} = 15/5 = 3 \qquad \bar{B} = 30/5 = 6$$

$$\sigma_A^2 = \frac{1}{5} \Sigma (a_i - \bar{A})^2 = \frac{10}{5} = 2$$

$$\sigma_A^3 = \frac{1}{5} \Sigma (b_i - \bar{B})^2 = \frac{40}{5} = 8$$

$$cov(A,B) = \frac{1}{5} \Sigma (a_i - \bar{A})(b_i - \bar{B}) = \frac{20}{5} = 4$$

$$p(A,B) = \frac{4}{\sqrt{2} \cdot \sqrt{8}} = 1$$

→ perfect +ve linear relationship

there's also the non-linear case which im skipping

euclidean distance

$$d(A,B) = \sqrt{\Sigma (a_i - b_i)^2}$$

$$A = [1 \quad 2 \quad 3] \qquad B = [4 \quad 5 \quad 6]$$

$$\sum (a_i - b_i)^2 = 3^2 + 3^2 + 3^2 = 27$$

$$\sqrt{\sum (a_i - b_i)^2} = \sqrt{27} = 3\sqrt{3}$$

manhattan distance

$$d(A,B) = \sum |a_i - b_i|$$

$$A = [1 \quad 2 \quad 3] \qquad B = [4 \quad 5 \quad 6]$$

$$d(A,B) = |-3| + |-3| + |-3| = 9$$

cosine similarity

$$d(A,B) = \frac{A \cdot B}{\|A\| \cdot \|B\|}$$

$$A = [1 \quad 2 \quad 3] \qquad B = [4 \quad 5 \quad 6]$$

$$A \cdot B = (1)(4) + (2)(5) + (3)(6) = 32$$

$$\|A\| = \sqrt{1^2 + 2^2 + 3^2} = \sqrt{14}$$

$$\|B\| = \sqrt{4^2 + 5^2 + 6^2} = \sqrt{77}$$

$$d(A,B) = \frac{32}{\sqrt{14} \cdot \sqrt{77}} \approx 0.975 \rightarrow \text{very similar}$$

if $B = [2\ 4\ 6] = 2A$, then $d(A,B) = 1$

jaccard similarity

$$d(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$A = \{1, 2, 3, 4, 5\}$ $B = \{4, 5, 6, 7\}$

$$d(A,B) = \frac{2}{7} = 0.287$$