Machine Learning

Complete Name: Carlos Castro

Understanding PCA and SVD

- **Part 1 – Theory**

1. **What is PCA?**
   Principal Component Analysis (PCA) is a statistical procedure that extracts the most important features of a dataset.
2. **What is SVD?**
   Singular Value Decomposition (SVD) is a method in linear algebra that decomposes a matrix into three simpler matrices
3. **How are they related?**
   PCA can be done using SVD. The principal components in PCA are the right singular vectors from SVD, and the singular values show how much variance each component explains.
4. **Compute the covariance matrix.**

• Compute the covariance matrix

$$A = \begin{pmatrix} 2 & 0 \\ 0 & 2 \\ 3 & 1 \\ 1 & 3 \end{pmatrix}$$

• The covariance matrix can be expressed in the following way

$$Q_n = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^T$$

where x_i is the ith row of the sample matrix

$\bar{x} = (1.5, 1.5)$

$x_1 = (2,0)$; $x_3 = (3,1)$

$x_2 = (0,2)$; $x_4 = (1,3)$

$$= \frac{1}{4}\left(\begin{pmatrix} 0.25 & -0.75 \\ -0.75 & 2.25 \end{pmatrix} + \begin{pmatrix} 2.25 & -0.75 \\ -0.75 & 0.25 \end{pmatrix} + \begin{pmatrix} 2.25 & -0.75 \\ -0.75 & 0.25 \end{pmatrix} + \begin{pmatrix} 0.25 & -0.75 \\ -0.75 & 2.25 \end{pmatrix} \right)$$

$$= \frac{1}{4}\begin{pmatrix} 5 & -3 \\ -3 & 5 \end{pmatrix}$$

$$= \begin{pmatrix} \frac{5}{4} & -\frac{3}{4} \\ -\frac{3}{4} & \frac{5}{4} \end{pmatrix}$$

5. **Calculate the eigenvalues and eigenvectors.**

Hacienda San José S/N Urcuquí - Ecuador | Cod. Postal 100119 | Telf.: +593 6 2 999 500 | info@yachaytech.edu.ec
www.yachaytech.edu.ec

We don't have a square matrix we use data of covariance matrix

$$C = \begin{pmatrix} \frac{5}{4} & -\frac{3}{4} \\ -\frac{3}{4} & \frac{5}{4} \end{pmatrix}$$

Eigenvalues of C:

$$\det \begin{pmatrix} \frac{5}{4}-\lambda & -\frac{3}{4} \\ -\frac{3}{4} & \frac{5}{4}-\lambda \end{pmatrix} = 0$$

$$\left(\frac{5}{4}-\lambda\right)^2 - \left(-\frac{3}{4}\right)^2 = 0$$

$$\frac{25}{4} - \frac{5}{2}\lambda + \lambda^2 - \frac{9}{8} = 0$$

$$\lambda_1 = 0.5$$

$$\lambda_2 = 2$$

*

Calculate the eigenvalues λ_1

$$\begin{pmatrix} \frac{5}{4}-0.5 & -\frac{3}{4} \\ -\frac{3}{4} & \frac{5}{4}-0.5 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} v_1 = -0.7071 \\ v_2 = -0.7071 \end{pmatrix}$$

$\lambda_2 = 2$

$$\begin{pmatrix} \frac{5}{4}-2 & -\frac{3}{4} \\ -\frac{3}{4} & \frac{5}{4}-2 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} -0.7071 \\ 0.7071 \end{pmatrix}$$

6. **Interpret which direction has the highest variance.**

The direction with the highest variance corresponds to the eigenvector with the largest eigenvalue

$$\lambda_{max} = 2$$

Its associated eigenvector

$$v = \begin{pmatrix} -0.7071 \\ 0.7071 \end{pmatrix} \quad //$$

- **Part 3 – Reflection**

1. **Why is variance important in PCA?**
   Variance is important in PCA because it measures how much information or spread exists in the data along a certain direction. PCA keeps the directions with the highest variance because they preserve the most meaningful structure of the dataset while reducing dimensionality.
2. **Why is SVD often used instead of computing eigenvalues directly?**
   SVD is often used instead of eigenvalue decomposition because it is more numerically stable and can handle cases where the covariance matrix is large or not full rank. SVD also avoids the explicit computation of the covariance matrix, which improves accuracy and efficiency.
3. **In what real-world situations would you use PCA/SVD?**
   PCA and SVD are used in many real-world applications such as image compression, noise reduction, and face recognition in computer vision, topic extraction in text analysis, and dimensionality reduction in recommender systems and machine learning pipelines.

Hacienda San José S/N Urcuquí - Ecuador I Cod. Postal 100119 I Telf.: +593 6 2 999 500 I info@yachaytech.edu.ec
www.yachaytech.edu.ec