UNIT V

PRINCIPAL COMPONENT ANALYSIS (PCA)

SYLLABUS:

Data Matrices - Projections - SSE Goal - Singular Value Decomposition - Best Rank-k Approximation Principal Component Analysis - Computation of PCA Components - Reduction of Two dimension data set to one dimension – Drawing Graph for PCA

5.1 Introduction

Principal Component Analysis (PCA) is a technique used to understand the **Shape** of a given Data**.**

Data Matrices

Let us consider a data in the matrix $A \in R^{n\times d}$ form having n rows and d columns. Here each row represents a **data point** and each column denote **attribute** of the data points. Also here d stands for **dimension** of the matrix A.

Examples of Data matrices:

1. Suppose 2 whether stations (located at different places) give **temperatures** at 3 different times in celcius degrees in the afternoon session. Then the Data matrix form shall be like this.

$$\begin{array}{cc} & \begin{array}{ccc} 2.30 & 3.00 & 3.30 \end{array} \\ \begin{array}{c} a_1 \\ a_2 \end{array} & \begin{pmatrix} 20 & 22 & 21 \\ 25 & 23 & 20 \end{pmatrix} \end{array}$$

2. Suppose n users rate d movies by giving scores between 1 – 5. Then a_i represents an ith user and d_j stands for jth movie and $A_{i,j}$ stands for the score given by that user to the jth movie.

For example if $A \in R^{3\times 4}$ and $A = \begin{pmatrix} 1 & 2 & 3 & 4 \\ 5 & 3 & 2 & 1 \\ 3 & 1 & 2 & 4 \end{pmatrix}$. The structure of this matrix is $A = \begin{pmatrix} A_{1,1} & A_{1,2} & A_{1,3} & A_{1,4} \\ A_{2,1} & A_{2,2} & A_{2,3} & A_{2,4} \\ A_{3,1} & A_{3,2} & A_{3,3} & A_{3,4} \end{pmatrix}$. Then the A matrix represents 3 users (a_1, a_2, a_3) give scores to 4 movies (d_1, d_2, d_3, d_4) between 1 to 5. Also the entry $A_{24} = 1$ stands that second user has given score 1 to the 4^{th} movie.

5.2 Projections

Let $u \in R^d$ be a given unit vector. Let $p \in R^d$ be any data point. Then the dot product $\langle u, p\rangle$ is the **norm of p projected onto the line through u. And** $\langle u, p\rangle$ is a scalar.

Multiply this scalar $\langle u, p \rangle$ with u, then we get $\boldsymbol{\pi_p(u)} = \langle \boldsymbol{u}, \boldsymbol{p} \rangle\, \boldsymbol{u}$, which is point on the line through u that is closest to data point p. This is **a projection of a data point p onto u.**

Let **F** be a k dimensional space and $p \in R^d$ be any data point. The projection on to F of a data point $p \in R^d$ is given by the formula

$$\boldsymbol{\pi_F(p)} = \sum_{i=1}^{k} \langle \boldsymbol{u_i}, \boldsymbol{p} \rangle\, \boldsymbol{u}_i$$

Note:
Let F be a d dimensional space. Let the basis of F be $U_F = \{u_1, u_2, u_3, \ldots . u_d\}$
$u_i = e_i = (0,0,0, \ldots ., 1, 0, 0,0)$.

5.3 SSE Goal:

Our goal is to **minimize the sum of squared errors**. Let F be a k dimensional subspace and A be a given data matrix. Then Error Sum of Squares can be calculated by using the given formula.

$$\textbf{SSE (A, F)} = \sum_{a_i \in A} \| \boldsymbol{a_i} - \boldsymbol{\pi_F(a_i)} \|^2$$

And k dimensional subspace F is $F^* = \arg\min\ \mathrm{SSE(A, F)}$.