Worksheet on "Principal component Analysis"

PRML – CS5691 (Jul–Nov 2023)

October 11, 2023

1. Consider a dataset of N points with each datapoint being a D-dimensional vector in $\mathbb{R}^D$. Let's assume that:

   - we are in a high-dimensional setting where $D >> N$ (e.g., D in millions, N in hundreds).
   - the $N \times D$ matrix X corresponding to this dataset is already mean-centered (so that each column's mean is zero, and the covariance matrix seen in class becomes $S = \frac{1}{N}X^TX$).
   - the rows (datapoints) of X are linearly independent.

   Under the above assumptions, please attempt the following questions.

   (a) Whereas X is rectangular in general, XX^T and X^TX are square. Show that these two square matrices have the same set of non-zero eigenvalues. Further, argue briefy why these equal eigenvalues are all positive and N in number, and derive the multiplicity of the zero eigenvalue for both these matrices.
   (Note: The square root of these equal positive eigenvalues $\{\lambda_i := \sigma_i^2\}_{i=1,\ldots,N}$ are called the singular values $\{\sigma_i\}_{i=1,\ldots,N}$ of X.)

   (b) We can choose the set of eigenvectors $\{u_i\}_{i-=1,\ldots,N}$ of XX^T to be an orthonormal set and similarly we can choose an orthonormal set of eigenvectors $\{v_j\}_{j=1,\ldots,D}$ for X^TX. Briefly argue why this orthonormal choice of eigenvectors is possible. Can you choose $\{v_i\}$ such that each v_i can be computed easily from u_i and X alone (i.e., without having to do an eigenvalue decomposition of the large matrix X^TX; assume $i = 1, \ldots, N$ so that $\lambda_i > 0$ and $\sigma_i > 0$)?
   (Note: $\{u_i\}, \{v_i\}$ are respectively called the left,right singular vectors of X, and computing them along with the corresponding singular values is called the Singular Value Decomposition or SVD of X.)

   (c) Applying PCA on the matrix X would be computationally difficult as it would involve finding the eigenvectors of $S = \frac{1}{N}X^TX$, which would take $O(D^3)$ time. Using answer to the last question above, can you reduce this time complexity to $O(N^3)$? Please provide the exact steps involved, including the exact formula for computing the normalized (unit-length) eigenvectors of S.

2. Source: CMU School of Computer Science (Fall 2008)
   Given 3 data points in 2-D space, $(1, 1)$, $(2, 2)$, and $(3, 3)$,

   (a) What is the first principle component?

   (b) If we want to project the original data points into 1-D space by the principle component you choose, what is the variance of the projected data?

   (c) For the projected data in (b), now if we represent them in the original 2-D space, what is the reconstruction error?

3. Source: CMU School of Computer Science (Fall 2008)
   Given 6 data points in 5-D space, $(1, 1, 1, 0, 0)$, $(-3, -3, -3, 0, 0)$, $(2, 2, 2, 0, 0)$, $(0, 0, 0, -1, -1)$, $(0, 0, 0, 2, 2)$, $(0, 0, 0, -1, -1)$. We can represent these data points by a 6×5 matrix X, where each row corresponds to a data point:

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

Note: Use numpy code to compute the eigen vectors. [can also be solved using Singular value decomposition]

(a) What is the sample mean of the data set?

(b) What is the first principle component for the original data points?

(c) If we want to project the original data points into 1-D space by the principle component you choose, what is the variance of the projected data?

(d) For the projected data in (c), now if we represent them in the original 5-D space, what is the reconstruction error?

4. Consider a dataset consisting of n data points with each datapoint being a D-dimensional vector in $\mathbb{R}^D$. What can you say about the covariance matrix and the principal components if there is no correlation between the features?

5. Given a dataset X:

$$X = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Calculate the covariance matrix and the corresponding eigenvectors. Determine the minimal number of principal components required to retain at least 90% of the variance in the dataset.